Parametric assumptions in single and multiple testing: when should we rely on them?

Alessio Farcomeni

"La Sapienza" University of Rome, Italy

Corresponding Author:
Alessio Farcomeni
Department of Experimental Medicine, "Sapienza" University of Rome
P.le Aldo Moro 5, 00185 Rome, Italy
E-mail: alessio.farcomeni@uniroma1.it

Summary
Testing for difference in location among two or more groups is an everyday problem in data analysis. Parametric (normality) assumptions are often taken for granted without much investigation. In single inference this may be a negligible issue if the sample size is large enough and, when necessary, a suitable transformation is applied to the numeric variable. In this paper we argue that the normality assumption should almost always be withdrawn in the complex setting of multiple testing. We support the claim with a small simulation and a case study on microRNA profiling of human medulloblastoma.

KEY WORDS: *t-test, Mann-Whitney test, multiple testing, nonparametric inference, ranks.*

Introduction

Comparing a numeric variable among two or more groups is one of the most common problems in biomedical statistics. In general, a statistic for location is compared, and the emergence of statistical significance leads to findings of interest to the biomedical community.

For the purposes of this paper, we will focus on the extremely simple and most common situation: that of two groups.

The *t*-test is used for comparing means, but it relies on normality. It is well known that the *t*-test is robust to small departures from the normality assumption, and that it becomes distribution-free as the sample size *n* grows.

A possible alternative is the Mann-Whitney test (1, 2), which does not require any normality assumption, and is deemed to compare medians (this is not strictly true but we will assume that it is for the time being).

Even though these two tests very often lead to the same conclusions, there are situations in which they can contradict each other.

As an example, let us consider recent data (3) documenting the effects of total or partial parathyroidectomy on serum levels of calcium, phosphate and parathyroid hormone (PTH) in dialysis patients. The sample comprised $n = 77$ patients, with no losses to follow up in the first year.

First, we compute the standardized difference between PTH levels before and one month after surgery. These standardized levels are compared between the 36 patients undergoing total removal of the thyroid and the 41 undergoing only partial removal. With the Mann-Whitney test statistic a *p*-value of 0.0013 is obtained. This *p*-value indicates that there is a differential effect of the kind of surgery on PTH serum levels, i.e. that total surgery leads to lower levels of PTH in blood after one month (median, -1.91 vs -1.71). On the other hand, were we to use the *t*-test, a *p*-value of 0.101 would be obtained, leading us to declare that there is no significant difference between the means of -1.70 and -1.50. This is related to the strong skewness of the standardized difference. We note that the Levene test for homogeneity of variances is not rejected with a *p*-value of 0.4117.

The reverse can often occur, i.e. the *t*-test leads to rejection but nonparametric testing does not.

We argue that the choice of the test should be made before seeing the outcome. Otherwise, if the two tests contradict each other, use of the significant one is data snooping and inflates the actual level of the test.

In our experience, the choice of test (when this is considered) is often driven by the habits of the data analyst, the type of test traditionally used in the biomedical sector for which the data analysis is being done, and only sometimes by a careful exploration of the normality of the continuous variable. We argue that even the last of these approaches is not always sensitive, and that in many applications it cannot even be performed. Further, normality is often assessed simply by eye-balling a histogram. This can be misleading in that distributions with heavy tails can easily be mistaken for normal distributions.

The main task of this paper is to explore, via a case study and a small simulation, the practical differences in opting for the route of *t*-testing versus Mann-Whitney. We will not consider, for the time being, the effects of transformations designed to make the *t*-test more grounded on the normality assumption. We are interested primarily in the effects of this choice in multiple testing, that is, when two or more tests are performed at the same time. Whereas when just a single test is performed parametric and nonparametric methods may lead to the same conclusions, when many tests are performed at the same time the outcomes may easily not coincide, as we will see below.

While tests for difference in location among populations have been used for many decades, the topic of this paper is of current interest due to the increasingly frequent need to use these tests in applications in which (i) the sample size is small, so that normality cannot be tested and the central limit theorem (CLT) will not necessarily hold, and (ii) the number of simultaneous tests of the same nature is high. These two factors (small sample size, use of a *vector* instead of just a single *p*-value) combine to make multiple testing different from the single test situation. Furthermore, in applications such as gene identification in DNA microarrays, some outliers can often be included in the data, which makes the *t*-test possibly unreliable.

Applications of tests for locations in multiple testing include identifying neuronal activity in the living brain (4-7), and the identification of differentially expressed genes in DNA microarray experiments (8-13). There is a plethora of other situations in which it is common for many tests to be performed at the same time, in bioinformatics, psychometrics (14), epidemiology (15), pharmacology (16), etc.

Many of these applications have arisen recently, posing new kinds of multiplicity problems and stimulating a tremendous interest and fast developments in multiple hypothesis testing. We have recently reviewed this aspect (17).

Our final claim will be that the *t*-test, in the previously discussed situations, may be frequently put aside in favour of its nonparametric counterpart.

In this paper, following a discussion of the application of location comparisons in multiple inference situations, and a brief background on multiple testing, our simulation studies will be presented, followed by a case study analysing original data on gene discovery.

Testing difference in location between two groups

The *t*-test is a standard approach used to verify the difference between the means of two normally distributed populations with the same but unknown variances. In practice, the difference in mean is compared to the experimental variability (the standard error) and if it is large enough with respect to differences that can reasonably be expected to be produced by chance, the null hypothesis that the two means are equal is rejected.

The data must come from a normally distributed population. It is customary, if necessary, to use a Box-Cox transformation (for instance, taking the logarithm) in order to improve the approximation to normality. When the sample size n is sufficiently large one can use the CLT to assume that the sample average $\bar{x}$ is approximately normally distributed. In that case, the true data generating distribution is not an issue. The main problem is that when n is small the distributional assumption cannot even be verified: common techniques for assessing normality (like histograms, q-q plots, or formal tests like the Kolmogorov-Smirnov) are not useful and the CLT cannot be invoked. In these situations one does not know whether the data are normal or not, and the safest route is to use tests for location which are distribution-free.

A simple alternative to *t*-testing is the Mann-Whitney statistic, which is based on ranks. The entire sample is ranked, and the ranks of one of the two groups are summed. The final Mann-Whitney statistic is $U = \sum_{i=1}^{n_1} R_{i1} - i$, or equivalently $U = (\sum_{i=1}^{n_1} R_{i1}) - \frac{n_1(n_1+1)}{2}$, where R_{i1} are the ranks of the first (arbitrarily chosen) group, and n_1 is the sample size of the first group. From the *U* statistic *p*-values are easily computed. In the case of small samples the distribution is tabulated, but for samples above about 20 the approximation is good using the normal distribution.

