Published in final edited form as:

Methods Mol Biol. 2010 ; 620: 105–153. doi:10.1007/978-1-60761-580-4_2.

Nonparametric Methods in Molecular Biology

Knut M. Wittkowski and Tingting Song

The Rockefeller University, Center for Clinical and Translational Science, 1230 York Ave Box 322, New York, NY 10021, U.S.A.

Knut M. Wittkowski: kmw@rockefeller.edu

Abstract

In 2003, the completion of the Human Genome Project[1] together with advances in computational resources[2] were expected to launch an era where the genetic and genomic contributions to many common diseases would be found. In the years following, however, researchers became increasingly frustrated as most reported ‘findings’ could not be replicated in independent studies[3]. To improve the signal/noise ratio, it was suggested to increase the number of cases to be included to tens of thousands[4], a requirement that would dramatically restrict the scope of personalized medicine. Similarly, there was little success in elucidating the gene–gene interactions involved in complex diseases or even in developing criteria for assessing their phenotypes. As a partial solution to these enigmas, we here introduce a class of statistical methods as the ‘missing link’ between advances in genetics and informatics. As a first step, we provide a unifying view of a plethora of non-parametric tests developed mainly in the 1940s, all of which can be expressed as u-statistics. Then, we will extend this approach to reflect categorical and ordinal relationships between variables, resulting in a flexible and powerful approach to deal with the impact of (1) multi-allelic genetic loci, (2) poly-locus genetic regions, and (3) oligo-genetic and oligo-genomic collaborative interactions on complex phenotypes.

Keywords

collaboration; coregulation; epistasis; Family-based Association Test (FBAT); gene-gene interaction; Genome-wide Association Study (GWAS); High Density Oligo-Nucleotide Assay (HDONA); multi-allelic; multi-locus; multi-variate; personalized medicine

1. Introduction

As functional genetics and genomics advance and prices for sequencing and expression profiling drop, new possibilities arise, but also new challenges. The initial successes in identifying the causes of rare, mono-causal diseases serve as a proof-of-concept that new diagnostics and therapies can be developed. Common diseases have even more impact on public health, but as they typically involve genetic epistasis, genomic pathways, and proteomic patterns, new requirements for database systems and statistical analysis tools are necessitated.

Biological systems are regulated by various, often unknown feedback loops so that the functional form of relationships between measurement and activity or efficacy is typically unknown, except within narrowly controlled experiments. Still, many statistical methods are based on the linear model[5], i.e., on the assumption that the above (unknown) relationship is linear. The linear model has the advantage of computational efficiency. Moreover, assuming independence and additivity yields conveniently bell shaped distributions and parameters of alluring simplicity. The prayer that biology be linear, independent, and additive, however, is rarely answered and the Central Limit Theorem (CLT) neither applies to small samples nor rescues from model misspecification.

When John Arbuthnot (1667–1735) argued that the chance of more male than female babies being born in London for the last 82 consecutive years was only $1/2^{82}$ “*if mere Chance govern’d*”, and, thus, “*it is Art, not Chance, that governs*”, [6] he was arguably the first to ever apply the concept of hypothesis testing to obtain what is now known as a ‘p-value’ and, interestingly, he applied it to a problem in molecular biology.

The test, now known as the sign or McNemar[7] test (see below), belongs to the class of ‘non-parametric’ tests, which differ from their ‘parametric’ counterparts, in that the distribution of the data is not assumed to be known, except for a single parameter to be estimated. As they require fewer unjustifiable assumptions to be made, non-parametric methods are more adequate for biological systems, in general. Moreover, they tend to be easier to understand. The median, for instance, is easily explained as the cut-off where as many observations are above as below, and the interquartile range as the range with 25% of the data both below and above. In contrast to the mean and standard deviation, these ‘quartiles’ do not change with (monotonous) scale transformations (such as log and inverse), are robust to outliers, and often reflect more closely to the goals of the investigator[8]: ‘Do people in this group *tend* to score higher than people in that?’ ‘Is the order on this variable *similar to the order* on that?’ If the questions are ordinal, it seems preferable to use ordinal methods to answer them[9].

The reason for methods based on the linear model, ranging from the t-test(s) and analysis of variance (ANOVA) to stepwise linear regression and factor analysis is not their biological plausibility, but merely their computational efficiency. While mean and standard deviation are easily computed with a pocket calculator, quartiles are not. Other disadvantages of ordinal methods are the relative scarcity of experimental designs that can be analyzed. Moreover, non-parametric methods are often presented as a confusing ‘hodgepodge’ of seemingly unrelated methods. On the one hand, equivalent methods can have the different names, such as the Wilcoxon (rank-sum)[10], the Mann–Whitney (u)[11], and the Kruskal–Wallis[12] test (when applied to two groups). On the other hand, substantially different tests may be attributed to the same author, such as Wilcoxon’s rank-sum and signed-rank tests[10].

In the field of molecular biology, the need for computational (bioinformatics) approaches in response to a rapidly evolving technology producing large amounts of data from small samples has created a new field of “in silico” analyses, the term a hybrid of the Latin *in silice* (from *silix*, *silicis* m.: hard stone[13], 14) and phases such as *in vivo*, *in vitro*, and *in utero*, which seem to be better known than, for instance, *in situ*, *in perpetuum*, *in memoriam*, *in nubibus*, and *in capite*. Many in silico methods have added even more confusion to the field. For instance, a version of the t-test with a more conservative variance estimate is often referred to as SAM (significance analysis for microarrays)[15], and adding a similar “fudge factor” to the Wilcoxon signed rank test yields SAM-RS[16]. Referring to ‘new’ in silico approaches by names intended to highlight the particular application or manufacturer, rather than the underlying concepts has often prevented these methods from being thoroughly evaluated. With sample sizes increasing and novel technologies (e.g., RNA-seq) replacing less robust technologies (e.g., microarrays)[17] that often relied on empirically motivated standardization and normalization, the focus can now shift from *ad hoc* bioinformatics approaches to well-founded biostatistical concepts. Below we will present a comprehensive theory for a wide range of non-parametric *in silice* methods based on recent advances in understanding the underlying fundamentals, novel methodological developments, and improved algorithms.

2. Univariate Unstructured Data

2.1. Independence

2.1.1. Introduction—For single-sample designs, independence of the observations constituting the sample is a key principle to guarantee when applying any statistical test to

observed data. All asymptotic (large-sample) versions of the tests discussed here are based on the CLT, which applies when (i) *many* (ii) *independent* observations contribute in an (iii) *additive* fashion to the test statistic. Of these three requirements, additivity is typically fulfilled by the way the test statistic is formed, which may be, for instance, based on the sum of the data, often after rank or log transformation. Independence, then, applies to how the data are being aggregated into the test statistic. Often, one will allow only a single child per family to be included, or at least only one of identical twins. Finally, the rule that ‘more is better’ underlies the Law of Large Numbers, which states that the accuracy of the estimates increases with the square root of the number of independent samples included.

Finally, one will try to minimize the effect of unwanted ‘confounders’. For instance, when one compares the effects of two interventions within the same subject, one would typically aim at a ‘cross-over’ design, where the two interventions are applied in a random order to minimize ‘carry-over’ effects, where the first intervention’s effects might affect the results observed under the subsequent intervention.

Another strategy is ‘stratification’, where subjects are analyzed as sub-samples forming ‘blocks’ of subjects that are comparable with respect to the confounding factor(s), such as genetic factors (e.g., race and ethnicity), exposure to environmental factors (smoking) or experimental conditions.

A third strategy would be to ‘model’ the confounding factor, e.g., by subtracting or dividing by its presumed effect on the outcome. As the focus of this chapter is on non-parametric (aka ‘model-free’) methods, this strategy will be used only sparingly, i.e., when the form of the functional relationship among the confounding and observed variables is, in fact, known with certainty.

2.1.2. The sign test for untied observations—The so-called sign test is the most simple of all statistical methods, applying to a single sample of binary outcomes from independent observations, which can be discrete, such as the sexes ‘M’ or ‘F’ in Arbuthnot’s case[6], ‘correct vs. false’, such as R.A. Fisher’s famous (Lady Tasting) Tea test[7], where a lady claims to be able to decide whether milk was added to tea, or *vice versa*. When applied to paired observations (‘increase’ vs. ‘decrease’), the sign test is often named after McNemar[18]. (Here, we assume that all sexes, answers, and changes can be unambiguously determined. The case of ‘ties’ will be discussed in Section 2.1.3 below).

When the observations are independent, the number of positive ‘signs’ follows a binomial distribution, so that the probability of results deviating at least as much as the result observed from the result expected under the null hypothesis H_0 (the ‘p-value’) is easily obtained.

Let X be the sum of positive signs among n replications, each having a ‘success’ probability of p (e.g., $p = 1/2$). Then, the probability of having k ‘successes’ (see Fig. 2.1) is

$$b_{n,p}(k) = P(X=k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

The (one-sided) p-value is easily computed as $\sum_{x=k}^n b_{n,p}(x)$. Let $x_o = \frac{1}{n} \sum_{i=1}^n x_i$, then, for binomial random variables, the parameter estimate $X/n = \hat{p}$ and its standard deviation

$$s_n(\hat{p}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - x_o)^2} = \sqrt{\frac{1}{n} \sqrt{\hat{p}(1-\hat{p})^2 + (1-\hat{p})(0-\hat{p})^2}} = \sqrt{\frac{1}{n} \sqrt{\hat{p}(1-\hat{p})}}$$

are also easily derived. More than 100 years after Arbuthnot[6], Gauss[19] proved the CLT, by which the binomial distribution can be approximated asymptotically (as.) with the familiar Gaussian ‘normal’ distribution

$$\sqrt{n} \frac{(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \sim_{\text{as.}} N(0, 1), \text{ i.e., } P \left(\sqrt{n} \frac{(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} > z \right) \xrightarrow{n \rightarrow \infty} \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

The former (exact) form of the sign test can be derived and applied using only elementary mathematical operations, while the latter, like the parametric equivalent, requires theoretical results that most statisticians find challenging to derive, including an integral without a closed form solution. For $p_0 = 1/2$ McNemar[18] noted that this yields a (two-sided) test statistic in a particularly simple form

$$\frac{(N_+ - N_-)^2}{N_+ + N_-} \sim_{\text{as.}} \chi_1^2$$

2.1.3. Sign tests for tied observations—In reality, the situations where the sign test is to be applied are often more complicated, because the sign of an observation (or pair of observations) may be neither positive nor negative. When such ‘ties’ are present, the McNemar test[18] yields the same results for two distinctively different situations:

- either we may have nine positive and one negative observations out of a total of ten,
- or we have also nine positive and one negative observation, but also 9990 tied observations. A ‘tie’ might be a subject being heterozygous or, in Arbuthnot’s case, having gonadal mosaicism, the rare conditions, where a subject has cells with both XX and XY chromosomes.

Ignoring ties is often referred to as ‘correction for ties’. Still, one may feel uncomfortable with a result being ‘significant’, although only nine observations ‘improved’ and one even ‘worsened’, while 9,990 were (more or less) ‘unchanged’.

This observation has long puzzled statisticians. In fact, several versions of ‘the’ sign test have been proposed[20]. Originally, Dixon, in 1946[21], suggested that ties be split equally between the positive and negative signs, but in 1951[22] followed McNemar[18] in suggesting that ties be dropped from the analysis.

To explicate the discussion, we divide the estimated proportion of positive signs by their standard deviation to achieve a common asymptotic distribution. Let N_- , N_+ , and N_0 denote the number of negative, positive, and tied observations, respectively, among a total of n . The original sign test[21] can then be written as

$$T^* = \frac{(N_+ + p_0 N_0) - p_0 n}{\sqrt{p_0(1-p_0)n}} = \frac{(N_+ + p_0 N_0)/n - p_0}{\sqrt{p_0(1-p_0)/n}} \Bigg|_{\text{if } p_0 = 1/2} = \frac{N_+ - N_-}{\sqrt{n}} \sim_{\text{as.}} N(0, 1)$$

and the alternative[18,22] as

$$T = \frac{(N_+ + p_0 N_-) - p_H n}{\sqrt{p_0(1-p_0)(n-N_-)}} = \frac{N_+/(n-N_-) - p_0}{\sqrt{p_0(1-p_0)/(n-N_-)}} \Bigg|_{p_0=1/2} = \frac{N_+ - N_-}{\sqrt{N_+ + N_-}} \sim_{as.} N(0, 1)$$

The first and last term in the above series of equations show that the ‘correction for ties’ increases the test statistic by a factor of $\sqrt{n/(n-N_-)}$, thereby yielding more ‘significant’ results. The center term is helpful in understanding the nature of the difference. T^* distributes the ties as the numbers of observations expected under H_0 and, then uses the unconditional variance[21]. T excludes the ties from the analysis altogether, thereby ‘conditioning’ the test statistic.

The sign test is often applied when the distribution of the data (or differences of data) cannot be assumed to be symmetric. (Otherwise, the paired t-test would be a robust alternative[5].) Assume, for simplicity of the argument, that the differences follow a triangular distribution with a median of zero and a symmetric discretization interval with (band-) width $2b$ around the origin (Fig. 2.2).