In practice, the order of the two groups is compared without taking into account the actual numerical differences. If the groups are separated, as in the case of Example 1 in Table 1, then it is obvious that there is a difference in location, and in fact the *U* statistic will be exactly zero if A is chosen as first group, and n_1n_2 if it is chosen as the second. It is straightforward to check that n_1n_2 is the maximum possible value for *U*. If the groups are mixed, as in the case of Example 2 in Table 1, then one can expect the two groups to come from the same population.

Table 1. Example data with separated groups (Example 1) and with mixed groups (Example 2) (data are ordered with respect to variable VAR).

VAR	Example 1	Example 2
0.05	A	A
0.10	A	B
0.12	A	A
0.15	A	B
0.20	A	A
0.40	A	B
1.90	B	A
2.00	B	B
2.30	B	A
2.35	B	B
2.40	B	A
2.50	B	B

The Mann-Whitney is thus intuitive, simple, and distribution-free (to derive conclusions about medians, it only assumes that the two distributions have the same shape apart from the tested shift). Furthermore, unlike the *t*-test, it is invariant with respect to monotone transformations, it can be used to compare locations of categorical ordered variables, and it is much less sensitive to spurious outliers. However, there are drawbacks. The most important is that if the data are truly normal, it is asymptotically about 4.5% less efficient than the *t*-test (to be precise, (π – 3) / π). This is exacerbated in very small sample situations: in Table 2 we show the minimum *p*-value that can be realized as a function of sample size, with n_1 being the sample size for the first group, and n_2 the sample size for the second group. The minimum *p*-values in Table 2 are the *p*-values obtained if the two groups are perfectly separated. It can be seen, in practice, that with as few as three observations per group, the test will never reject at level $\alpha = 0.05$ or lower, no matter what the true shift. To reject the null hypothesis at level $\alpha = 0.05$ with a Mann-Whitney test at least four observations per group are needed, and five for significance at level $\alpha = 0.01$. In unbalanced situations, an even larger total sample size is needed. We note that with *t*-testing, two observations per group suffice to achieve arbitrarily small *p*-values.

In real situations there can be a moderate overlap between the groups due to volatility or small dimension

Table 2. Minimum *p* achievable with Mann-Whitney as a function of sample size.

					n_2				
n_1	2	3	4	5	6	7	8	9	10
2	0.333	0.200	0.133	0.095	0.071	0.056	0.044*	0.036*	0.030*
3		0.100	0.057	0.036*	0.024*	0.017*	0.012*	0.009**	0.007**
4			0.029*	0.016*	0.009**	0.006**	0.004**	0.003**	0.002**
5				0.008**	0.004**	0.002**	0.002**	0.001**	0.001**
6					0.002**	0.001**	0.001**	0.000**	0.000**

Abbreviations and symbols: n_1 = sample size for the first group; n_1 = sample size for the second group; * = significant at level $\alpha = 0.05$, ** = significant at level $\alpha = 0.01$.

Table 3. *p*-value with Mann-Whitney as a function of sample size with second group above first except for one single observation, which corresponds to the minimum value.

	n_2								
n_1	2	3	4	5	6	7	8	9	10
2	1.0000	0.8000	0.5333	0.3810	0.2857	0.2222	0.1778	0.1455	0.1212
3	1.0000	0.7000	0.4000	0.2500	0.1667	0.1167	0.0848	0.0636	0.0490*
4	1.0000	0.6286	0.3429	0.1905	0.1143	0.0727	0.0485*	0.0336*	0.0240*
5	1.0000	0.5714	0.2857	0.1508	0.0823	0.0480*	0.0295*	0.0190*	0.0127*
6	1.0000	0.5476	0.2571	0.1255	0.0649	0.0350*	0.0200*	0.0120*	0.0075**
7	1.0000	0.5167	0.2303	0.1061	0.0513	0.0262*	0.0140*	0.0079**	0.0046**
8	1.0000	0.4970	0.2141	0.0932	0.0426*	0.0205*	0.0104*	0.0055**	0.0031**
9	1.0000	0.4818	0.1986	0.0829	0.0360*	0.0164*	0.0079**	0.0040**	0.0021**
10	1.0000	0.4685	0.1878	0.0753	0.0312*	0.0136*	0.0062**	0.0030**	0.0015**

Abbreviations and symbols: n_1 = sample size for the first group; n_1 = sample size for the second group; * = significant at level $\alpha = 0.05$, ** = significant at level $\alpha = 0.01$.

of the effect (i.e., small difference in location), which makes observed *p*-values even higher. To give an idea, Table 3 gives the p-value that is achieved in the situation in which groups are perfectly separated except for one single observation from the second group being below the minimum in the first group.

The case of multiple testing

Let us consider a multiple testing situation in which *m* tests are being performed. For each test, significance is assessed via a *p*-value, which can arise from any kind of test. While the methods for multiple testing apply in the same way in all cases, the choice of error rate, correction, and test can affect power strongly.

In the usual (single) test setting, one controls the probability of false rejection (Type I error) while looking for a procedure that possibly minimizes the probability of observing a false negative (Type II error).

In the multiple case, despite the small probability of each uncorrected level α test falsely rejecting the null hypothesis, as *m* increases the total number of false discoveries will obviously increase dramatically. Corrections are needed to control specific Type I error measures. There are various functions of false positive counts that can serve as possible generalizations of the probability of Type I error. Control of the chosen Type I error rate can be loosely defined to be achieved when the error rate is bounded above by a pre-specified α, which usually is fixed at 0.05, 0.01 or 0.1.

More formally, let us suppose M_0 of the *m* null hypotheses are true, and M_1 are false. Table 4 shows the possible outcomes in testing *m* hypotheses: we denote with *R* the number of rejections, with $N_{0|1}$ and $N_{1|0}$ the exact (unknown) number of errors made after testing, and with $N_{1|1}$ and $N_{0|0}$ the number of correctly rejected and correctly retained null hypotheses.

Table 4. Outcomes in testing *m* hypotheses.

	H_0 not rejected	H_0 rejected	Total
H_0 True	$N_{0\|0}$	$N_{1\|0}$	M_0
H_0 False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$m - R$	R	m

A classical multiple Type I error rate is the *family-wise error rate* (FWE), i.e. the probability of a least one Type I error:

$$FWE = \Pr(N_{1|0} \geq 1).$$

The FWE can be controlled with the famous Bonferroni correction, simply by rejecting only the hypotheses corresponding to *p*-values below α / *m*. There are many improvements that can be made to the Bonferroni, mainly resulting in data-dependent cut offs given by step-down or step-up procedures.

In step-down procedures the *p*-values are compared in order, from smallest to largest, with a rank-specific cut-off. Once a *p*-value is greater than its cut-off, the corresponding hypothesis is not rejected, and neither are all those higher than it. Step-up procedures are similar. The p-values are examined from the largest to the smallest. Once a p-value is found to be smaller than its rank-specific cut-off, the corresponding hypothesis is rejected, together with all the smaller ones.

Since the significance of one hypothesis is related not only to its corresponding p-value but also to the p-values of all the other hypotheses, small changes in the p-value vector may cause the number, and list, of rejected hypotheses to vary wildly. These changes can be due to many factors, one of which is the choice between parametric and nonparametric tests for computing p-values.

In this paper we will use Holm's step-down method (18) for controlling the FWE. Holm's step-down method starts by fixing the step-down constant $\alpha / (m - j + 1)$, that is, the j-th p-value is compared with $\alpha / (m - j + 1)$ in a step-down fashion. This controls the FWE at level α.

Control of the FWE guarantees that with high probability the list of rejections will be free of false rejections. All rejections can thus be expected with high probability to be true findings. This is a very interesting feature for practitioners, but it has a drawback. In fact, when the number of tests is large (for instance, in the order of the thousands), FWE control can become overly conservative, basically resulting in a disappointingly low number of rejections. For this reason, Benjamini and Hochberg (19) suggest using a more liberal error measure known as the *false discovery rate* (FDR), which can be defined as the expected proportion of the number of erroneously rejected hypotheses to the number of rejections, if any. Control of the FDR turns out to give a much better balance between false rejections and number of correct rejections (that is, power) when m is large. Simulation studies comparing FWE and FDR control are reviewed elsewhere (17).

FDR at level α can be controlled using Benjamini and Yekutieli's approach (20), which fixes the step-up constant: $\frac{j\alpha}{m\sum_{i=1}^{m}(1/i)}$.

The proposed corrections, (Holm's step-down and Benjamini and Yekutieli's step-up) control the corresponding Type I error rate in finite samples and with arbitrary dependence among the p-values. The only requirement is that the p-values be *valid*, so that there is no bias as could be present when the t-test is applied to a variable far from normality. A detailed review on statistical methods for multiple testing can be found elsewhere (17).

Simulations

We simulate data from three distributions: a standard normal distribution, a t-distribution with three degrees of freedom, and an exponential distribution with rate equal to 1. In all cases we add a constant shift of δ to the second group, and simulate in the balanced $n_1 = n_2$ situation. The t_3 distribution is heavy tailed but symmetric, and would very likely be recognized as normal by any user judging normality from a histogram. The exponential is highly skewed.

We generate $B = 5000$ data sets and record the out-

Table 5. Probability of Type II error for Student's t-test (β_T) and Mann-Whitney test (β_{MW}), and proportion of simulated data sets where they show agreement for different n and different δ under normality. Nominal error levels: $\alpha = 0.01$ and $\alpha = 0.05$.

			$\alpha = 0.01$			$\alpha = 0.05$	
$n_1 = n_2$	δ	β_T	β_{MW}	$I_{T=MW}$	β_T	β_{MW}	$I_{T=MW}$
5	0.20	0.99	0.99	1.00	0.95	0.97	0.98
10	0.20	0.98	0.99	0.99	0.92	0.93	0.98
20	0.20	0.98	0.98	0.99	0.91	0.91	0.97
50	0.20	0.94	0.95	0.98	0.83	0.84	0.96
100	0.20	0.88	0.88	0.97	0.71	0.72	0.94
5	2.50	0.35	0.41	0.83	0.08	0.14	0.92
10	2.50	0.01	0.01	0.99	0.00	0.00	1.00
20	2.50	0.00	0.00	1.00	0.00	0.00	1.00
50	2.50	0.00	0.00	1.00	0.00	0.00	1.00
100	2.50	0.00	0.00	1.00	0.00	0.00	1.00

Table 6. Probability of Type II error for Student's t-test (β_T) and Mann-Whitney test (β_{MW}), and proportion of simulated data sets where they show agreement for different n and different δ under t_3-distributed data. Nominal error levels: $\alpha = 0.01$ and $\alpha = 0.05$.

		$\alpha = 0.01$			$\alpha = 0.05$		
$n_1 = n_2$	δ	β_T	β_{MW}	$I_{T=MW}$	β_T	β_{MW}	$I_{T=MW}$
5	0.20	1.00	0.99	0.99	0.97	0.97	0.97
10	0.20	0.99	0.99	0.99	0.95	0.94	0.97
20	0.20	0.98	0.98	0.99	0.93	0.92	0.96
50	0.20	0.97	0.96	0.97	0.90	0.87	0.93
100	0.20	0.96	0.93	0.95	0.86	0.80	0.89
5	2.50	0.65	0.62	0.85	0.39	0.38	0.89
10	2.50	0.25	0.19	0.90	0.11	0.06	0.93
20	2.50	0.06	0.01	0.95	0.02	0.00	0.98
50	2.50	0.01	0.00	0.99	0.00	0.00	1.00
100	2.50	0.00	0.00	1.00	0.00	0.00	1.00

Table 7. Probability of Type II error for Student's t-test (β_T) and Mann-Whitney test (β_{MW}), and proportion of simulated data where they show agreement for different n and different δ under exponentially distributed data. Nominal error levels: $\alpha = 0.01$ and $\alpha = 0.05$.

		$\alpha = 0.01$			$\alpha = 0.05$		
$n_1 = n_2$	δ	β_T	β_{MW}	$I_{T=MW}$	β_T	β_{MW}	$I_{T=MW}$
5	0.20	0.99	0.99	0.99	0.96	0.95	0.97
10	0.20	0.99	0.97	0.98	0.93	0.91	0.96
20	0.20	0.98	0.94	0.97	0.90	0.83	0.91
50	0.20	0.94	0.85	0.90	0.82	0.65	0.82
100	0.20	0.87	0.64	0.76	0.70	0.40	0.69
5	2.50	0.31	0.29	0.88	0.10	0.19	0.91
10	2.50	0.03	0.03	0.97	0.01	0.01	0.99
20	2.50	0.00	0.00	1.00	0.00	0.00	1.00
50	2.50	0.00	0.00	1.00	0.00	0.00	1.00
100	2.50	0.00	0.00	1.00	0.00	0.00	1.00

come of each of the two tests. In Tables 5, 6 and 7 we show the results for different n_1 and δ in the single test settings. β denotes the proportion of data sets for which the test is not rejected at level α. This is the Type II error rate (1-power) whenever $\delta > 0$. By $I_{T=MW}$ we mean the proportion of data sets for which the two tests lead to the same decision.