Then, with T , ‘significance’ increases with the discretization bandwidth, i.e., with the inaccuracy of the measurements; the test statistic increases from 0.00 to 2.34 (Table 2.1). With T^* , in contrast, the estimate for the probability of a positive sign remains within a narrow limit between 0.4 and 0.5 and the test statistic never exceeds the value of 1.0.

To resolve the seeming discrepancy between theoretical results suggesting a specific treatment of ties as ‘optimal’ and the counterintuitive consequences of using this strategy, one needs to consider the nature of ties. In genetics, for instance, tied observations can often be assumed to indicate identical phenomena (e.g., mutation present or absent[23]). Often, however, a thorough formulation of the problem refers to an underlying continuous or unmeasurable factor, rather than the observed discretized variable. For instance, ties may be due to rounding of continuous variables (temperature, Fig. 2.2) or to the use of discrete surrogate variables for continuous phenomena (parity for fertility), in which case the unconditional sign test should be used. Of course, when other assumptions can be reasonably made, such as the existence of a relevance threshold[24], or a linear model for the paired comparison preference profiles [25], ties could be treated in many other ways[26].

2.2. Applications

2.2.1. Family based association studies—Family-based association tests (FBAT) control for spurious associations between disease and specific marker alleles due to population stratification[27]. Thus, the transmission/disequilibrium test (TDT) for biallelic markers, proposed in 1993 by Spielman *et al.*[28] has become one of the most frequently used statistical methods in genetics. Part of its appeal stems from its computationally simple form $(b-c)^2/(b+c)$, which resembles the conditional sign test (Section 2.1.3). Here, b and c represent the number of transmitted wild-type (P) and risk (Q) alleles, respectively, whose parental origin can be identified in affected children.

Let n_{XY} denote the number of affected XY children and let $XY \sim X'Y'$ denote a parental mating type ($X, Y, X', Y' \in \{P, Q\}$; the order of genotypes and parental genders is arbitrary). Alleles transmitted from homozygous parents are non-informative, so that children with two homozygous parents are excluded from the analysis as ‘exact ties’[29,30]. If one parent is homozygous, only the allele transmitted from the other parent is informative. When both parents are heterozygous, the number of PP and QQ children, n_{PP} and n_{QQ} , respectively,

represents both the number of P or Q alleles transmitted from either parent; n_{PQ} represents the number of P alleles transmitted from one and the number of Q alleles transmitted from the other parent, yet the origin of the alleles is not known.

To compare “the frequency with which [among heterozygous parents] the associated allele P or its alternate Q is transmitted to the affected offspring”[28], the term $b - c$ in the numerator of the TDT can be decomposed into the contributions from families stratified by the parental mating types

$$b - c = [n_{PQ} - n_{QQ}]_{PQ-QQ} + [2(n_{PP} - n_{QQ}) + (n_{PQ} - n_{PQ})]_{PQ-PQ} + [n_{PP} - n_{PQ}]_{PP-PQ}$$

Of course $[n_{PQ} - n_{PQ}]_{PQ-PQ} = 0$, so that this equation can be rewritten as[31]

$$b - c = [n_{PQ} - n_{QQ}]_{PQ-QQ} + 2[n_{PP} - n_{QQ}]_{PQ-PQ} + [n_{PP} - n_{PQ}]_{PP-PQ}. \quad (2.1)$$

As an ‘exact tie’[29,30], the term $[n_{PQ} - n_{PQ}]_{PQ-PQ}$ can be as ignored when computing the variance of (2.1), because they are as non-informative as the alleles transmitted from homozygous parents.

As noted above, independence of observations is a key concept in building test statistics. While “the contributions from both parents are independent”[28], the observations are not. Because the effects of the two alleles transmitted to the same child are subject to the same genetic and environmental confounders, one does not have a sample of independently observed alleles to build the test statistic on, but “a sample of n affected children”[28]. As a consequence, the factor ‘2’ in (2.1) does not increase the sample size[32], but is a weight indicating that the PP and the QQ children are ‘two alleles apart’, which implicates a larger risk difference under the assumption of co-dominance. Each of the three components in (2.1) follows a binomial distribution with probability of success equal to $1/2$, so that the variance of (2.1) follows from “the standard approximation to a binomial test of the equality of two proportions”[28]

$$\sigma_0^2(b - c) = [n_{PQ} + n_{QQ}]_{PQ-QQ} + 4[n_{PP} - n_{QQ}]_{PQ-PQ} + [n_{PP} - n_{PQ}]_{PP-PQ}. \quad (2.2)$$

Dividing the square of estimate (2.1) by its variance (2.2) yields a stratified McNemar test (SMN) which combines the estimates and variances of three McNemar tests in a fashion typical for stratified tests [33].

$$\frac{(b - c)^2}{\sigma_0^2(b - c)} = \frac{\left([n_{PQ} - n_{QQ}]_{PQ-QQ} + 2[n_{PP} - n_{QQ}]_{PQ-PQ} + [n_{PP} - n_{PQ}]_{PP-PQ}\right)^2}{[n_{PQ} - n_{QQ}]_{PQ-QQ} + 4[n_{PP} - n_{QQ}]_{PQ-PQ} + [n_{PP} - n_{PQ}]_{PP-PQ}} \sim as \chi_1^2 \quad (2.3)$$

(‘Stratification’, i.e., blocking children by parental mating type, here is merely a matter of computational convenience. Formally, one could treat each child as a separate block, with the same results.) The TDT, in contrast, divides the same numerator by a “variance estimate”[28]

$$\hat{\sigma}_0^2(b-c) = b+c = \left[n_{PQ} + n_{QQ} \right]_{PQ \sim QQ} + \left\{ \begin{array}{c} 2 \left[n_{PP} + n_{QQ} \right]_{PQ \sim PQ} \\ + 2 \left[n_{PQ} \right]_{PQ \sim PQ} \end{array} \right\} + \left[n_{PP} + n_{PQ} \right]_{PP \sim PQ} \quad (2.4)$$

which relies on the counts of non-informative PQ children. Replacing the observed $n_{PP}+n_{QQ}$ by its estimate n_{PQ} under H_0 would require an adjustment, similar to the replacement of the Gaussian by the t-distribution when using the empirical standard deviation with the t-test[34]. Hence the SMN has advantages over the TDT for finite samples[31], in part, because it has a smaller variance (more and, thus, lower steps, see Fig. 2.3) under H_0 (see Table 2.2).

Because the TDT estimates the variance under the assumption of co-dominance, it overestimates the variance (has lower power) for alleles with high dominance, i.e., when being heterozygous for the risk allele carries (almost) as much risk as being homozygous[31]. On the other hand, the TDT underestimates the variance (yielding higher power) for recessive alleles[31].

2.2.2. Extensions of the Stratified McNemar—Understanding the role of parental mating types, the sensitivity of the SMN can be easily focused on either dominant or recessive alleles. When screening for recessive alleles, one excludes trios where one parent is homozygous for the wild type and assigns equal weights to the remaining trios[35]. Conversely, when screening for dominant alleles, one excludes trios where one parent is homozygous for putative risk allele.

A better understanding of the statistical principles underlying family-based associations studies not only leads to test statistics with better small sample properties and better targeted tests for alleles with low or high dominance but also to extensions which open new areas of applications.

For decades it has been known that the HLA-DRB1 gene is a major factor in determining the risk for multiple sclerosis[36]. As HLA-DRB1 is one of the few genes where more than two alleles per locus are routinely observed, the SMN is easily generalized to multi-allelic loci by increasing the number of parental mating-type strata and identifying within each stratum the number of informative children[35]. In early 2009, applying the extension of the SMN for multi-allelic loci (Table 2.3) allowed the narrowing down of risk determinants to amino acid 13 at the center of the HLA-DRB1 P4 binding pocket, while amino acid 60, which had earlier been postulated based on structural features[37], was seen as unlikely to play a major role [35,38] (Fig. 2.4).

2.2.3. Microarray Quality Control—Another simple method based on u-statistics is the median on which the Bioconductor package Harshlight is based, a program to identify and mask artifacts on Affymetrix microarrays[35,39,40]. After extensive discussion[41,42], this approach was recently adopted for Illumina bead arrays as “*BASH: a tool for managing BeadArray spatial artifacts*”[43] (Fig. 2.5).

2.2.4. Implementation and Availability—Fig. 2.6 demonstrates the relationships between the different versions of the sign and McNemar tests and how easily even their exact versions, where all possible permutations of data need to be considered, can be implemented in code compatible with R and S-PLUS. For dominant and recessive alleles, the parameters wP/wQ are set to .0/.5, and .5/.0, respectively. HarshLight and BASH (as part of the beadarray package) are available from <http://bioconductor.org>.

3. Univariate Data

3.1. u-Scores

3.1.1. Definitions—Our aim is to first develop a computationally efficient procedure for scoring data based on ordinal information only. We will not make any assumptions regarding the functional relationship between variable and the latent factor of interest, except that the variable has an orientation, i.e., that an increase in this variable is either ‘good’ or ‘bad’. Without loss of generality, we will assume that for each of the variables ‘more’ means ‘better’.

Here, the index k is used for subjects. Whenever this does not cause confusion, we will identify patients with their vector of $L \geq 1$ observations to simplify the notation. The scoring mechanism is based on the principle that each patient $\{x_k = (x_{k1}, \dots, x_{kL})'\}_{k=1, \dots, m_j}$ is compared to every other patient in a pairwise fashion. More specifically, a score $u(x_k)$ is assigned to each patient x_k by counting the number of patients being inferior and subtracting the number of subjects being superior

$$u(x_k) = \sum_{k'} I(x_{k'} < x_k) - \sum_{k'} I(x_{k'} > x_k) \quad (2.5)$$

3.1.2. Computational Aspects—When Gustav Deuchler[44] in 1914 developed what more than 30 years later should become known as the Mann–Whitney test[11] (although he missed one term in the asymptotic variance), he proposed a computational scheme of striking simplicity, namely to create a square table with the data as both the row and the column headers (Fig. 2.7). With this, the row sums of the signs of paired comparisons yields the u-score.

Since the matrix is symmetric, only one half (minus the main diagonal) need to be filled. Kehoe and Cliff[46] pointed out even more redundancies. The interactive program Interord computes which additional entries could be implied by transitivity from the entries already made.

3.2. Randomization and Stratification

3.2.1. Introduction—When several conditions are to be compared, the key principle to be observed is that of randomization. In most situations, randomization is the only practical strategy to ensure that the results seen can be causally linked to the interventions to be compared.

When observational units vary, it is often useful to reduce the effect of confounding variables through stratification. A special case of stratification is the sign test, when two interventions are to be compared within a stratum of two closely related subjects (e.g., twins). Here we will consider situations where more than two subjects form a stratum, e.g., where subjects are stratified by sex, race, education, or the presence of genetic or environmental risk factors. The fundamental concept underlying the sign test is easily generalized to situations where a stratum contains more than two observations.

Under H_0 , the expectation of $T_j = \sum_{k=1}^{M_j} u(x_{jk})$ is zero. For the two-sample case, Mann and Whitney, in 1947 [11], re-invented the ‘u-test’ (this time with the correct variance). As u-scores are a linear function of ranks ($u = 2r - (n + 1)$), their test is equivalent[47] to the rank-sum test proposed in 1954 by Wilcoxon[10]. The Wilcoxon–Mann/Whitney (WMW) test, in turn, is a special case of the 1952 Kruskal–Wallis (KW) rank test for >2 groups[12]. The notation used below will prove to be particularly useful to generalize these results.

The test statistic W_{KW} is based on the sum of u-scores $U_j = \sum_{k=1}^{M_j} u(X_{jk})$ within each group $j = 1, \dots, p$. It can be computed as a 'quadratic form'

$$W_{KW} = \mathbf{U}' \mathbf{V}^- \mathbf{U} = \sum_{j,j'=1}^p U_j v_{jj'}^- U_{j'} \sim_{as} \chi_{p-1}^2 \quad (2.6)$$

where $\mathbf{V}^-$ is a generalized inverse of the variance-covariance matrix $\mathbf{V}$. For the KW test, the matrix $\mathbf{V}$, which describes the experimental design and the variance among the scores are:

$$\mathbf{V}_{KW} = s_0^2 \left(\begin{pmatrix} M_1 & & \\ & \ddots & \\ & & M_p \end{pmatrix} - \begin{pmatrix} \frac{M_1 M_1}{M_+} & & \frac{M_1 M_p}{M_+} \\ & \ddots & \\ \frac{M_p M_1}{M_+} & & \frac{M_p M_p}{M_+} \end{pmatrix} \right), s_0^2 = \frac{1}{M_+ - 1} \begin{cases} \sum_{x=1}^{M_+} u^2(x) & \text{unconditional} \\ \sum_{j=1}^p \sum_{k=1}^{M_j} u^2(X_{jk}) & \text{conditional on ties} \end{cases} \quad (2.7)$$

Traditionally, the WMW and KW tests have been presented with formulae designed to facilitate numerical computation at the expense of conceptual clarity:

$$\mathbf{V}_{KW}^- = \frac{1}{s_0^2} \begin{pmatrix} \frac{1}{M_1} & & \\ & \ddots & \\ & & \frac{1}{M_p} \end{pmatrix}, s_0^2 = \frac{M_+ (M_+ + 1)}{3} \begin{cases} 1 & \text{unconditional} \\ 1 - \sum_{g=1}^G (W_g^3 - W_g) / (M_+^3 - M_+) & \text{conditional on ties} \end{cases}$$

where W_g indicates the size of the g -th tie (group of identical observations). (The constant '3' in the denominator of s_0^2 is replaced by '12' for ranks). The unconditional version (equivalent to the conditional version in the absence of ties) then yields the well known and computationally simple form

$$W_{KW} = \frac{3}{M_+ (M_+ + 1)} \sum_{j=1}^p \frac{T_j^2}{M_j} \sim_{as} \chi_{p-1}^2.$$

With the now abundant computer power one can shift focus to the underlying design aspects and features of the score function represented in the left and right part of (2.7), respectively. In particular, the form presented in (2.6) can be more easily extended to other designs (below).