As expected, under normality (Table 5) the t-test is more powerful in all cases, but the two tests substantially agree. The lower bound for the proportion of times they agree is 0.83, with the lowest sample size (=5) and largest δ (=2.5). So, not much of an improvement is obtained by assuming normality when this is uncertain, given that the Mann-Whitney test performs more or less the same as the t-test under normality. On the other hand, when the data are not normal (Tables 6 and 7), the Mann-Whitney test is more powerful and the two tests can agree with lower probability (as low as 69%).

We now turn to the multiple testing situation. For reasons of space we report results only for $\alpha = 0.05$. We generate data sets for testing m hypotheses, 10% of which are false nulls with a difference in location of δ. We report the actual value of the controlled Type I error measure, and the average number of true rejections ($\overline{N}_{1|1}$) for each kind of test, together with the proportion of simulations (over $B = 1000$) in which the two tests lead to the same list of rejected hypotheses. The results are presented in Tables 8, 9, and 10.

Under normality, irrespective of the controlled error measure, use of the t-test often but not always leads to higher power. The difference in power is rarely substantial though, reaching a maximum of around

Table 8. FWE, FDR, average number of correct rejections ($\bar{N}_{1|1}$) over the 0.1 m possible for Student's t-test and Mann-Whitney test, together with proportion of simulated data sets where they show agreement, for different m, n and δ under normality (nominal $\alpha = 0.05$).

			t-test		Mann-Whitney			*t*-test		Mann-Whitney		
m	n	δ	*FWE*	$\bar{N}_{1\|1}$	*FWE*	$\bar{N}_{1\|1}$	$I_{T=MW}$	*FDR*	$\bar{N}_{1\|1}$	*FDR*	$\bar{N}_{1\|1}$	$I_{T=MW}$
20	10	0.20	0.041	0.013	0.040	0.015	0.98	0.010	0.006	0.010	0.007	0.99
20	20	0.20	0.049	0.022	0.044	0.023	0.97	0.010	0.009	0.013	0.008	0.99
20	50	0.20	0.043	0.085	0.042	0.072	0.95	0.012	0.033	0.009	0.029	0.97
20	100	0.20	0.043	0.168	0.041	0.153	0.92	0.014	0.087	0.009	0.075	0.96
20	10	2.50	0.048	2.909	0.046	2.858	0.91	0.011	2.874	0.009	2.786	0.89
20	20	2.50	0.049	3.000	0.050	3.000	0.98	0.012	3.000	0.012	3.000	0.98
20	50	2.50	0.044	3.000	0.045	3.000	0.98	0.011	3.000	0.011	3.000	0.98
20	100	2.50	0.039	3.000	0.033	3.000	0.97	0.010	3.000	0.008	3.000	0.97
250	10	0.20	0.040	0.007	0.032	0.006	0.96	0.004	0.002	0.009	0.001	0.99
250	20	0.20	0.038	0.021	0.030	0.018	0.97	0.000	0.006	0.000	0.005	1.00
250	50	0.20	0.032	0.075	0.027	0.064	0.94	0.006	0.020	0.005	0.014	0.99
250	100	0.20	0.041	0.255	0.042	0.205	0.89	0.003	0.077	0.004	0.068	0.96
250	10	2.50	0.037	19.997	0.045	18.770	0.05	0.007	23.267	0.007	22.231	0.21
250	20	2.50	0.040	25.992	0.034	25.985	0.96	0.007	25.999	0.007	25.999	0.89
250	50	2.50	0.045	26.000	0.039	26.000	0.97	0.008	26.000	0.006	26.000	0.90
250	100	2.50	0.045	26.000	0.041	26.000	0.97	0.007	26.000	0.007	26.000	0.88
1000	10	0.20	0.035	0.010	0.041	0.013	0.96	0.005	0.002	0.000	0.000	0.99
1000	20	0.20	0.039	0.022	0.035	0.018	0.97	0.008	0.001	0.010	0.001	1.00
1000	50	0.20	0.044	0.097	0.026	0.074	0.93	0.008	0.023	0.003	0.014	0.98
1000	100	0.20	0.044	0.384	0.034	0.310	0.81	0.006	0.102	0.003	0.064	0.94
1000	10	2.50	0.029	58.116	0.033	48.002	0.01	0.005	88.164	0.005	84.028	0.01
1000	20	2.50	0.043	100.847	0.048	100.766	0.87	0.006	100.992	0.006	100.987	0.71
1000	50	2.50	0.042	101.000	0.044	101.000	0.97	0.006	101.000	0.005	101.000	0.71
1000	100	2.50	0.045	101.000	0.047	101.000	0.97	0.006	101.000	0.005	101.000	0.72

10% ([58-48]/M_1, where $M_1 = 100$) with $m = 1000$, $n = 10$ and $\delta = 2.5$. When the true data generating distribution is not normal, as expected the Mann-Whitney performs better than the t-test, with a difference in power that can be substantial.

Most important, irrespective of the error measure and true data distribution, the two tests can disagree, even almost never showing the same list of rejected hypotheses. This problem, which does not arise in the single inference situation, is more and more present as the number of tests m grows. As can be appreciated from Table 9, using the t-test with a symmetric but heavy tailed distribution leads to overly conservative control of the FWE. Again, this does not arise in the single inference situation.

As a final remark, it can be noted that use of the Benjamini-Yekutieli procedure for FDR control is conservative (the true Type I error rate is much lower than the nominal). This is well known. Still, the Benjamini-Yekutieli procedure is the only one that controls the FDR under arbitrary dependence for any finite sample size, and therefore the only one that can be blindly used in applications.

A real data example: MicroRNA profiling in human medulloblastoma

Medulloblastoma (MB) is the most frequent brain malignancy observed in childhood and originates from aberrant development of cerebellar progenitor neurons.

MB multimodal treatments (surgical resection,

Table 9. FWE, FDR, average number of correct rejections ($\bar{N}_{1|1}$) over the 0.1 m possible for Student's t-test and Mann-Whitney test, together with proportion of simulated data sets where they show agreement, for different m, n and δ under t_3 distributed data (nominal $\alpha = 0.05$).

			t-test		Mann-Whitney			t-test		Mann-Whitney		
m	n	δ	FWE	$\bar{N}_{1\|1}$	FWE	$\bar{N}_{1\|1}$	$I_{T=MW}$	FDR	$\bar{N}_{1\|1}$	FDR	$\bar{N}_{1\|1}$	$I_{T=MW}$
20	10	0.20	0.017	0.004	0.036	0.006	0.97	0.004	0.001	0.007	0.001	1.00
20	20	0.20	0.021	0.011	0.036	0.015	0.97	0.005	0.003	0.011	0.006	0.99
20	50	0.20	0.038	0.020	0.051	0.033	0.93	0.005	0.010	0.008	0.011	0.99
20	100	0.20	0.034	0.046	0.042	0.097	0.91	0.008	0.024	0.014	0.037	0.97
20	10	2.50	0.018	1.718	0.029	1.857	0.65	0.005	1.447	0.007	1.522	0.67
20	20	2.50	0.029	2.702	0.047	2.938	0.74	0.007	2.639	0.010	2.921	0.73
20	50	2.50	0.022	2.973	0.034	3.000	0.95	0.005	2.970	0.008	3.000	0.95
20	100	2.50	0.029	2.997	0.048	3.000	0.95	0.007	2.997	0.011	3.000	0.95
250	10	0.20	0.013	0.003	0.031	0.006	0.97	0.002	0.001	0.009	0.001	0.99
250	20	0.20	0.013	0.007	0.047	0.015	0.95	0.003	0.001	0.004	0.001	0.99
250	50	0.20	0.021	0.018	0.033	0.034	0.95	0.004	0.003	0.006	0.006	0.99
250	100	0.20	0.020	0.043	0.037	0.116	0.87	0.001	0.009	0.003	0.033	0.97
250	10	2.50	0.012	7.340	0.030	8.056	0.03	0.002	8.224	0.006	9.267	0.03
250	20	2.50	0.023	19.563	0.048	23.361	0.01	0.003	21.532	0.007	24.815	0.03
250	50	2.50	0.029	25.548	0.047	26.000	0.61	0.005	25.697	0.007	26.000	0.64
250	100	2.50	0.031	25.915	0.033	26.000	0.88	0.005	25.937	0.007	26.000	0.76
1000	10	0.20	0.009	0.003	0.038	0.012	0.96	0.000	0.001	0.000	0.001	1.00
1000	20	0.20	0.012	0.006	0.044	0.019	0.95	0.001	0.000	0.003	0.004	0.99
1000	50	0.20	0.017	0.017	0.029	0.035	0.95	0.004	0.004	0.004	0.007	0.99
1000	100	0.20	0.026	0.065	0.026	0.161	0.85	0.002	0.007	0.006	0.031	0.97
1000	10	2.50	0.007	16.569	0.038	19.060	0.01	0.001	27.298	0.005	32.371	0.01
1000	20	2.50	0.012	65.525	0.047	81.682	0.01	0.003	82.181	0.005	95.648	0.01
1000	50	2.50	0.023	98.256	0.041	100.999	0.06	0.004	99.671	0.005	101.000	0.17
1000	100	2.50	0.024	100.653	0.035	101.000	0.68	0.004	100.807	0.005	101.000	0.48

chemotherapy, and/or radiotherapy) have improved survival, however MB is still incurable in about a third of cases and survivors commonly have severe treatment-induced long-term side effects. The molecular aspects of tumorigenic pathways of MB are still poorly understood.

A study was recently conducted to identify specific microRNA (miRNA) signatures distinguishing tumours from normal tissues, which could be used to develop early detection and new risk-adapted therapeutic strategies based on molecular classification (21).

Surgical specimens of primary MBs were collected from n_1 = 34 patients with Institutional Review Board approval. A number of $n_2 = 14$ samples of normal human cerebellum were purchased from Biocat (Heidelberg, Germany), Ambion (Applied Biosystems, Foster City, CA) and BD Biosciences (San Jose, CA); thus, the total sample size was n = 48. Quantitative analysis of 250 miRNAs was performed on RNA samples using the specific stem-loop primers for reverse transcription (RT) followed by real-time PCR. The final measurements were log-transformed.

A number of m = 250 tests were then performed simultaneously, in which the two groups (tumour vs normal) were compared. We were interested in forming a list of the subset of the 250 miRNAs which are significantly differentially expressed.

Table 11 gives the number of genes selected by the t-test (R_T), by the Mann-Whitney (R_{MW}), the number of genes selected using the t-test that are not significant using Mann-Whitney (T_+) and the number genes selected by the Mann-Whitney that are not significant

Table 10. FWE, FDR, average number of correct rejections ($\bar{N}_{1|1}$) over the 0.1 m possible for Student's t-test and Mann-Whitney test, together with proportion of simulated data sets where they show agreement, for different m, n and δ under exponentially distributed data (nominal $\alpha = 0.05$).

			t-test		Mann-Whitney			t-test		Mann-Whitney		
m	n	δ	FWE	$\bar{N}_{1\|1}$	FWE	$\bar{N}_{1\|1}$	$I_{T=MW}$	FDR	$\bar{N}_{1\|1}$	FDR	$\bar{N}_{1\|1}$	$I_{T=MW}$
20	10	0.20	0.016	0.003	0.045	0.018	0.94	0.0010	0.001	0.0100	0.003	0.99
20	20	0.20	0.023	0.021	0.040	0.052	0.94	0.0045	0.006	0.0165	0.024	0.97
20	50	0.20	0.025	0.057	0.041	0.210	0.82	0.0080	0.016	0.0145	0.087	0.92
20	100	0.20	0.03	0.163	0.038	0.615	0.60	0.0070	0.091	0.0083	0.390	0.75
20	10	2.50	0.014	2.744	0.047	2.716	0.77	0.0024	2.668	0.0091	2.628	0.76
20	20	2.50	0.030	2.990	0.049	2.999	0.96	0.0077	2.989	0.0113	2.999	0.96
20	50	2.50	0.038	3.000	0.055	3.000	0.95	0.0092	3.000	0.0132	3.000	0.95
20	100	2.50	0.044	3.000	0.049	3.000	0.95	0.0103	3.000	0.0122	3.000	0.95
250	10	0.20	0.005	0.003	0.027	0.017	0.96	0.0030	0.000	0.0030	0.001	0.99
250	20	0.20	0.004	0.009	0.046	0.049	0.92	0.0000	0.001	0.0060	0.017	0.98
250	50	0.20	0.021	0.068	0.033	0.362	0.71	0.0020	0.013	0.0090	0.105	0.91
250	100	0.20	0.025	0.279	0.029	1.605	0.25	0.0060	0.073	0.0033	0.901	0.55
250	10	2.50	0.005	19.212	0.044	17.060	0.01	0.0015	21.744	0.0072	19.660	0.05
250	20	2.50	0.015	25.745	0.045	25.901	0.76	0.0028	25.900	0.0077	25.977	0.79
250	50	2.50	0.017	26.000	0.041	26.000	0.97	0.0040	26.000	0.0058	26.000	0.87
250	100	2.50	0.035	26.000	0.042	26.000	0.95	0.0060	26.000	0.0067	26.000	0.82
1000	10	0.20	0.003	0.003	0.042	0.015	0.94	0.000	0.001	0.000	0.000	1.00
1000	20	0.20	0.007	0.009	0.039	0.072	0.90	0.001	0.001	0.006	0.012	0.98
1000	50	0.20	0.024	0.066	0.037	0.438	0.65	0.003	0.010	0.005	0.102	0.92
1000	100	0.20	0.019	0.405	0.031	2.912	0.07	0.001	0.105	0.003	1.651	0.39
1000	10	2.50	0.012	61.228	0.043	54.553	0.01	0.001	83.004	0.006	75.089	0.01
1000	20	2.50	0.011	98.941	0.045	99.831	0.14	0.002	100.564	0.006	100.895	0.40
1000	50	2.50	0.015	101.000	0.036	101.000	0.96	0.004	101.000	0.005	101.000	0.62
1000	100	2.50	0.023	101.000	0.030	101.000	0.97	0.005	101.000	0.006	101.000	0.54

using the t-test (MW_+), controlling the FWE or FDR at different α levels.