3.2.2. Complete balanced design—To extend u-tests to stratified designs, we add an index $i = 1, \dots, n$ for the blocks to be analyzed. With this, the (unconditional) sign test can be formally rewritten as

$$\mathbf{V}_{ST} = n s_0^2 \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \right) = n s_0^2 (\mathbf{I} - \frac{1}{2} \mathbf{J}), \mathbf{V}_{ST}^- = \frac{1}{n s_0^2} \mathbf{I}, s_0^2 = \sum_{j=1}^2 u^2(j) = 2.$$

$$W_{ST} = \frac{1}{2n} \sum_{j=1}^2 T_j^2 = \frac{1}{n} (N_+ - N_-)^2 \sim_{as} \chi_{p-1}^2$$

Coming from a background in econometrics, the winner of the 1976 Nobel prize in economics, Milton Friedman, had presented a similar test[48] in 1937, yet for more than two conditions.

$$\mathbf{V}_{\text{FM}} = ns_0^2 \left(\begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{p} & & \\ & \ddots & \\ \frac{1}{p} & & \frac{1}{p} \end{pmatrix} \right) = ns_0^2 \left(\mathbf{I} - \frac{1}{p} \mathbf{J} \right), \mathbf{V}_{\text{FM}}^- = \frac{1}{ns_0^2} \mathbf{I}, s_0^2 = \frac{1}{p-1} \sum_{j=1}^p u^2(j) = \frac{p(p+1)}{3}.$$

$$W_{\text{FM}} = \frac{3}{np(p+1)} \sum_{j=1}^p T_{+j}^2 \sim_{as} \chi_{p-1}^2$$

In either case, non-informative blocks are excluded with ‘conditioning on ties’.

In the field of voting theory, the blocks represent voters and the u-scores are equivalent to the Borda counts, originally proposed in 1781[49], around the time of Arbuthnot’s contributions, although the concept of summing ranks across voters was already introduced by Ramon Llull (1232–1315)[50].

3.2.3. Incomplete and/or unbalanced designs—In the 1950s[51,52], it was demonstrated that the Kruskal–Wallis[12] and Friedman[48] tests, and also the tests of Durban [53] and Bradley–Terry[54] can be combined and extended by allowing for both several strata and several observations within each cell (combination of stratum and group).

However, when blocks represent populations of different size (proportionate to m_i , say) and/or have missing data ($M_i < m_i$), the problem arises how to deal with unbalanced and/or incomplete designs. Between 1979 and 1987, several approaches for weighting blocks were suggested[33]. In 1988, Wittkowski[33] used the marginal likelihood principle to prove that Mike Prentice’s “intuition” [55] was correct, i.e., that u-scores (or ranks) should be scaled by a factor of $(m_+ + 1)/(M_+ + 1)$ to reflect differences in population sizes and missing data. In 2005, Alvo *et al.*[56,57] confirmed these results using a different approach.

$$u(x_{ijk}) = \sum_{j' \neq k} \mathbf{I}(x_{ij'k} < x_{ijk}) - \sum_{j' \neq k} \mathbf{I}(x_{ij'k} > x_{ijk}) \quad U_{ij} = \frac{m_i + 1}{M_i + 1} \sum_{k=1}^{M_{ij}} u(x_{ijk})$$

$$\mathbf{V}_i = s_{i0}^2 \left(\begin{pmatrix} M_{i1} & & \\ & \ddots & \\ & & M_{ip} \end{pmatrix} - \begin{pmatrix} \frac{M_{i1}M_{i1}}{M_{i+}} & & \frac{M_{i1}M_{ip}}{M_{i+}} \\ & \ddots & \\ \frac{M_{ip}M_{i1}}{M_{i+}} & & \frac{M_{ip}M_{ip}}{M_{i+}} \end{pmatrix} \right), s_{i0}^2 = \frac{1}{M_{i+}-1} \left(\frac{m_{i+}+1}{M_{i+}+1} \right)^2 \left\{ \begin{array}{ll} \sum_{x=1}^{M_{i+}} u^2(x) & \text{unconditional} \\ \sum_{j=1}^p \sum_{k=1}^{M_{ij}} u^2(x_{ijk}) & \text{conditional on ties} \end{array} \right.$$

(2.8)

In contrast to the above special cases, no generalized inverse with a closed form solution is known, so that one has to rely on a numerical solution:

$$\mathbf{W} = \mathbf{U}'_+ \mathbf{V}_+^- \mathbf{U}_+ = \sum_{i=1}^n \sum_{j,j'=1}^p U_{ij} v_{jj'}^- U_{ij'} \sim_{as} \chi_{p-1}^2 \quad (2.9)$$

When the conditions to be compared are genotypes, a situation may arise, where the genotypes of some subjects are not known. Still, phenotype information from those subjects can be used when computing the scores by assigning these subjects to a pseudo group $j = 0$, which is used for the purpose of scoring only.

3.3. Applications

3.3.1. Binary Data / Mantel Haenszel—Genome-wide association studies (GWAS) often aim at finding a locus, where the proportions of alleles differ between cases and controls. For most human single nucleotide polymorphisms (SNPs), only two alleles have been seen. Thus, the data can be organized in a 2×2 table,

	Control	Cases	
Allele 0	$M_1^{(0)}$	$M_2^{(0)}$	$M_+^{(0)}$
Allele 1	$M_1^{(1)}$	$M_2^{(1)}$	$M_+^{(1)}$
	M_1	M_2	M_+

While data for the sign test can also be arranged in a 2×2 table, the question here is not, whether $M_1^{(1)} \neq M_2^{(0)}$, but whether $M_1^{(1)}/M_1 \neq M_2^{(1)}/M_2$. Thus, the appropriate test here is not the sign test, but Fisher's exact or, asymptotically, the χ^2 test for independence or homogeneity. This χ^2 test is asymptotically equivalent to the WMW test for two response categories. As with the sign test, ties are considered 'exact' in genetics and, thus, the variance conditional on the ties is applied.

When the data are stratified, e.g., by sex, one obtains a 2×2 table for each block and the generalization of the χ^2 test is known as the Cochran–Mantel–Haenszel (CMH) test, which can also be seen as a special case of the W-test

$$W_{\text{CMH}} = \frac{\left(\sum_{i=1}^n \frac{m_{i+}+1}{M_{i+}+1} M_{i1} M_{i2} (P_{i1} - P_{i2}) \right)^2}{\sum_{i=1}^n \left(\frac{m_{i+}+1}{M_{i+}+1} \right)^2 \frac{M_{i+}^2}{M_{i+}-1} M_{i1} M_{i2} P_{i+}(1-P_{i+})} = \frac{\left(\sum_{i=1}^n \frac{m_{i+}+1}{M_{i+}+1} (M_{i2}^{(1)} M_{i1}^{(0)} - M_{i1}^{(1)} M_{i2}^{(0)}) \right)^2}{\sum_{i=1}^n \left(\frac{m_{i+}+1}{M_{i+}+1} \right)^2 \frac{1}{M_{i+}-1} M_{i+}^{(1)} M_{i+}^{(0)} M_{i1} M_{i2}} \sim_{\text{as}} \chi_1^2$$

3.3.2. Degrees of freedom in GWAS—For most statistical applications, the degrees of freedom for χ^2 tests is the rank of the variance-covariance matrix $\mathbf{V}$. For genetic and genomic screening studies, however, where the dimension of the variance-covariance matrix may vary depending on the number of groups present for a particular gene or locus. Clearly, it would be inappropriate in a screening study to decrease the degrees of freedom (df) for the χ^2 distribution if one (or more) of the groups is (are) missing, but revert to the full df when a single observation for this group is available for another gene or locus.

3.3.3. Implementation and Availability—Among the many obstacles that have prevented statistical methods based on u-scores from being used more widely is that they are traditionally presented as an often confusing hodgepodge of procedures, rather than a unified approach backed by a comprehensive theory. For instance, the Wilcoxon rank sum and the Wilcoxon signed rank test are easily confused. Both were published in the same paper [10], but they are not special cases of a more general approach. The Lam–Longnecker test [58], on the other hand, directly extends the WMW to paired data. Finally, the Wilcoxon rank sum and the Mann–Whitney u test are equivalent [47], although they were independently developed based on different theoretical approaches.

Above, we have demonstrated that several rank or u-tests can be presented as a quadratic form $W = \mathbf{U}'_+ \mathbf{V}_+^{-1} \mathbf{U}_+ \sim_{as} \chi^2_{p-1}$, where the variance-covariance matrix $\mathbf{V}_+$ reflects the various designs. The `muStat` package (available for both R and S-PLUS) takes advantage of this fact, both by providing the unified function `mu.test()` (in addition to replacement functions of the original R and S-PLUS functions, for compatibility) and by being based on a single copy of code (Fig. 2.8), although internally an optimized version may be used for special cases. (The case of censored data will be handled as a special case of multivariate data, see Section 4.2.1).

To ensure that p-values are comparable within a study (see Section 3.3.2), the function `mu.test()` allows the number of degrees to be fixed ($df > 0$) or to be determined by the design matrix ($df = 0$), in addition to it being computed from the observed data ($df = -1$).

4. Multivariate Data

4.1. Introduction

Few biological systems can be sufficiently characterized by a single variable only. A single measure often does not appropriately reflect the effect of all relevant genetic or environmental risk factors, clinical or epidemiological interventions, or personal preferences. Sometimes the definite measure is not easily obtained, so that several surrogate measures need to be evaluated. At other times, e.g., when assessing a complex syndrome or a chronic disease, a definite measure may not even exist.

Still, most statistical methods for multivariate data are based on the (generalized) linear model, either explicitly, as in regression, factor, discriminant, and cluster analysis, or implicitly, as in neural networks. One scores each variable individually on a comparable scale, either present/absent, low/intermediate/high, 1–10, or z-transformation, and then defines a global score as the linear combination (weighted average) of these scores. Thus, it is assumed that it is known how to transform each variable to a common scale, so that a weighted average of these transformed variables can be meaningfully interpreted and that these weights are constant.

The linear model became popular mainly because its mathematical elegance lead to computational efficiency and parameters of alluring simplicity. When applied to real-world data, however, this approach may have shortcomings, because biological systems are typically regulated by various, often unknown, feedback loops so that the functional form of relationships between measurement and activity or efficacy is typically unknown, except within narrowly controlled experiments. Since the relative importance of the variables, the correlation among them, and the functional relationship of each variable with the immeasurable latent factor ‘safety’, ‘activity’ or ‘effectiveness’ are typically unknown, construct validity [59] cannot be established on theoretical grounds and one needs to resort to empirical ‘validation’, choosing weights and functions to provide a reasonable fit with a ‘gold standard’ when applied to a sample, a process of questionable validity by itself [60]. The Delphi oracle, where women intoxicated by fumes predicted the future was often difficult to interpret. The ‘Delphi method’ [61] approach to scoring systems, where weights and functions are agreed upon by a group of experts may facilitate comparison between studies, yet comparability along a scale with questionable validity may still yield questionable results. The diversity of scoring systems used attests to the subjective nature of this process.

Non-parametric methods, in general, are designed for ordinal data, where a one-unit difference may not carry the same ‘meaning’ across the range of possible values [62], and, thus, avoid artifacts created by making unrealistic assumptions. Consequently, non-parametric methods are particularly well suited for analyses in human phenomics [63]. The marginal likelihood principle (MrgL) provides a general framework to extended rank tests to missing data [33, 64–

66]. In 1992, it was demonstrated that MrgL procedures for censored data can be generalized to cover multivariate ordinal observations[67], in general. While this approach proved eminently useful[68–72], the computational effort was prohibitive.

Drawing on the analogy of the Wilcoxon rank sum[10] (MrgL) with the Mann–Whitney test [73] (u-statistics), a 1914 algorithm[44] yielded a computationally more efficient approach, which can easily be extended to more complex partial orderings[62,74] and designs[31,33, 75]. Below, it will be demonstrated how μ -scores (u-scores for multivariate data) cover situations where a one-unit difference may carry a different ‘meaning’ across variables[76] and, thus, can integrate information even when the events counted are incomparable or the variables’ scales differ, as long as each variable has the same ‘orientation’.

4.2. Partial and Incomplete Orderings

We will now extend the notation introduced in Section 3.1 (using $i = 1, \dots, n$ for blocks, $j = 1, \dots, p$ for groups, and $k = 1, \dots, m_{ij}$ for replications) by adding an additional index $\ell = 1, \dots, L$ for the variables to be analyzed. The scoring mechanism is based on the principle that each subject

$$\{x_{jk} = (x_{jk1}, \dots, x_{jkL})\}_{j=1, \dots, p; k=1, \dots, m_j}$$

is compared to every other subject in a pairwise manner. For stratified designs, these comparisons will be made within each stratum only.

When the observed outcomes can be assumed to be correlated with an unobservable latent factor, a partial ordering[67] among the patients can easily be defined. If the second of two patients has values higher in at least one variable and at least as high among all variables $\ell = 1, \dots, L$ (or equivalently, lower in none), it will be called ‘superior’:

$$x_{jk} < x_{j'k'} \iff \left\{ \forall_{\ell=1, \dots, L} x_{jk\ell} \leq x_{j'k'\ell} \wedge \exists_{\ell=1, \dots, L} x_{jk\ell} < x_{j'k'\ell} \right\}. \quad (3.1)$$

Many partial orderings can be defined, which, by definition, are transitive ($a < b \wedge b < c \Rightarrow a < c$). Orderings like (3.1), which treat all variables equally, are called ‘regular’. The partial ordering for an (interval) censored variable is just one example. Several more examples will be given below.

Even though a partial ordering does not guarantee that all patients can be ordered on a pairwise basis, all patients can be scored. One can assign a score in exactly the same fashion as described in (2.5), using the partial order (3.1) instead of the simple univariate order. By definition, μ -scores are ‘intrinsically valid’, i.e., independent of the choice of (non-zero) weights and (monotonous) transformations assigned to the variables.

4.2.1. Censored Data—Many phenotypes are ‘censored’. A typical case is ‘survival’, where some subjects may still be alive at the end of a study. Similarly, a subject may be surveyed as not yet having cancer. Other examples are events that are not directly observable, like the recurrence of cancer, where often only the last date with a negative and the first date with a positive result are known. With genetic studies, censored information arises, for instance, when the prevalence of a disease that develops over time is observed in subjects that vary in age, such as in late-onset diabetes, cardiovascular diseases, cancer, etc. For such studies, a negative observation merely means that this subject does not yet have developed the disease, rather than that the subject is ‘immune’.

In 1965, Gehan[77,78] demonstrated how u-score can be applied to censored (including interval-censored) variables (see Fig. 2.9). Assuming that the variables represent the last date negative (LDN) and the first date positive (FDP), Subject A experiences the event under investigation ‘later’ than subject B if $LDN(A) > FDP(B)$.

In clinical or epidemiological research, such data may arise when the exact date of an event, e.g., infection or recurrence, is not known, but the event is known to have happened between the date of the last negative test x_{jk1} and the date of the first positive test x_{jk2} . Right censored data are a special case ($x_{jk2} = x_{jk1}$: event, $x_{jk2} = \infty$: censoring). Thus, pairs of intervals can be ordered, if they do not overlap. For left- and right-censored observations, LDN and FDP are $-\infty$ and $+\infty$, respectively.

Tests for censored data are often presented in an equivalent form, where the first variable is the time point, and the second is the ‘censoring indicator’. The above representation, however, is more easily generalized, e.g., to ‘interval censored’ data, where $x_{jk1} < x_{jk2} < \infty$.

To further clarify the relation between censored and multivariate data, it is convenient to consider the most general case, interval censored observations. In clinical or epidemiological research, such data may arise when the exact date of an event, e.g., infection or recurrence, is not known, but the event is known to have happened between the date of the last negative test x_{jk1} and the date of the first positive test x_{jk2} . Right censored data are a special case ($x_{jk2} = x_{jk1}$: event, $x_{jk2} = \infty$: censoring). Thus, pairs of intervals can be ordered, if they do not overlap, or, equivalently, if both time points in one subject are earlier than both time points in the other subject:

Schemper[79] generalized the Gehan test to more than two groups. Also, from the work of Hoeffding[73] and Lehmann[80], u-scores for censored data can be analyzed for all the stratified designs mentioned above.

4.2.2. Ambiguous paired comparisons—If only a single variable needs to be considered ($L = 1$) and all observations are different, the order is complete (Fig. 2.10a). If identical observations (ties) are present, two cases need to be considered, as with the sign test, above. On the one hand, ties may be due to the underlying phenomena and on the other hand may be caused by discretization or by observing a discrete surrogate variable for a continuous phenomenon. In both cases, there are three possibilities for each pair of patients. In the former case, they are $<$, $>$, or $=$ (Fig. 2.10b), in the latter, where ties reflect some ambiguity[29], they are $<$, $>$, or $\equiv$ (Fig. 2.10c). Intervals, however, can only be ordered, if they are disjoint, so that some paired comparisons may be undetermined. In Fig. 2.10d, for instance, it is not known, if the patient infected between the first and the third follow-up visit (1..3) was infected earlier than the patient infected between the second and the third visit (2..3). The same rationale applies to situations with several ($L > 1$) variables (Fig. 2.10e). For multivariate data, the order between two subjects is undetermined if $x_{jk\ell} < x_{j'k'\ell}$ for some variable ℓ , while $x_{jk\ell'} > x_{j'k'\ell'}$ for another variable ℓ' . In either case, the ordering may be ‘partial’, rather than ‘complete’.

From Fig. 2.10d and e, the Gehan test[77, 78] can be easily extended to multivariate data, using the approach originally suggested by Hoeffding[81], because it relies on the existence of a partial order only, irrespective of how that partial ordering was created.

4.2.3. Information Content—As an alternative to the ‘weak’ partial ordering (3.1), one might require its ‘strong’ cousin [82]

$$x_{jk} \ll x_{j'k'} \iff \left\{ \forall_{\ell=1,\dots,L} x_{jk\ell} < x_{j'k'\ell} \right\}. \quad (3.2)$$

At first sight, the strong order (3.2) may seem more appropriate for discretized variables, because the true order of the observations discretized into a tie is unknown. If each variable is presumed to be a surrogate for the same latent factor, however, the strong partial order highlights a potential downside of μ -scores, namely that the number of ambiguous paired comparisons increases with the number of variables, unless the variables are highly correlated. As a result, a larger sample may be needed to achieve the desired power. Allowing for ties to be broken by observations in other variables again yields the weak partial ordering. Thus, the weak regular partial ordering (3.1) will be called ‘natural’ for applications where each variable can be assumed to be a surrogate for the same underlying latent factor.

Formally, we will define the proportion of paired comparison with subject A that can be decided as the ‘information content’ of subject A’s score. Unless the variables are highly correlated, so that adding more variables merely breaks ties, information content often decreases with the number of variables and the rigidity of the partial ordering. On the other hand, the more information about the underlying models is available to justify combining several variables into a linear combination, the more paired comparisons can be decided.

Below, we will discuss two non-parametric approaches to increase information content by reflecting prior knowledge. The first is to transform the data using less stringent assumptions than those imposed by the linear model. Then, we will demonstrate how allowing for hierarchical structures among variables can increase information content when variables are known to belong to different factors (e.g., safety and efficacy) or sub-factors[45,83]. In turn, identifying structures that maximize information content then leads to a non-parametric alternative to traditional factor analyses, including more efficient questionnaires.

4.2.4. Computational Aspects—Deuchler’s[44] univariate algorithm to depict a complete ordering as a symmetric matrix is easily extended to partial and imperfect (non-transitive) orderings for data profiles comprising several variables. Morales *et al.*[45] separated this process into two steps, as outlined in Fig. 2.11. First, Deuchler’s[44] univariate paired

comparisons are represented as the matrices $U^{(\ell)} = \left(\left(u_{kk'}^{(\ell)} \right) \right)$ (middle row) with $u_{kk'}^{(\ell)} = ?$ if $x_k^{(\ell)}$ or $x_{k'}^{(\ell)}$ are missing and $u_{kk'}^{(\ell)} = I \left(x_{k'}^{(\ell)} < x_k^{(\ell)} \right) - I \left(x_k^{(\ell)} > x_{k'}^{(\ell)} \right)$ otherwise [76], extending Deuchler’s[44] algorithm to imperfect (not necessarily partial) orderings allowing for missing data.

The matrix U obtained by the ‘AND’ operation

$$U = \oplus_{\ell} U^{(\ell)} = \left(\left(u_{kk'}^{(\ell)} \right) \right), \text{ where } u_{jj'} = \begin{cases} 1 & \exists \ell: u_{kk'}^{(\ell)} = 1 \wedge \forall \ell: u_{kk'}^{(\ell)} \neq -1 \\ 0 & \exists \ell: u_{kk'}^{(\ell)} = 0 \wedge \forall \ell: \left| u_{kk'}^{(\ell)} \right| \neq 1 \\ -1 & \exists i: u_{kk'}^{(i)} = -1 \wedge \forall i: u_{kk'}^{(i)} \neq 1 \\ ? & \text{otherwise} \end{cases}$$

with ‘?’ indicating ambiguity, is the same as the matrix one would obtain by applying the imperfect ordering defined earlier[76] and, thus, the scores obtained from $U = \oplus_{\ell} U^{(\ell)}$ (bottom of Fig. 2.11) are the non-hierarchical μ -scores.

Recently, Cherchye and Vermeulen[84] proposed to replace the matrix of paired comparisons, which has entries ‘+1’, ‘0’, ‘?’, and ‘-1’, by a computationally simpler ‘greater-equal’ (GE) matrix of binary entries $u_{kk'} = I(x_{k\ell} \geq x_{k'})$. Combining these two approaches, one can compute univariate GE matrices $u_{kk'\ell} = I(x_{k\ell} \geq x_{k'\ell})$ first and then combine them into a multivariate GE matrix. Several of these matrices can then be combined, again, thereby providing a high level of flexibility.

When comparing the severity of ‘damage’ seen in ultrasound or radiologic images[85], for instance, it may not be possible to define all relevant variables. In these situations, one might present a rater with pairs of images (A, B) to be judged as $A < B$, $A > B$, or $A \cong B$. As in the univariate case (see Section 3.1.2), an interactive system could then reduce the number of questions to be asked by inferring which paired comparisons can be implied from the answers already made. If the number of pairs that each rater can be presented is to be limited, the unanswered questions would be considered ambiguous.

4.2.5. Implementation and Availability—In the muStat package for R and S-PLUS packages are available from <http://cran.r-project.org> and <http://csan.insightful.com>, respectively, these steps are implemented as the functions `mu.PwO` and `mu.AND`, followed by computation of scores in information content (Fig. 2.12).

Using the asymptotic results of Hoeffding[73], the resulting μ -scores can then be analyzed with `mu.test` (Fig. 2.8) or through bootstrapping [86].

4.3. Relationships Between Variables

4.3.1. Graded Variables—Often, a simple transformation of the variables may suffice to reflect additional knowledge. For graded variables, where one unit impacts less in a ‘lower grade’ than in a ‘higher grade’ variable[74,84] one can split each value of variable (ℓ) (sorted by grades) into the value of the lowest grade variable Δ_ℓ and incremental values of the higher grade variables $\Delta_{\ell'=2 \dots \ell}$.

$$\begin{aligned}x_{k,(1)} &= x_{k,(1)} \Delta_1 \\x_{k,(2)} &= x_{k,(2)} \Delta_1 + x_{k,(2)} \Delta_2 \\&\dots \\x_{k,(L)} &= x_{k,(L)} \Delta_1 + x_{k,(L)} \Delta_2 + \dots + x_{k,(L)} \Delta_L\end{aligned}$$

Thus, the profile of counts sorted by grade $(x_{k,(1)}, x_{k,(2)}, \dots, x_{k,(L)})$ can be expressed as the column sums $(x_{k,(\geq 1)}, x_{k,(\geq 2)}, \dots, x_{k,(=L)})$ where $x_{k,(\geq \ell)} = \sum_{\ell'=\ell}^L x_{k,(\ell')}$. The partial ordering for graded variables

$$(x_{k,(1)}, \dots, x_{k,(L)}) < (x_{k', (1)}, \dots, x_{k', (L)}) \iff \forall_{\ell=1}^L x_{k,(\geq \ell)} \leq x_{k', (\geq \ell)} \wedge \exists_{\ell=1}^L x_{k,(\geq \ell)} < x_{k', (\geq \ell)} \quad (3.3)$$

is equivalent to the regular ordering (3.1) applied to the cumulative variables $x_{k,(\geq \ell)}$.

Although each profile’s outcomes are decomposed into additive components Δ_ℓ , substantially weaker assumptions are made than with linear weight (lw) scores, because the additive components can be unknown and may even differ between pairs, noting that for subjects far apart (subject k is lower than subject k' for each of the variables) or incomparable (some variables higher for subject k and some variables higher for subject k') the weights are irrelevant. Thus, the weights only need to be ‘locally similar’, rather than ‘globally constant’.

4.3.2. Hierarchically Structured Variables—If the variables are related to different ‘factors’ and the order between subjects A and B is ambiguous with respect to variables related to one factor (e.g., genetics), unambiguous results with respect to another factor (e.g., environment) can ‘overwrite’ this ambiguity. The advantage of creating the matrices reflecting the univariate orderings first (`mu.PwO`) and combining them in a separate step (`mu.AND`) before computing the scores (`mu.Sum`), is that incorporating knowledge about the sub-factor hierarchy

through hierarchically combining the matrices typically reduces loss of information content [45] (number of unambiguous paired comparisons contributing to a score).