Table 11. Number of genes selected using the t-test (R_T) and the Mann-Whitney (R_{MW}), genes selected by the t-test but not by Mann-Whitney (T_+), and the reverse (MW_+), for different Type I error measures at different α levels.

Type I	α	R_T	R_{MW}	T_+	MW_+
FWE	10%	73	86	6	19
FWE	5%	67	72	10	15
FWE	1%	51	55	12	16
FDR	10%	111	125	6	20
FDR	5%	96	114	5	23
FDR	1%	73	83	9	19

It can be seen that while there is a mild effect of the error measure on the number of tests rejected, there seems to be no effect on the overlap between the lists, which is disappointing. The t-test selects between 5 and 12 genes that are significant using the Mann-Whitney, and Mann-Whitney between 15 and 23 that are not selected using the t-test.

In light of our theoretical considerations and simulations, we claim that the Mann-Whitney test here gives more reliable conclusions, and the t-test may be less reliable. In order to support these claims, we focus on genes selected by controlling the FWE at level $\alpha = 0.01$. The list of genes selected by the t-test, with median, first and third quartile (respectively, Q_2, Q_1 and Q_3), mean μ and standard deviation σ for each group, is reported in Table 12. In the last column (U/D) it is indicated with a minus sign if the gene is down-regulated in tumour samples, and with a plus sign otherwise. Note that we compute these statistics

Table 12. Selected mRNA sequences using the *t*-test and controlling FWE at level $\alpha = 0.01$.

	Control group					Tumour group					
mRNA	Q_2	Q_1	Q_3	μ_C	σ_C	Q_2	Q_1	Q_3	μ_C	σ_C	U/D
miR127	0.68	0.56	0.96	0.79	0.44	0.03	0.02	0.11	0.09	0.13	–
miR103	0.66	0.57	0.93	0.73	0.24	0.13	0.06	0.25	0.18	0.16	–
hsalet7a	0.94	0.79	1.26	1.01	0.37	0.31	0.22	0.49	0.39	0.31	–
hsalet7b	0.43	0.31	0.64	0.51	0.26	0.13	0.06	0.21	0.14	0.09	–
hsalet7d	0.51	0.32	0.89	0.59	0.32	0.06	0.05	0.10	0.08	0.04	–
hsalet7e	0.51	0.37	0.91	0.62	0.30	0.09	0.05	0.15	0.12	0.10	–
hsalet7f	1.03	0.81	1.27	1.04	0.38	0.25	0.14	0.41	0.28	0.16	–
hsalet7g	0.97	0.74	1.17	0.92	0.32	0.20	0.13	0.31	0.23	0.12	–
haslet7i	0.92	0.73	1.05	0.90	0.27	0.11	0.07	0.18	0.12	0.07	–
miR107	0.75	0.60	1.15	0.95	0.55	0.12	0.06	0.16	0.13	0.08	–
miR124a	0.81	0.60	1.14	0.88	0.39	0.07	0.02	0.18	0.11	0.10	–
miR128a	1.11	0.99	1.40	1.26	0.42	0.05	0.01	0.11	0.11	0.17	–
miR128b	1.01	0.83	1.31	1.10	0.30	0.04	0.01	0.12	0.15	0.27	–
miR132	1.41	1.07	1.52	1.33	0.33	0.19	0.09	0.31	0.22	0.17	–
miR133b	1.48	0.91	2.09	1.57	0.84	0.09	0.02	0.15	0.23	0.47	–
miR134	0.81	0.68	0.93	0.77	0.21	0.04	0.02	0.11	0.10	0.14	–
miR138	2.01	1.11	2.81	2.07	1.13	0.06	0.03	0.19	0.21	0.37	–
miR143	1.22	1.01	1.35	1.21	0.30	0.43	0.27	0.75	0.55	0.40	–
miR149	0.83	0.75	0.99	0.82	0.19	0.11	0.05	0.27	0.25	0.32	–
miR150	0.62	0.48	0.92	0.68	0.27	0.15	0.08	0.32	0.22	0.20	–
miR154	0.99	0.60	1.33	0.95	0.46	0.05	0.03	0.13	0.10	0.14	–
miR184	1.10	0.97	1.33	1.21	0.42	0.01	0.01	0.04	0.03	0.03	–
miR190	1.15	0.94	1.56	1.25	0.45	0.18	0.07	0.50	0.35	0.47	–
miR191	1.12	0.94	1.32	1.14	0.28	0.39	0.23	0.84	0.55	0.51	–
miR212	1.01	0.72	1.06	0.99	0.36	0.23	0.14	0.36	0.30	0.23	–
miR22	1.05	0.91	1.64	1.25	0.52	0.35	0.19	0.48	0.46	0.44	–
miR28	1.44	1.11	1.58	1.34	0.35	0.54	0.20	0.68	0.51	0.36	–
miR30a3p	1.69	1.42	1.89	1.67	0.45	0.19	0.08	0.40	0.25	0.21	–
miR30b	0.96	0.50	1.12	0.93	0.46	0.33	0.16	0.45	0.33	0.21	–
miR30c	1.02	0.63	1.30	1.01	0.40	0.38	0.20	0.55	0.38	0.24	–
miR320	0.67	0.59	0.79	0.73	0.24	0.16	0.10	0.29	0.25	0.24	–
miR323	1.01	0.59	1.46	1.05	0.50	0.03	0.01	0.11	0.10	0.16	–
miR3243p	0.77	0.51	0.99	0.74	0.28	0.16	0.09	0.29	0.21	0.16	–
miR326	0.57	0.38	0.82	0.61	0.29	0.06	0.03	0.13	0.10	0.10	–
miR330	0.62	0.22	0.95	0.60	0.37	0.03	0.02	0.06	0.05	0.06	–
miR331	0.69	0.48	0.91	0.71	0.32	0.17	0.10	0.25	0.22	0.20	–
miR337	1.12	1.09	1.18	1.15	0.12	0.08	0.01	0.17	0.16	0.25	–
miR346	0.32	0.29	0.76	0.50	0.32	0.04	0.03	0.07	0.07	0.10	–
miR3693p	1.20	0.95	1.51	1.30	0.43	0.09	0.01	0.22	0.18	0.27	–
miR3695p	1.23	0.98	1.36	1.13	0.35	0.11	0.02	0.23	0.19	0.30	–
miR370	0.38	0.29	0.78	0.54	0.31	0.05	0.03	0.07	0.06	0.06	–
miR376a	1.06	0.90	1.35	1.13	0.32	0.15	0.04	0.31	0.22	0.24	–
miR3803p	1.26	0.96	1.54	1.26	0.47	0.06	0.02	0.11	0.16	0.31	–
miR382	0.53	0.39	0.96	0.64	0.31	0.04	0.02	0.10	0.08	0.11	–
miR383	0.34	0.25	0.59	0.46	0.32	0.00	0.00	0.02	0.02	0.04	–
miR425	0.54	0.46	0.87	0.66	0.27	0.23	0.15	0.35	0.29	0.23	–
miR494	1.07	0.82	1.18	1.01	0.35	0.11	0.05	0.23	0.17	0.17	–
miR95	0.67	0.61	0.88	0.73	0.19	0.07	0.02	0.13	0.10	0.10	–
miR661	1.06	0.56	1.27	0.96	0.46	0.19	0.13	0.31	0.25	0.24	–
miR18a	1.18	0.58	2.27	1.94	1.82	6.63	2.39	12.02	7.65	6.01	+
miR301	1.00	0.80	1.72	1.21	0.62	8.96	3.89	15.93	25.09	51.84	+