If the variables X_1 and X_2 in Fig. 2.11 were related to the same ‘factor’, while the variables Y_1 and Y_2 are related to another ‘factor’, one could replace $U_{NH} = \bigoplus_{\ell} U^{(\ell)}$ by $U_H = \bigoplus \{\bigoplus_{\ell: X_1, X_2} U^{(\ell)}, \{\bigoplus_{\ell: Y_1, Y_2} U^{(\ell)}\}\}$. Fig. 2.13 demonstrates how reflecting the structure increases information content by resolving the ambiguity related to comparing A vs. D.

Fig. 2.13, shows that reflecting more hierarchical information can never decrease and typically increases information content, because it reduces the effect of ‘noise’ that may have caused paired comparisons within a factor to be ambiguous. If all ambiguities are resolved, the μ -scores become ranks, which are uniformly spaced across the widest possible range.

4.3.3. Doubly Interval-Censored Variables—Since Gehan, u-scores have been used to handle single in interval censored data, where only the last date the subject is known to have been negative (LDN) and the first date the subject is known to have been positive (FDP) are available. Subject A experiences the event under investigation ‘later’ than subject B if LDN (A) > FDP(B). For left- and right-censored observations, LDN and FDP are $-\infty$ and $+\infty$, respectively. μ -Scores, consequently, can easily handle censored (including interval-censored) variables (Fig. 2.14). Adding a hierarchical ordering to interval-censored data also provides a solution to analyzing doubly interval censored data, where, for instance, both the date of exposure and the date of disease manifestation are interval censored (Fig. 2.15).

4.3.4. Genetical Structure (Epistasis between Diplotypes)—With SNP arrays, which can have an even larger number of variables (currently >100K) than expression arrays, the known sequence of the SNPs on the chromosome (ordinal structure) can be utilized to reduce complexity. From Fig. 2.16, a disease locus in close proximity to a particular SNP (e.g., SNP A) is likely to be highly correlated (in linkage disequilibrium, LD) with this SNP.

A disease locus approximately equidistant to two adjacent loci (SNP C and D) may be in LD with both of the adjacent SNPs. Traditionally, one would look at the plot of the univariate $-\log_{10}(p)$ by locus and give more confidence to a locus, if several neighboring loci are also in LD. With μ -scores, one can now directly aggregate evidence from neighboring SNPs by assuming that each chromosome consists of intervals around and between SNPs. (Note that we will not require the demarcations between these intervals to be known.) In particular, we will treat each genetic interval as two variables, each contributing to the same disease locus as ‘factor’ (either side of Fig. 2.13).

With high density SNPs, several adjacent intervals may form diplotypes in LD with a single disease locus, the most informative diplotypes being determined by the LD/noise ratio.

Epistasis is defined as an interaction between diplotypes that is associated with a phenotype. Within muStat, epistasis is indicated as a hierarchical structure among diplotypes (Fig. 2.17).

4.3.5. Summary—While u-statistics for univariate data[11,18] and censored data[77–79] are widely used, u-statistics for multivariate data[73,80] are rarely applied, presumably because they were not presented in an easy-to-use form and transistors had just been put to practical use[87]. A more fundamental problem may have been an even more important hindrance. As the number of variables increases, information content (the proportion of paired comparisons that can be decided) drops fast, until all μ -scores (u-scores for multivariate data) become NA, especially with a ‘strong’ order[82], as compared to its ‘weak’ counterpart[67]. As pointed out recently[88], averaging univariate u-scores[89] or using the lexicographical order avoid this

problem, yet require that the relative importance of the variables be constant and known or that less important variables contribute only by breaking ties.

Two recently developed extensions of u-statistics resolve this conundrum. First, the process of scoring multivariate data was separated into first determining the univariate orders first [44] and then combining them [45] into an incomplete order. As the proposed combination of incomplete orders results in yet another incomplete order, a ‘tree’ of incomplete orders can be defined, wherein hierarchical (sub-) factor structures can reflect functional, topological, or temporal relationships. Conversely, ‘ μ -factor analysis’ can identify hierarchical factor structures that maximize information content [90]. Second, cumulation of count variables [62] allows for ‘graded’ variables (e.g., ‘mild’ < ‘moderate’ < ‘severe’) but avoids the infinite relative weights implicitly assigned with a lexicographical order. μ -Scores can be used for various analyses, including testing differences between groups defined by simple genotypes with respect to complex phenotypes, correlating complex genotypes with complex phenotypes or identifying genetic variables that explain (correlate best with) a complex phenotype.

Once GE matrices (or scores) are created, further strategies can be employed to incorporate design or model knowledge. For instance, if variables can be ordered, but the outcomes cannot be cumulated (as in Section 4.3.1), Kendall’s [91] correlation coefficient r_K or, equivalently, the Jonckheere–Terpstra test [92,93] could be used.

For paired observations the Wilcoxon signed rank test [10] (WSR) is often proposed as a ‘non-parametric’ alternative to the paired t-test. Since the WSR is based on the arithmetic difference between the paired observations, however, the results are not independent of scale transformations. The Lam–Longnecker test [58], instead, is directly based on the WMW test, except that the variance is reduced by a factor $1 - r_S$, where r_S is Spearman’s rank correlation coefficient. Replacing r_S by the closely related r_K yields a truly ‘non-parametric’ alternative to the WSR, which fits seamlessly into the above unified concept of u-tests.

4.4. Applications

4.4.1. μ -Discriminant analysis: Collaboration vs Coregulation (supervised)—The focus of many traditional methods is on ‘correlation’ as an indicator of ‘coregulation’, i.e., of variables along the same ‘pathway’. Often, however, different subjects may respond along different pathways, i.e., several pathways may be able to “share the load” [94]. Then, including several variables (or pathways) may provide better discrimination between categories than using individual variables, until the variables included begin to model noise, in which case the information content of the resulting scores would begin to decline (see Section 4.2.3).

μ -Scores enable a paradigm shift by introducing a novel concept of interaction. With the linear model, finding the best discriminating function was often futile, because the choice of the best discriminating set depended on the linearizing transformations and the relative weights chosen. Selecting one among the many ‘best solutions’ generated by the various combinations of subjective transformations and weights was difficult at best. Moreover, the mere assumption that the relative importance among two variables would not depend on the magnitude of these and other variables could easily bias the results. μ -Scores are independent of (monotonous) transformations and (positive) weights, so that solutions are less dependent on the assumptions made and, thus, allow, for the first time, a new type of questions to be approached. Instead of focusing on ‘coregulated’ variables within a pathway, one could search for variables representing ‘collaborating’ pathways.

This paradigm shift has direct implications for applications in systems biology. In Spangler *et al.* [94], who compared activity of dopaminergic receptors between rats addicted to sugar and control rats, the response in the CPU (Fig. 2.18, bottom) was clearly more ‘coordinated’ than

in the NAC (Fig. 2.18, top). In the NAC, bi-variate profiles including D3 and either of D1, D2, pD, and pT discriminated better than their components, and the best discrimination was seen for tri-variate profiles containing both D3 and pT discriminated even better ($p < .0001$). In the CPU, in contrast, D2 ($p = 0.008$) discriminated better than any combination.

A recent study comparing patients with Fanconi anemia type C (FA-C) to normal controls [96] identified in a first step 200 differentially expressed genes in uni-variate comparisons (significance indicated by size of nodes in Fig. 2.19). These genes showed some clustering, albeit without hinting to any suggestive interpretation. By multivariate analyses, the genes AURKA and RRM2 emerged as the member of many of the most significant pairs and triplets, respectively. Both genes are targets of approved drugs in cancer treatment, suggesting that these drug might become the first treatments for FA-C patients.

Thus, adding cooperation as a novel concept in system biology opens an additional dimension for the interpretation of functional relationships between molecular variables. By integrating the muStat output with systems such as Cytoscape (<http://www.cytoscape.org/>), collaborative relationships are easily visualized.

4.4.2. Genome-Wide Association Studies—After the human genome was decoded in 2004, it was widely expected that medicine would enter a new era of personalized medicine, where common diseases, such as hypertension, obesity, and cancer could be successfully treated. Many of the early results, however, could not be confirmed in subsequent studies.

Fig. 2.20 exemplifies the problems experienced. Traditionally, each SNP is analyzed using either the Cochran–Armitage χ^2 test for trend[97] with weights (1,1,0), (1,2,3), and (0,1,1) for dominant, additive, and recessive effects, respectively (the 2×2 ‘allelic’ test based on counts of alleles is less appropriate[32] for the same reason as the TDT, see Section 2.2.1, and, thus, omitted). To the dismay of investigators, even the most significant loci often turned out to be false positive, while important loci are often overlooked.

In Fig. 2.20, for instance, comparing 177 cases of children with Childhood Absence Epilepsy (CAE) are compared against three sets of 354 controls, matched against different subsets of Ancestry Informative Markers (AIMs). Not surprisingly, the traditional and the μ -score analyses yield similar results. Only one SNP has at least one $\max(\chi^2) / \mu$ p-value below 10^{-3} . Hence, this locus would not have been drawn any attention.

With μ -scores for diplotypes, in contrast, a median p-value below $< 10^{-5}$ (rare for such a small sample size) and the regular triangular pattern of shading seen in Fig. 2.20 clearly stand out. One locus implicated ($10^{-5.0}$ – $10^{-6.4}$) is close to the SNP that would have been ignored with traditional methods, the other is even more “significant” ($10^{-5.2}$ – $10^{-6.5}$). The former “significance triangle” points at a single exon and the latter at the promoter region of a splice variant, which is expressed in embryonal brain only and is involved in neuronal development, so that it clearly is a strong candidate gene.

GWAS is primarily a screening strategy to generate hypotheses, which then need to be studied further. As a criterion in a selection procedure[98] (balancing not the risk of false positive results, but the number of loci selected vs the risk of overlooking the most relevant loci), rather than as indicating the confidence in a particular result, p-values must not be judged by their absolute value (mainly a function of sample size). Instead, one may give the high priority for further investigation to those hypotheses, that that have both low p-values and high biologic plausibility.

Given the high “false positive rate” in the early GWAS studies, sample size requirements have since been increased to thousands or even tens of thousands of cases. As an alternative, μ -

scores allow to increase the signal/noise ratio by increasing the number of neighboring SNPs that are comprehensively analyzed. As demonstrated in Fig. 2.20, investing more computational effort into better statistical methods can reduce the sample size requirements for GWAS substantially, even for common, complex disease, such as CAE.

Fig. 2.21 shows how representing genetic structure increases sensitivity for detecting association, in general, and epistasis, in particular. Among the 13 chromosomes selected, univariate analysis (bottom border of both diagrams) points to a short range at the begin of Chr 3 and, possibly, Chr 12 as a whole. Moving from SNPs to diplotypes by itself (top border of right diagram) merely shifts the focus on Chr 3 slightly higher. Epistasis between SNPs, as indicated in the square areas, suggests that the range identified on Chr 3 might interact with Chr 11 as a whole and suggests epistasis between Chr 10 and Chr 11. Allowing for epistasis between diplotypes confirms these results, and, in addition, points to ranges on Chr 5 as interacting with ranges on Chr 3, 4, 11, and 12. Moreover, one now sees that the SNPs on several chromosomes are separated into ‘clusters’ interspersed by SNPs that seem to be unrelated.

4.4.3. Complex Phenotypes—The above examples tried to explain relatively simple phenotypes, such as addiction to sugar (Fig. 2.18), presence vs absence of a disease (Fig. 2.19), or degree of atherosclerosis in mice (Fig. 2.21). Even with some ‘monogenic’ diseases, such as Fanconi anemia, however, the phenotype involved can be anything but simple. With FA, the phenotype contains various congenital malformations (often binary), life span, time of cancer manifestation or hematological failure (each of them often censored), and laboratory measurements of chromosomal stability (mostly quantitative) (Fig. 2.22 top).

4.4.4. μ -Factor Analysis: Hierarchical coregulation structures (unsupervised)—Being able to increase information content by introducing a hierarchical structure offers the opportunity to reverse the goals, i.e., to find the hierarchical structure for a given set of variables that maximizes information content. Diana *et al.* [90], for instance, compared three different models (hierarchical structures) with respect to the information content of the resulting scores.

Clearly, an exhaustive search among all possible hierarchical factor structures is not computationally feasible, even for relatively small numbers of variables, unless a limited number (up to a few thousand) is preselected. In Diana *et al.* [90], this approach was used to decide whether commuters categorize transportation modes by:

- **C1: Constraint:** bicycle < motorcycle < car (driver, pax) < transit (bus, tram, metro)
- **C2: Typology:** two-wheeled (bicycle, motorcycle), car (driver, pax), transit (bus, tram, metro)
- **C3: Autonomy:** passenger (car pax, bus, tram, metro) < driver (bicycle, motorcycle, car driver).

In this case, the data suggested autonomy as the concept most relevant for individual decision making. In molecular biology, one might easily use the same approach as an alternative to traditional factor (or ‘cluster’) analysis.

Fig. 2.23 shows another possible use of this method, namely to classify object with respect to how much they gain from the use of the hierarchy. With RDM, in particular, it is clear that the hierarchical and non-hierarchical μ -scores are essentially the same, but there exists a population whose scores are improved by using the hierarchy.

4.4.5. μ - Signatures and Personalized Medicine—A frequent problem in molecular biology is to find a molecular ‘signature’ that discriminates between diseases that present with

similar phenotypes, but require different therapeutic interventions. Better targeted diagnostics could improve patient health by avoiding a trial-and-error approach.

In a study of 33 cellular surface markers and six plasma biomarkers as diagnostic markers for two hemophagocytic syndromes[99], univariate u-tests discriminated poorly, with 30.0%, 21.0%, 30.2%, 37.7%, 38.5%, and 31.8% misclassifications, respectively. By μ -scores, the subset of biomarkers that discriminated best in this population ($p=10^{-15}$, 9.4% misclassifications) consisted of two plasma and two surface markers.

In contrast to traditional methods, discrimination based on μ -scores does not necessarily improve when more variables are added. In fact, less significance with more variables may indicate that noise is being modeled. Moreover, μ -scores are independent of (monotonous) transformations and (positive) weights. Thus, neither the selection of the variables nor the choice of transformations of weights need to be ‘validated’.

For complex diseases, finding signatures that can be applied to the population as a whole has proven to be difficult, if not impossible[100], in part, because different pathways tend to be involved in different sub-populations. As a signature based on μ -scores is ‘intrinsically valid’ for the population in which it is developed (requiring validation for neither the number of variables chosen nor the transformations applied to them), μ -scores may provide a solution for this conundrum. Traditionally, when the same ‘signature’ is used for a wide range of patients, representativity of the population of cases for the subjects to be diagnosed needs to be ‘validated’. However, when a signature is developed based on a sub-population of cases specifically selected to match the subject based on criteria considered relevant by the caring physician (a process, which could involve μ -scores for similarity), the need to ‘validate’ representativity is substantially lessened, so that a personalized signature can be generated *at hoc*. Most importantly, this sub-population of cases is likely to be more homogeneous with respect to the risk factors involved, so that highly predictive signatures will be easier to identify.

4.4.6. Implementation and Availability—In the muStat package for R and S-PLUS, the function `mu.PwO` generates Deuchler’s univariate orderings, `mu.AND` combines them into an incomplete ordering, and `mu.Sums` computes scores and weights from an incomplete ordering (Fig. 2.12). The pseudo-code below demonstrates calculation of the hierarchical (UH) and the non-hierarchical (UNH) μ -scores of Fig. 2.22. The parentheses indicate the hierarchy and variables separated by colon indicate the bounds of interval censored observations.

[illegible]

When using the elementary functions of Fig. 2.12, the statements `mu.Score(x, frml)` and `mu.Sums(mu.AND(mu.PwO(x, frml), frml))$score` are equivalent to the following pseudo-code:

```
mu.Sums( mu.AND(cbind(
  mu.AND(cbind(
    mu.AND(mu.PwO(x[,c(BLnB,BLpC)])),
    mu.AND(cbind(
      mu.AND(mu.PwO(x[,c(DEBnB,DEBpC)])),mu.PwO(x[,Mosaic])),
      mu.PwO(x[,LS0],x[,LS1])),
      mu.AND(cbind(
        mu.PwO(x[,CA0],x[,CA1]), mu.PwO(x[,LK0],x[,LK1])),
        mu.AND(cbind(
          mu.PwO(x[,HM0],x[,HM1]),
          mu.AND(cbind(
            mu.PwO(x[,P0],x[,P1]),... mu.PwO(x[,w0],x[,w1])),
            mu.PwO(x[,BMT0],x[,BMT1])))),
            mu.AND(cbind(
              mu.AND(cbind(mu.PwO(x[,Hzp]),mu.PwO(x[,c(EAR,...)]))),
              mu.AND(cbind(mu.PwO(x[,Htp],mu.PwO(x[,c(BMRk,...)]))))))
            )Score
```

5. Conclusions

Non-parametric tests are often considered the second choice, to be used primarily, as Friedman implied, “to avoid the assumption of normality”[48]. Being based on least square estimates, rather than maximum likelihood, ANOVA is asymptotically distribution-free. As Scheffé demonstrated in Chapter 10 of *The Analysis of Variance*[5], “Nonnormality has little effect on inferences about means”, in general, and ANOVA in particular, because the CLT holds for even moderately sized samples. Hence, when methods based on both the linear model and u-statistics are available, the choice between these approaches should primarily rely on the hypothesis to be tested, rather than the empirical distribution of the residuals[101]. If differences of one unit have the same ‘meaning’ across the scale, the linear model would be more appropriate. If, like in many biological applications, no truly ‘linearizing’ transformation exists, tests based on u-statistics often give more meaningful results[9].

Another reason for using methods based on the linear model is that, with the exceptions of some special cases[102,103], non-parametric methods have not yet been generalized to complex factorial designs. On the other hand, factor and sub-factor structures for graded and censored variables often require restrictive assumptions to be made for the linear model to apply. Hence, a new distinction between parametric and non-parametric methods emerges. The former are more appropriate to reflect controlled structures among independent variables, while the latter may be better suited to reflect uncontrolled structures among dependent variables.

When Fisher introduced structured experimental designs[7], he did so with emphasis on agricultural problems where the assumption of linearity between water, sunshine, and nutrients and the single outcome (yield) is often easily justified. In molecular biology, in contrast, experiments are often simple, but the confounding variables and outcomes can be complex with a structure that is only partially known. Thus, molecular biology may eventually prove as a field where the recent advances in non-parametric statistics are especially useful. In particular, the recent generalizations to structured multi-variate data may, together with sequencing methods and grid or cloud computing may become a equally important recent developments allowing personalized medicine to become a successful strategy for improving health while reducing overall health care cost.

Acknowledgments

The work was supported in part by Grant Number UL1RR024143 from the U.S. National Center for Research Resources (NCRR). Of the many colleagues who have contributed to this chapter through discussions and suggestions, I would like to thank, in particular, Jose F. Morales, Ephraim Sehayek, Sreeram Ramagopalan, and Martina Durner for their input on the biological background, Sreeram Ramagopalan, Bill Raynor, and Norman Cliff for their helpful comments, an anonymous reviewer for an inspiring discussion, and Daniel Eckardt for help with Latin grammar.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003;422:835–847. [PubMed: 12695777]
2. Butler D. The Grid: Tomorrow's computing today. *Nature* 2003;422:799–800. [PubMed: 12712163]
3. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008;299:1335–1344. [PubMed: 18349094]
4. Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry* 2009;166:540–556. [PubMed: 19339359]
5. Scheffé, H. *The Analysis of Variance*. New York, NY: Wiley; 1959.
6. Arbuthnot J. An argument for divine providence taken from the constant regularity observ'd in the births of both sexes. *Philos Trans R Soc London* 1710;27:186–190.

7. Fisher, RA. The Design of Experiments. Edinburgh: Oliver & Boyd; 1935.
8. Cliff N. Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behav Res* 1996;31:331–350.
9. Cliff, N. Ordinal Methods for Behavioral Data Analysis. Mahwah, NJ: Lawrence Erlbaum; 1996.
10. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1954;1:80–83.
11. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947;18:50–60.
12. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 1952;47:583–631.
13. Lewis, CT.; Short, C. A Latin Dictionary. Oxford: Clarendon; 1879.
14. Georges, KE. Ausführliches lateinisch-deutsches Handwörterbuch. Hannover: Hahn; 1918.
15. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001). *Proc Natl Acad Sci U S A* 2001;98:10515–10515.
16. van de Wiel MA. Significance Analysis of Microarrays using Rank Scores. *Kwantitatieve Methoden* 2004;71:25–37.
17. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57–63. [PubMed: 19015660]
18. McNemar Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* 1947;12:153–157.
19. Gauss, CF. *Theoria combinationis observationum erroribus minimis obnoxiae*. Goettingen: Dieterich; 1823.
20. Coakley CW, Heise MA. Versions of the sign test in the presence of ties. *Biometrics* 1996;52:1242–1251.
21. Dixon WJ, Mood AM. The statistical sign test. *J Am Stat Assoc* 1946;41:557–566. [PubMed: 20279351]
22. Dixon, WJ.; Massey, FJJ. *An Introduction to Statistical Analysis*. New York, NY: McGraw-Hill; 1951.
23. Rayner JCW, Best DJ. Modelling Ties in the Sign Test. *Biometrics* 1999;55:663–665. [PubMed: 11318231]
24. Rao PV, Kupper LL. Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. *J Am Stat Assoc* 1967;62:194–204.
25. David, HA. *The Method of Paired Comparisons*. 2nd ed.. London: Griffin; 1988.
26. Stern HAL. A continuum of paired comparisons models. *Biometrika* 1990;77:265–273.
27. Yan T, Yang YN, Cheng X, DeAngelis MM, Hoh J, Zhang H. Genotypic association analysis using discordant-relative-pairs. *Ann Hum Genet* 2009;73:84–94. [PubMed: 19040657]
28. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–516. [PubMed: 8447318]
29. Wittkowski KM. Versions of the sign test in the presence of ties. *Biometrics* 1998;54:789–791.
30. Wittkowski KM. An asymptotic UMP sign test for discretized data. *The Statistician* 1989;38:93–96.
31. Wittkowski KM, Liu X. A statistically valid alternative to the TDT. *Hum Hered* 2002;54:157–164. [PubMed: 12626848]
32. Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997;53:1253–1261. [PubMed: 9423247]
33. Wittkowski KM. Friedman-type statistics and consistent multiple comparisons for unbalanced designs. *J Am Stat Assoc* 1988;83:1163–1170.
34. Student. On the probable error of a mean. *Biometrika* 1908;6:1–25.
35. Ramagopalan SV, McMahon R, Dymant DA, Sadovnick AD, Ebers GC, Wittkowski KM. An extension to a statistical approach for family based association studies provides insights into genetic risk factors for multiple sclerosis in the HLA-DRB1 gene. *BMC Medical Genetics* 2009;10:10. [PubMed: 19193207]

36. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PIW, Gabriel SB, Mirel DB, Iverson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL. Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 2007;357:851–862. [PubMed: 17660530]
37. Barcellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G, Briggs F, Cree BC, Begovich AB, Villoslada P, Montalban X, Uccelli A, Savettieri G, Lincoln RR, DeLoa C, Haines JL, Pericak-Vance MA, Compston A, Hauser SL, Oksenberg JR. Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. *Hum Mol Genet* 2006;15:2813–2824. [PubMed: 16905561]
38. Ramagopalan SV, Ebers GC. Multiple sclerosis: major histocompatibility complexity and antigen presentation. *Genome Med* 2009;1:105. [PubMed: 19895714]
39. Suárez-Fariñas M, Haider A, Wittkowski KM. “Harshlighting” small blemishes on microarrays. *BMC Bioinformatics* 2005;6:65. [PubMed: 15784152]
40. Suarez-Farinas M, Pellegrino M, Wittkowski KM, Magnasco MO. Harshlight: a “corrective make-up” program for microarray chips. *BMC Bioinformatics* 2005;6:294. [PubMed: 16336691]
41. Arteaga-Salas JM, Harrison AP, Upton GJG. Reducing spatial flaws in oligonucleotide arrays by using neighborhood information. *Stat Appl Genet Mol Biol* 2008;7:19.
42. Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJG, Harrison AP. An overview of image-processing methods for Affymetrix GeneChips. *Brief Bioinform* 2008;9:25–33. [PubMed: 18057073]
43. Cairns JM, Dunning MJ, Ritchie ME, Russell R, Lynch AG. BASH: a tool for managing BeadArray spatial artefacts. *Bioinformatics* 2008;24:2921–2922. [PubMed: 18953044]
44. Deuchler G. Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. *Z pädagog Psychol* 1914;15:114–131. 145–159, 229–242.
45. Morales JF, Song T, Auerbach AD, Wittkowski KM. Phenotyping genetic diseases using an extension of μ -scores for multivariate data. *Stat Appl Genet Mol Biol* 2008;7:19.
46. Kehoe JF, Cliff N. Interord: a computer-interactive Fortran IV program for developing simple orders. *Educ Psychol Meas* 1975;35:675–678.
47. Kruskal WH. Historical notes on the Wilcoxon unpaired two-sample test. *J Am Stat Assoc* 1957;52:356–360.
48. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 1937;32:675–701.
49. Iain, M.; Urken, AB. On elections by ballot. In: Iain, M.; Urken, AB., editors. *Classics of Social Choice*. Ann Arbor, MI: University of Michigan Press; 1995. p. 83-89.
50. Hägerle G, Puckelsheim F. Lull's writings on electoral systems. *Stud Lulliana* 2001;41:3–38.
51. Benard A, Van Elteren PH. A Generalization of the Method of m Rankings. *Indagationes Mathematicae* 1953;15:358–369.
52. van Elteren P, Noether GE. The asymptotic efficiency of the χ^2 -test for a balanced incomplete block design. *Biometrika* 1959;46:475–477.
53. Durbin J. Incomplete blocks in ranking experiments. *Br J Psychol* 1951;4:85–90.
54. Bradley RA, Milton ET. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 1952;39:324–345.
55. Prentice MJ. On the problem of m incomplete rankings. *Biometrika* 1979;66:167–170.
56. Alvo M, Cabilio P. General scores statistics on ranks in the analysis of unbalanced designs. *Can J Stat* 2005;33:115–129.
57. Gao X, Alvo M. A unified nonparametric approach for unbalanced factorial designs. *Journal of the American Statistical Association* 2005;100:926–941.
58. Lam FC, Longnecker MT. A modified Wilcoxon rank sum test for paired data. *Biometrika* 1983;70:510–513.
59. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;52:281–302. [PubMed: 13245896]
60. Popper, KR. *Logik der Forschung*. Wien: Julius Springer; 1937.
61. Delbecq, A. *Group techniques for program planning*. Glenview, IL: Scott Foresman; 1975.