Table 13. Selected mRNA sequences using the Mann-Whitney test and controlling FWE at level $\alpha = 0.01$.

	Control group					Tumour group					
mRNA	Q_2	Q_1	Q_3	μ_C	σ_C	Q_2	Q_1	Q_3	μ_C	σ_C	U/D
miR26a	0.65	0.25	1.08	0.68	0.42	0.13	0.10	0.19	0.15	0.11	–
miR26b	0.60	0.49	1.00	0.73	0.37	0.24	0.13	0.38	0.27	0.16	–
miR127	0.68	0.56	0.96	0.79	0.44	0.03	0.02	0.11	0.09	0.13	–
miR103	0.66	0.57	0.93	0.73	0.24	0.13	0.06	0.25	0.18	0.16	–
hsalet7a	0.94	0.79	1.26	1.01	0.37	0.31	0.22	0.49	0.39	0.31	–
hsalet7b	0.43	0.31	0.64	0.51	0.26	0.13	0.06	0.21	0.14	0.09	–
hsalet7d	0.51	0.32	0.89	0.59	0.32	0.06	0.05	0.10	0.08	0.04	–
hsalet7e	0.51	0.37	0.91	0.62	0.30	0.09	0.05	0.15	0.12	0.10	–
hsalet7f	1.03	0.81	1.27	1.04	0.38	0.25	0.14	0.41	0.28	0.16	–
hsalet7g	0.97	0.74	1.17	0.92	0.32	0.20	0.13	0.31	0.23	0.12	–
haslet7i	0.92	0.73	1.05	0.90	0.27	0.11	0.07	0.18	0.12	0.07	–
miR107	0.75	0.60	1.15	0.95	0.55	0.12	0.06	0.16	0.13	0.08	–
miR124a	0.81	0.60	1.14	0.88	0.39	0.07	0.02	0.18	0.11	0.10	–
miR125a	0.53	0.39	0.87	0.70	0.44	0.14	0.10	0.21	0.17	0.11	–
miR128a	1.11	0.99	1.40	1.26	0.42	0.05	0.01	0.11	0.11	0.17	–
miR128b	1.01	0.83	1.31	1.10	0.30	0.04	0.01	0.12	0.15	0.27	–
miR129	0.83	0.17	1.16	0.72	0.53	0.01	0.00	0.06	0.08	0.17	–
miR132	1.41	1.07	1.52	1.33	0.33	0.19	0.09	0.31	0.22	0.17	–
miR133b	1.48	0.91	2.09	1.57	0.84	0.09	0.02	0.15	0.23	0.47	–
miR134	0.81	0.68	0.93	0.77	0.21	0.04	0.02	0.11	0.10	0.14	–
miR138	2.01	1.11	2.81	2.07	1.13	0.06	0.03	0.19	0.21	0.37	–
miR143	1.22	1.01	1.35	1.21	0.30	0.43	0.27	0.75	0.55	0.40	–
miR149	0.83	0.75	0.99	0.82	0.19	0.11	0.05	0.27	0.25	0.32	–
miR150	0.62	0.48	0.92	0.68	0.27	0.15	0.08	0.32	0.22	0.20	–
miR151	1.17	1.08	1.47	1.30	0.41	0.39	0.33	0.54	0.41	0.18	–
miR154	0.99	0.60	1.33	0.95	0.46	0.05	0.03	0.13	0.10	0.14	–
miR190	1.15	0.94	1.56	1.25	0.45	0.18	0.07	0.50	0.35	0.47	–
miR192	0.89	0.39	0.99	0.84	0.55	0.19	0.08	0.27	0.25	0.27	–
miR194	0.39	0.31	0.93	0.66	0.56	0.11	0.05	0.20	0.14	0.13	–
miR212	1.01	0.72	1.06	0.99	0.36	0.23	0.14	0.36	0.30	0.23	–
miR219	0.37	0.14	0.62	0.54	0.56	0.04	0.01	0.09	0.06	0.07	–
miR22	1.05	0.91	1.64	1.25	0.52	0.35	0.19	0.48	0.46	0.44	–
miR2995p	0.47	0.34	1.10	0.72	0.47	0.03	0.01	0.09	0.06	0.08	–
miR29a	0.53	0.13	0.77	0.53	0.41	0.04	0.03	0.11	0.08	0.10	–
miR30a3p	1.69	1.42	1.89	1.67	0.45	0.19	0.08	0.40	0.25	0.21	–
miR30b	0.96	0.50	1.12	0.93	0.46	0.33	0.16	0.45	0.33	0.21	–
miR30c	1.02	0.63	1.30	1.01	0.40	0.38	0.20	0.55	0.38	0.24	–
miR320	0.67	0.59	0.79	0.73	0.24	0.16	0.10	0.29	0.25	0.24	–
miR323	1.01	0.59	1.46	1.05	0.50	0.03	0.01	0.11	0.10	0.16	–
miR3243p	0.77	0.51	0.99	0.74	0.28	0.16	0.09	0.29	0.21	0.16	–
miR3245p	0.84	0.36	1.30	0.84	0.48	0.16	0.07	0.24	0.20	0.17	–
miR326	0.57	0.38	0.82	0.61	0.29	0.06	0.03	0.13	0.10	0.10	–
miR328	0.33	0.28	0.76	0.49	0.32	0.04	0.02	0.07	0.06	0.06	–
miR330	0.62	0.22	0.95	0.60	0.37	0.03	0.02	0.06	0.05	0.06	–
miR331	0.69	0.48	0.91	0.71	0.32	0.17	0.10	0.25	0.22	0.20	–
miR346	0.32	0.29	0.76	0.50	0.32	0.04	0.03	0.07	0.07	0.10	–
miR370	0.38	0.29	0.78	0.54	0.31	0.05	0.03	0.07	0.06	0.06	–
miR381	0.64	0.20	0.96	0.58	0.37	0.05	0.03	0.13	0.14	0.22	–
miR382	0.53	0.39	0.96	0.64	0.31	0.04	0.02	0.10	0.08	0.11	–
miR383	0.34	0.25	0.59	0.46	0.32	0.00	0.00	0.02	0.02	0.04	–
miR425	0.54	0.46	0.87	0.66	0.27	0.23	0.15	0.35	0.29	0.23	–