62. Wittkowski KM, Song T, Anderson K, Daniels JE. U-Scores for Multivariate Data in Sports. *J Quant Anal Sports* 2008;4:7.
63. Freimer N, Sabatti C. The human phenome project. *Nat Genet* 2003;34:15–21. [PubMed: 12721547]
64. Wittkowski, KM. Institut für Medizinische Statistik. Göttingen, D.: Georg-August-Universität; 1980. Ein nichtparametrischer Test im Stufenblockplan [A nonparametric test for the step-down design; p. 87]
65. Wittkowski, KM. Semiquantitative Merkmale in der nichtparametrischen Statistik. In: Köhler, CO.; Wagner, E.; Tautu, P., editors. *Der Beitrag der Informationsverarbeitung zum Fortschritt der Medizin*. Berlin, D.: Springer; 1984. p. 100-105.
66. Wittkowski KM. Small sample properties of rank tests for incomplete unbalanced designs. *Biom J* 1988;30:799–808.
67. Wittkowski KM. An extension to Wittkowski. *J Am Stat Assoc* 1992;87:258.
68. Einsele H, Ehninger G, Hebart H, Wittkowski KM, Schuler U, Jahn G, Mackes P, Herter M, Klingebiel T, Löffler J, et al. Polymerase chain reaction monitoring reduces the incidence of cytomegalovirus disease and the duration and side effects of antiviral therapy after bone marrow transplantation. *Blood* 1995;86:2815–2820. [PubMed: 7670117]
69. Talaat, M.; Wittkowski, KM.; Husein, MH.; Barakat, R. A new procedure to access individual risk of exposure to cercariae from multivariate questionnaire data. In: Barlow, R.; Brown, JW., editors. *Reproductive Health and Infectious Diseases in the Middle East*. Aldershot, U.K.: Ashgate; 1998. p. 167-174.
70. Susser E, Desvarieux M, Wittkowski KM. Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. *Am J Public Health* 1998;88:671–674. [PubMed: 9551017]
71. Wittkowski KM, Susser E, Dietz K. The protective effect of condoms and nonoxynol-9 against HIV infection. *Am J Public Health* 1998;88:590–596. 972. [PubMed: 9551000]
72. Banchereau J, Palucka AK, Dhodapkar M, Kurkeholder S, Taquet N, Rolland A, Taquet S, Coquery S, Wittkowski KM, Bhardwaj N, Pineiro L, Steinman R, Fay J. Immune and clinical responses after vaccination of patients with metastatic melanoma with CD34+ hematopoietic progenitor-derived dendritic cells. *Cancer Res* 2001;61:6451–6458. [PubMed: 11522640]
73. Hoeffding W. A class of statistics with asymptotically normal distribution. *Ann Math Stat* 1948;19:293–325.
74. Wittkowski KM. Novel Methods for Multivariate Ordinal Data applied to Genetic Diplotypes, Genomic Pathways, Risk Profiles, and Pattern Similarity. *Comp Sci Stat* 2003;35:626–646.
75. Wittkowski KM, Liu X. Beyond the TDT: Rejoinder to Ewens and Spielman. *Hum Hered* 2004;58:60–61.
76. Wittkowski KM, Lee E, Nussbaum R, Chamian FN, Krueger JG. Combining several ordinal measures in clinical studies. *Stat Med* 2004;23:1579–1592. [PubMed: 15122738]
77. Gehan EA. A generalised two-sample Wilcoxon test for doubly censored samples. *Biometrika* 1965;52:650–653. [PubMed: 5858975]
78. Gehan EA. A generalised Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965;52:203–223. [PubMed: 14341275]
79. Schemper M. A nonparametric k-sample test for data defined by intervals. *Statistica Neerlandica* 1983;37:69–71.
80. Lehmann EL. Consistency and unbiasedness of certain nonparametric tests. *Ann Math Stat* 1951;22:165–179.
81. Hoeffding, W. *The Collected Works of Wassily Hoeffding*. New York, NY: Springer; 1994.
82. Rosenbaum PG. Coherence in observational studies. *Biometrics* 1994;50:368–374. [PubMed: 8068837]
83. Song T, Coffran C, Wittkowski KM. Screening for gene expression profiles and epistasis between diplotypes with S-Plus on a grid. *Stat Comput Graph* 2007;18:20–25.
84. Cherchye L, Vermeulen F. Robust rankings of multidimensional performances: an application to Tour de France racing cyclists. *J Sports Econ* 2006;7:359–373.

85. Quaia E, D'Onofrio M, Cabassa P, Vecchiato F, Caffarri S, Pittiani F, Wittkowski KM, Cova MA. Diagnostic value of hepatocellular nodule vascularity after microbubble injection for characterizing malignancy in patients with cirrhosis. *Am J Roentgenol* 2007;189:1474–1483. [PubMed: 18029888]
86. Ramamoorthi RV, Rossano MG, Paneth N, Gardiner JC, Diamond MP, Puscheck E, Daly DC, Potter RC, Wirth JJ. An application of multivariate ranks to assess effects from combining factors: Metal exposures and semen analysis outcomes. *Stat Med* 2008;27:3503–3514. [PubMed: 18314933]
87. Shockley W, Bardeen J, Brattain WH. The electronic theory of the transistor. *Science* 1948;108:678–679.
88. Häberle L, Pfahlberg A, Gefeller O. Assessment of Multiple Ordinal Endpoints. *Biom J* 2009;51:217–226. [PubMed: 19197963]
89. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984;40:1079–1087. [PubMed: 6534410]
90. Diana M, Song T, Wittkowski K. Studying travel-related individual assessments and desires by combining hierarchically structured ordinal variables. *Transportation* 2009;36:187–206. [PubMed: 20953273]
91. Kendall MG. A new measure of rank correlation. *Biometrika* 1938;30:81–93.
92. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. *Biometrika* 1954;41:133–145.
93. Terpstra TJ. The asymptotic normality and consistency of Kendall's test against trend when ties are present in one ranking. *Indagationes Mathematicae* 1952;14:327–333.
94. Spangler R, Wittkowski KM, Goddard NL, Avena NM, Hoebel BG, Leibowitz SF. Opiate-like Effects of Sugar on Gene Expression in Reward Areas of the Rat Brain. *Mol Brain Res* 2004;124:134–142. [PubMed: 15135221]
95. Wittkowski KM, Seybold MP, Schneider EM. Neue nicht-parametrische Methoden und Tools für die Auswertung multivariater Daten in der Klinischen Forschung und Diagnostik. *Dtsch Z Klin Forsch* 2008;12:22–27.
96. Morales JF, Song T, Wittkowski KM, Auerbach AD. A statistical systems biology approach to FANCC gene expression suggests drug targets for Fanconi anemia. (submitted).
97. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11:375–386.
98. Lehmann EL. Some model I problems of selection. *Ann Math Stat* 1961;32:990–1012.
99. Seybold MP, Wittkowski KM, Schneider EM. Biomarker analysis using a non-parametric selection procedure to discriminate the phagocytic syndromes HLH (hemophagocytic lymphohistiocytosis) and mas (macrophage activation syndrome). *Shock* 2008;29:90–90.
100. Kraft P, Hunter DJ. Genetic risk prediction -- Are we there yet? *N Engl J Med* 2009;360:1701–1703. [PubMed: 19369656]
101. Wittkowski KM. Statistical knowledge-based systems — critical remarks and requirements for approval. *Computer Methods and Programs in Biomedicine* 1990;33:255–259. [PubMed: 2282791]
102. Akritas MG, Arnold SF, Brunner E. Nonparametric hypotheses and rank statistics for unbalanced factorial designs, Part I. *J Am Stat Assoc* 1997;92:258–265.
103. Brunner E, Munzel U, Puri ML. Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis* 1999;70:286–317.

Fig. 2.1. Binomial distribution for $n = 10$ under the hypothesis $p = 0.5$

Fig. 2.2. Triangular distribution, discretized at zero with bandwidth $b=0.5$

Fig. 2.3. Cumulative distribution function for the example of Table 2.2
TDT: dashed lines and open symbols, SMN: solid lines and full symbols.

Fig. 2.4. Extension of the stratified McNemar to multi-allelic loci identifies amino acid 13 (at the center of the P4 pocket) as a major risk factor for multiple sclerosis
(modified from *BMC Medical Genetics*, Ramagopalan *et al.* available from <http://www.biomedcentral.com/1471-2350/10/10/> © 2009, with permission from BioMed Central Ltd).

Fig. 2.5. Quality control images generated by Harshlight for Affymetrix chips (top, modified from *BMC Bioinformatics*[39,40] © 2005, with permission from BioMed Central Ltd) and by the “Bead Array Subversion of Harshlight” (bottom, modified from *Bioinformatics*[43] © 2008, with permission from Oxford University Press)

[illegible]

Fig. 2.6. Implementation of the conditional and unconditional sign tests as part of the “muStat” package for both R and S-PLUS
 SMN: stratified $M_{C/NEMAR}$, TDT: transmission disequilibrium test, MCN: $M_{C/NEMAR}$, DMM: $D_{XON/MOOD/MASSEY}$. The function `MC()` avoids scoping problems in S-PLUS and is part of the compatibility package `muS2R`.

	Y1	Y2
A	3	2
B	3	1
C	2	1
D	2	1

Y1	3	3	2	2
3	0	0	1	1
3	0	0	1	1
2	-1	-1	0	0
2	-1	-1	0	0

Y2	2	1	1	1
2	0	1	1	1
1	-1	0	0	0
1	-1	0	0	0
1	-1	0	0	0

Fig. 2.7. Matrices of paired comparisons for two variables (Y1, Y2), each observed in four subjects (modified from *Stat Appl Genet Mol Biol*, Morales *et al.*[45] available at <http://www.bepress.com/sagmb/vol7/iss1/art19/>, with permission from The Berkeley Electronic Press, © 2008).

```

#-----
# mu.test          (y, groups, blocks, ...) # most general form
# mu.friedman.test (y, groups, blocks, ...) # one observation per cell
# mu.kruskal.test  (y, groups, ...)        # single block
# mu.wilcox.test   (y, groups, ...)        # single block/two groups
#-----
# y,                # data (NA allowed)
# groups,           # groups (unbalanced allowed)
# blocks,           # blocks (unequal sz allowed)
# score            = "rank",               # NULL: y already scored
# paired           = FALSE,               # wilcox only
# exact            = NULL,                # wilcox only
# optim            = TRUE                  # optimize for special cases
# df               = -1,                   # >0: as entered, =0: by design, -1: by data
#-----
mu.test          <- function(y, groups, blocks, ...)
mu.friedman.test <- function(y, groups, blocks, ...)
mu.wilcox.test   <- function(y, groups, blocks = NULL, paired = F, ...)
mu.kruskal.test  <- function(y, groups, blocks = NULL, ...)
...
ok <- is.orderable(y)
mi <- rowSums(m <- xTable(blocks, groups)) # exp. block size
Mi <- rowSums(M <- xTable(blocks[ok], groups[ok])) # obs. block size
wi <- mi+1 # Prentice weights
Tijk <- Centered(
  Score(applyBy(y, blocks, wRank)/(Mi[blocks]+1),
    blocks, mi)
  * wi[blocks])
T1 <- qapply(Tijk, groups, sum)
V0 <- structure(dim=c(P,P),
  (1/(Mi-1))*qapply(Tijk^2, blocks, sum) %*% ( # blkvar
    t(apply(M,1, function(x,P) diag(x))) - # design
    (1/Mi) * t(apply(M,1, function(x) outer(x,x))) ) # matrix )
W <- T1 %*% ginv(V0) %*% T1

```

Fig. 2.8. Implementation of (conditional) rank tests within the “muStat” package for both R and S-PLUS

The notation corresponds directly to the details published in 1988[33], as given in Equations (2.8) and (2.9). The code has been simplified for clarity. See the muStat package for details and for definitions of the functions xTable, applyBy, and qapply, which are optimized variance of the standard functions table and apply.

	X1	X2
A	1	2
B	2	2
C	?	1
D	1	2

X1	X2	censored	U
1	2	? -1 1 ?	0
2	2	1 0 1 1	3
?	1	-1 -1 ? -1	-3
1	2	? -1 1 ?	0

Fig. 2.9. Matrix of paired comparisons a single interval censored data observed in four subjects (modified from *Stat Appl Genet Mol Bio*, Morales *et al.*[45] available at <http://www.bepress.com/sagmb/vol7/iss1/art19/>, with permission from The Berkeley Electronic Press, © 2008).

Fig. 2.10. Orderings: a) simple, b) exact, c) inexact, d) interval, e) multivariate (modified from *Comp Sci Stat*, Wittkowski[74], available from <http://ideas.repec.org/p/pramprapa/4570.html> with permission from the Interface Foundation of North America, Inc, and *Statistics in Medicine*, Wittkowski *et al.*[76] © 2004, with permission from John Wiley & Sons, Inc.)

Fig. 2.11. Ambiguity caused by discordant paired comparisons across variables

Hypothetical example with four variables X1, X2, Y1, and Y2, observed in four subjects A, B, C, and D. The node on top shows the data, the center row the four univariate partial (X1) and complete (X2, Y1, Y2) orderings, The node at the bottom shows the data, the multivariate partial ordering, the μ -scores (U) and their information content (W). Each matrix shows whether the row element is smaller (-1), identical (0), or larger (1) than the column element. Ambiguous paired comparisons are indicated as '?'. Resolving ambiguities (such as those in X1) if the corresponding paired comparisons in the other matrices are unambiguous increases information content, but may result in non-transitivity and, thus, an imperfect order. On the other hand, ambiguities can arise (such as in the upper right and lower left corner of the bottom matrix) if some paired comparisons are negative, while others are positive (modified from *Stat Appl Genet Mol Biol*, Morales *et al.*[45] available at <http://www.bepress.com/sagmb/vol7/iss1/art19/>, with permission from The Berkeley Electronic Press, © 2008).

```

mu.PwO <- function(x, y=x)
if (length(y)>1) apply(rbind(x,y),2,mu.PwO,nrow(x))
else as.numeric(NAtoZer(outer(x[1:y],x[-(1:y)],">=")))

mu.AND <- function(PwO, frm1=NULL)
if (is.null(frm1)) {
  GE <- sq.array(PwO); AND <- GE[,1]^0; nNA <- AND[,1]*0
  for (i in 1:dim(GE)[3]) {
    nNA <- nNA + diag(GEi <- GE[,i])
    AND <- AND * (GEi + (1-GEi)*(1-t(GEi))) }
  return(as.numeric(AND * ((c(nNA) %o% c(nNA)) > 0))) }
else { # ... deal with the formula ... (code omitted)
  return(...) }

mu.Sums <- function(PwO) {
  GE <- sq.matrix(PwO)
  wght <- colSums(GE|t(GE))
  list(score = (rowSums(GE) - colSums(GE)) *ifelse(wght==0,NA,1),
        weight = wght) }

```

Fig. 2.12. Simplified code for the core functions of the muStat packages

mu.PwO creates a representation of a partial ordering, mu.AND combines several partial orderings, and mu.Sums evaluates a partial ordering.

Fig. 2.13. Resolving ambiguity from discordant paired comparisons

Using hierarchical structure in the example of Fig. 2.11. Note that adding the intermediate step resolves the ambiguity in the lower left and upper right corner of the matrices representing the partial orderings (modified from *Stat Appl Genet Mol Biol*, Morales *et al.*[45] available at <http://www.bepress.com/sagmb/vol7/iss1/art19/>, with permission from The Berkeley Electronic Press, © 2008).

Fig. 2.14. Reflecting censoring during the creation of matrices of paired comparisons
Subjects A and B can be ordered if $X1(A) > X2(B)$ or $Y2(A) < Y1(B)$, unless one observation is uncensored, in which case “ $\geq$ ” or “ $\leq$ ” suffice. (modified from *Stat Appl Genet Mol Biol*, Morales *et al.*[45] available at <http://www.bepress.com/sagmb/vol7/iss1/art19/>, with permission from The Berkeley Electronic Press, © 2008)

Fig. 2.15. Doubly interval censored data as a special case of hierarchically structured interval censored data

O_0 and O_1 indicate the last negative and first positive date of onset, respectively, while R_0 and R_1 indicate the last negative and first positive date of response, respectively. When neither the onset nor the response ranges overlap, objects A and B can be ordered.

Fig. 2.16. Hypothetical relative distance of disease loci to a single ‘proximal’ SNP or a pair of “distal” SNPs and corresponding muStat formulae (frml)

Top: conceptual interval structure. Alternating intervals around a SNP and between two adjacent SNPs (first row), diplotype ranges of various length (second and third row). Bottom: muStat formulae. To conserve space, single-SNP intervals at the end of a diplotype, which may or may not increase association are not capitalized. Note that a given SNP has either always positive or always negative polarity within a diplotype.

Fig. 2.17. Epistasis between Diplotypes

Left: Sequence of hierarchical (⊕) and combinatorial (&) steps. Right: corresponding mu.AND formulae. (See Fig. 2.16 for details.)

Fig. 2.18. Coregulation vs Collaboration

Dopaminergic activity (D1/D2/D3: dopamine 1/2/3, pE: preproenkephalin, pD: prodynorphin, pT: preprotachykinin) in the brain of rats addicted to sugar (closed circles) vs control rats (open circles). Top: *nucleus accumbens* (NAC), Bottom: *caudatus putamen* (CPU) (Modified from *DZKF*, Wittkowski *et al.*[95], with permission from Geelen & Geelen, © 2008)

Fig. 2.19. Collaboration between pairs and trios of genes against a background of genes clustered by pairwise correlation indicate

AURKA (top) and RRM2 (bottom) are identified through gene–gene interactions as relevant for FANCC. Gene ontology information suggests that subjects may differ with respect to how three major pathways (signal transduction, repair/recombination/replication, protein metabolism/transcription) are involved. Background: MDS (multidimensional scaling) map (with $\sqrt{1-r^2}$ as distance) showing genes clustered by coregulation and significance of the univariate p-values (size of nodes). Foreground: The network indicates the “most significant” pairs (top) and trios (bottom) of genes with font size indicating the smallest of the uni-, and oligo-variate p-values[96].

Fig. 2.20. GWAS comparing 132 children with Childhood Absence Epilepsy to three sets of differently matched normal controls

Data: $-\log_{10}(\text{p-value})$ C2: max chi-square (recessive, trend, dominant), 1...7: μ -test by length of diplotype. Shading: >3 (C2), $>4/5/6$ (μ -test).

Fig. 2.21.
Heatmap of correlations between atherosclerosis and μ -scores of combinations of diplotypes on mouse chromosomes 8–Y (Left: individual SNPs; Right: diplotypes of up to 4 SNPs)

Fig. 2.22. Reflecting a hierarchical structure among variables increases information content and improves discrimination

Hierarchical data structure of the Fanconi anemia phenotype (top), hierarchy increasing information content (center), and discriminatory power (bottom)[45] (middle and bottom modified from *Stat Appl Genet Mol Biol*, Morales *et al.*[45] available at <http://www.bepress.com/sagmb/vol7/iss1/art19/>, with permission from The Berkeley Electronic Press, © 2008)

Fig. 2.23. Model selection by Information content ($C3 > C1 > NH$)

Top: Subjective Mobility (SM), Bottom: Relative Desired Mobility (RDM). Inserts of correlations between non-hierarchical (NH) and either C1 or C3 hierarchical scores show two categories of subjects, especially with respect to RDM[90]. (modified from *Transportation*, Diana *et al.*[45], with permission from Springer, © 2009)

Table 2.1Expected estimates and test statistics ($n=25$) by bandwidth b

b	$p_{=}$	p_{+}	$\frac{p_{+}}{1-p_{=}}$	T_{25}	$\frac{p_{+}}{1-p_{=}}$	$\frac{p_{+}}{1-p_{=}}$	T_{25}^{*}	$\frac{T_{25}^{*}}{T_{25}^{*}}$
0.00	0.00	0.50	0.50	0.00	0.50	0.00	0.00	1.0
0.25	0.21	0.39	0.49	0.06	0.49	0.05	0.05	1.1
0.50	0.41	0.27	0.46	0.28	0.48	0.21	0.21	1.3
0.75	0.62	0.14	0.37	0.78	0.45	0.48	0.48	1.6
0.90	0.75	0.06	0.23	1.38	0.43	0.69	0.69	2.0
0.95	0.79	0.03	0.14	1.68	0.42	0.77	0.77	2.2
0.99	0.82	0.01	0.03	1.98	0.42	0.84	0.84	2.4
1.00	0.83	0	0	2.07	0.41	0.86	0.86	2.4
1.50	0.93	0	0	2.34	0.46	0.36	0.36	3.7
>2.41	1.00	0	-	-	0.50	0.00	0.00	8.2

Table 2.2
Exact distribution of the TDT (top) and SMN (bottom) for a sample of three children.

g_1	g_2	g_3	b	p_1	p_2	p_3	f	n_P	n_Q	$n_P - n_Q$	$n_P + n_Q$	x	f
PP	PP	PP	1	0.25	0.25	0.25	0.016	6	0	6	6	2.449	0.016
PP	PP	PQ	3	0.25	0.25	0.50	0.094	5	1	4	6	1.633	0.094
PP	PQ	PQ	3	0.25	0.50	0.50	0.188	4	2	2	6	0.816	0.188
PP	PP	QQ	3	0.25	0.25	0.25	0.047	4	2	2	6	0.816	0.047
PQ	PQ	PQ	1	0.50	0.50	0.50	0.125	3	3	0	6	0.000	0.125
PP	PQ	QQ	6	0.25	0.50	0.25	0.188	3	3	0	6	0.000	0.188
PP	QQ	QQ	3	0.25	0.25	0.25	0.047	2	4	-2	6	-0.816	0.047
PQ	PQ	QQ	3	0.50	0.50	0.25	0.188	2	4	-2	6	-0.816	0.188
PQ	QQ	QQ	3	0.50	0.25	0.25	0.094	1	5	-4	6	-1.633	0.094
QQ	QQ	QQ	1	0.25	0.25	0.25	0.016	0	6	-6	6	-2.449	0.016
							1.000						
							0.000	1.000					

g_1	g_2	g_3	b	p_1	p_2	p_3	f	n_{PP}	n_{QQ}	$n_{PP} - n_{QQ}$	$n_{PP} + n_{QQ}$	x	f
PP	PP	PP	1	0.25	0.25	0.25	0.016	3	0	3	3	1.732	0.018
PP	PP	QQ	3	0.25	0.25	0.25	0.047	2	1	1	3	0.577	0.054
PP	QQ	QQ	3	0.25	0.25	0.25	0.047	1	2	-1	3	-0.577	0.054
QQ	QQ	QQ	1	0.25	0.25	0.25	0.016	0	3	-3	3	-1.732	0.018
PP	PP	PQ	3	0.25	0.25	0.50	0.094	2	0	2	2	1.414	0.107
PP	PQ	QQ	6	0.25	0.50	0.25	0.188	1	1	0	2	0.000	0.214
PQ	QQ	QQ	3	0.50	0.25	0.25	0.094	0	2	-2	2	-1.414	0.107
PP	PQ	PQ	3	0.25	0.50	0.50	0.188	1	0	1	1	1.000	0.214
PQ	PQ	QQ	3	0.50	0.50	0.25	0.188	0	1	-1	1	-1.000	0.214
PQ	PQ	PQ	1	0.50	0.50	0.50	0.075	0	0	0	0	0	0
							0.875						
							0.000	1.000					

Legend: g_i child i genotype; b : number of combinations; p_i : $E_Q(g_i)$; f : $b \times \Pi p_i$; n_X/n_{XY} : number of alleles/children; x : $(n_P - n_Q) / \sqrt{n_P + n_Q} / (n_{PP} - n_{QQ}) / \sqrt{n_{PP} + n_{QQ}}$

Table 2.3
Nucleotides (N#: nucleotide number, #A: amino acid number) found in 94 HLA-DRB1 SNPs discriminating the 13 main two-digit allelic groups
The nucleotide found in the HLA-DRB1*15 (15) allele, is highlighted across allelic groups (modified from *BMC Medical Genetics*, Ramagopalan *et al.*[35] available from <http://www.biomedcentral.com/1471-2350/10/10/> © 2009, with permission from BioMed Central Ltd).

Allelotype:		02	01	04	03	05			06	08	10	09	07	
Allele	Nucleotide	15	16		17/18	11	12	13	14					
-1	85	T	G	G	G	G	G	G	G	G	G	G	G	
	124	A	A	T	C	G	T	G	TG	G	T	T	T	
	125	G	G	T	A	C	C	G	CG	G	T	T	A	
60	126	G	G	T	T	T	T	T	T	T	T	T	T	
	265	T	T	T	T	T	T	T	CT	T	T	T	T	
	266	A	A	A	A	A	AC	AC	A	A	A	C	C	
96	373	C	C	G	T	C	C	C	C	C	C	C	C	
	375	A	A	G	T	T	T	T	T	T	A	T	T	
	390	A	A	G	G	G	G	G	G	G	G	G	G	
133	485	T	T	G	G	G	G	G	G	G	G	G	G	
142	511	A	A	G	G	G	G	G	G	G	G	G	G	
166	584	G	G	G	G	G	G	G	G	G	A	G	G	
	585	A	A	G	G	G	G	G	G	G	G	G	G	
#Cases:		3994	208	1396	1982	1873	1068	146	1519	243	408	80	98	1653