on the original and not on the log scale, which was used instead for testing. The list of genes selected by the Mann-Whitney test is shown in Table 13.

Due to the use of the FWE and low α, we should be very confident about the selected genes. However, there emerge some contradictions.

In particular, controlling FWE at level $\alpha = 0.01$, the two tests disagree about miR-125a, in that with the *t*-test no significance is declared, whereas with the Mann-Whitney there seems to be differential expression with regard to that miR. This is particularly interesting since miR-125a has been biologically validated, and we thus have some post screening evidence of the fact that, for this gene, the Mann-Whitney test gives a more reliable result. This result, on a single gene, cannot be used to validate the Mann-Whitney over the *t*-test in general, but it can certainly be taken as mild evidence from a real data application.

The implications of our findings about MB are that most miRNAs display overall down-regulated expression, suggesting a tumour suppression function. This is another feature supporting the use of the Mann-Whitney test in this application. The down-regulation of MB tumour samples could, in fact, be expected before the experiment, and while the list in Table 13 shows only down-regulated genes, Table 12 contains two up-regulated genes (miR18a and miR301), which could be false discoveries due to bias in assuming normality even after log-transformation.

Ferretti et al. conclude (21) are that an altered expression of microRNAs controlling granular cell differentiation events might be involved in cerebellum tumorigenesis. This property has been validated for miR-125a whose rescued expression might inhibit the proliferation of MB cells.

Conclusions

Our main conclusion is that while in single inference, under normality or small departures from it, choice between parametric and nonparametric testing may not be particularly crucial, in that the same conclusion will often be achieved, in multiple testing wild differences can emerge. Multiple testing procedures take into account all the *p*-values, with the result that even small differences may lead to very different conclusions. This is exacerbated under non-normality, when the *t*-test may be biased. The only apparent drawback of using rank-based nonparametric tests is that they never reject when the sample size is very small (Table 2).

George Box used to say: “all the models are wrong (but some are useful)”. We draw support from his remark to claim that tests for location based on normality (which Box would call a wrong model) should not be undertaken lightly in the frequent small-sample many-tests situation. Our final recommendation is always to use distribution-free tests in such cases.

References

1. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics. 1947; 18: 50-60.
2. Conover WJ. Practical Nonparametric Statistics. New York: Wiley, 1998.
3. Mazzaferro S, Pasquali M, Farcomeni A, Vestri AR, Filippini A, Romani AM, Barresi G, Pugliese F. Parathyroidectomy as a therapeutic tool for targeting the recommended NKF-K/DOQI ranges for serum calcium, phosphate and parathyroid hormone in dialysis patients. Nephrology Dialysis Transplantation 2008; Advance Access published on February 16, 2008.
4. Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 1996; 4: 58-71.
5. Ellis SP, Underwood MD, Arango V. Mixed models and multiple comparisons in analysis of human neurochemical maps. Psychiatry Res 2000; 9: 111-119.
6. Merriam EP, Genovese CR, Colby CL. Spatial updating in human parietal cortex. Neuron 2003; 39: 361-373.
7. Logan BR, Rowe DB. An evaluation of thresholding techniques in fMRI analysis. Neuroimage 2004; 22: 95-108.
8. Drigalenko EI, Elston RC. False discoveries in genome scanning. Genet Epidemiol 1997; 14: 779-784.
9. Weller JI, Song JZ, Heyen DW, Lewin HA, Ron M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics. 1998; 150: 1699-1706.
10. Heyen DW, Weller JI, Ron M, Band M, Beever JE, Feldmesser E, Da Y, Wiggans GR, VanRaden PM, Lewin HA. A genome scan for QTL influencing milk production and health traits in dairy cattle. Physiol Genomics 1999; 1: 165-175.
11. Bovenhuis H, Spelman RJ. Selective genotyping to de-

tect quantitative trait loci for multiple traits in outbred populations. J Dairy Sci 2000; 83: 173-180.

12. Mosig MO, Lipkin E, Khutoreskaya G, Tchourzyna E, Soller M, Fridmann A. A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics 2001; 157: 1683-1698.
13. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19: 368-375.
14. Vedantham K, Brunet A, Boyer R, Weiss DS, Metzler TJ, Marmar CR. Post-traumatic stress disorder, trauma exposure, and the current health of Canadian bus drivers. Can J Psychiatry 2001; 46: 149-155.
15. Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. Am J Epidemiol 1998; 147: 615-619.
16. Schlaeppi M, Edwards K, Fuller RW, Sharma R. Patient perception of the diskus inhaler: a comparison with the turbuhaler inhaler. Br J Clin Pract 1996; 50: 14-19.
17. Farcomeni A. A review of modern multiple hypothesis testing with particular attention to the false discovery proportion. Stat Methods Med Res 2007 [Epub ahead of print]
18. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 1979; 6: 65-70.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society (Ser B) 1995; 57: 289-300.
20. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 2001; 29: 1165-1188.
21. Ferretti E, De Smaele E, Po A, Di Marcotullio L, Tosi E, Espinola MSB, Di Rocco C, Riccardi R, Giangaspero F, Farcomeni A, Nofroni I, Laneve P, Gioia U, Caffarelli E, Bozzoni I, Screpanti I, Gulino A. MicroRNA profiling in human medulloblastoma. Currently under submission 2008.