New Statistical Paradigms Leading to Web-Based Tools for Clinical/Translational Science

Knut M. Wittkowski

The Rockefeller University, General Clinical Research Center, Department of Experimental Design and Biostatistics
1230 York Ave Box 322, New York, NY 10021, U.S.A

1 Introduction

As the field of functional genetics and genomics is beginning to mature, we become confronted with new challenges. The constant drop in price for sequencing and gene expression profiling as well as the increasing number of genetic and genomic variables that can be measured makes it feasible to address more complex questions. The success with rare diseases caused by single loci or genes has provided us with a proof-of-concept that new therapies can be developed based on functional genomics and genetics.

Common diseases, however, typically involve genetic epistasis, genomic pathways, and proteomic pattern. Moreover, to better understand the underlying biological systems, we often need to integrate information from several of these sources. Thus, as the field of clinical research moves toward complex diseases, the demand for modern data base systems and advanced statistical methods increases. At the same time, as with many emerging fields, better understanding of the underlying concepts allows for similarities with other, more established fields to be revealed and, thus, some of the techniques developed earlier to be revisited.

As biological systems are often controlled by a variety of regulatory feedback loops, many of which may be unknown, the assumption that the functional form of the relationship between a measurement and activity or efficacy is known may not be valid, except, at best, within very narrowly controlled experimental settings. Since twenty years ago (Wittkowski 1980), new **non-parametric methods** have been developed to avoid artifacts created by using methods based on unrealistic assumptions. New bioinformatics tools now help to make these methods more widely available. The earlier work introduced the marginal likelihood principle (MrgL) as a technique to extend rank tests to partially ordered univariate data, in general, and missing data, in particular (Wittkowski 1980; 1984; 1988b; 1988e).

From extensive consulting experience, it soon was realized that few biological systems can be sufficiently characterized by a single variable only. In 1992, rank tests for censored data were generalized to **multivariate ordinal observations** (Wittkowski 1992a). While this approach proved eminently useful (Einsele, Ehninger 1995; Susser, Desvarieux 1998; Talaat, Wittkowski 1998; Wittkowski, Susser 1998), the computational effort that comes with the MrgL principle was prohibitive. Only after drawing on the analogy of the Mann-Whitney (u statistics) and the Wilcoxon test (MrgL), a more computationally efficient approach became available, this time based on u-statistics (Hoeffding 1948) and a computational strategy that had been earlier devised in Tübingen (Deuchler 1914), see, e.g., (King, Jim 2003)

Recently, we have begun to extend this approach by allowing for more complex designs in genetics (Wittkowski, Liu 2002; Wittkowski, Liu 2004) and more problem-specific partial orderings (Wittkowski 2003). The first **Web tools** based on these results are now available (muStat.rockefeller.edu). For small data sets, spreadsheets can be downloaded, for larger data sets, we are moving parallelized computational services from a cluster to a grid.

As the complexity of experimental designs and the choice of statistical methods increases, so does the need for data management and decision support. The PANOS system (Wittkowski 1985) provided for the first **data model and a knowledge representation concept** able to support the choice of both parametric and non-parametric methods. A subset of these results can be seen in JMP (SAS Institute Inc. 2002). More recently, we developed a similar data model for topological (genetics) and functional relations (genomics) between variables (SNPs and genes, respectively) to be represented.

As with many knowledge based systems, **knowledge acquisition** remained a bottleneck. Drawing upon modern rapid prototyping systems, we have now begun to implement a Web based tool for acquisition of knowledge on the biological background and the experimental design (WISDOM).

2 U Statistics for Changes in Function

Complex diseases are typically characterized by several variables. In Phase I/II studies, were surrogate activity variables need to be considered in lieu of clinical outcomes, it is especially unlikely that a single variable can be found to be sufficient. With traditional linear weight scoring systems, one transforms each variable individually (linear, logarithmic, categorization) to obtain a score on a comparable scale (present / absent, low/intermediate/high, 1 to 10, z-score) and then defines a global score as the linear combination (weighted average) of these individual scores.

Relying on the linear model has advantages. First, algorithms are computationally efficient. Mean and standard deviation, for instance, are easily computed with a pocket calculator, while quartiles, the analogues based on u statistics, are not. Second, the assumption of independent, additive main effects and interactions coerces fitted models whose alluring simplicity often turns out to be an artifact of the linear model not easily allowing for more complex, non-linear relations. Finally, the assumption of independent additive errors yields the convenient bell shaped distribution of errors. The prayer that biology be linear, independent, and additive, however, is rarely answered and the central limit theorem does not provide for a rescue from model misspecification. In particular, the relative importance of the variables, and, thus the weight they need to be assigned, is typically not known *a-priori*.

Since relative importance, correlation, and functional relationship between variables are typically unknown, construct validity (Cronbach, Meehl 1955) cannot be established on theoretical grounds. Instead, researchers often resort to empirical 'validation', choosing weights and functions that provide a reasonable fit with a 'gold standard' when applied to a sample. While this allows for a comparison between studies where researchers agreed on the same scoring system, comparability along a scale with questionable validity may still yield questionable results. The diversity of scoring systems used attests to their subjective nature.

Recently, we have utilized multivariate u statistics to overcome these obstacles, providing the first approach generating clinical scores that are 'intrinsically valid', i.e., that do not need to rely on empirical validation, a process of questionable validity by itself (Popper 1937). While this approach proved eminently useful, we soon realized that further work was needed, because additional information about relationships between the variables often needs to be reflected.

Within the linear model, multivariate data can be reduced to a score by applying a simple transformation such as the average, difference, or ratio. In other cases, one estimates a more complex parameter, like the slope of a regression line or a maximum likelihood estimate. If one does not feel comfortable with the assumptions of linearity and independence underlying the linear model u statistics can provide solutions for closely related questions. For univariate data, the Wilcoxon-Mann-Whitney u test (Mann, Whitney 1947; Wilcoxon 1954), for instance, corresponds to the Student t-test (Student 1908). The question then arises, how to extend u statistics to reflect known relations between variables. Interestingly, one of the first application of multivariate u statistics, the u-test for (interval) censored data (Gehan 1965a; 1965b; Schemper 1983) can be seen as a u-test for bivariate data, if one replaces the natural partial ordering (Wittkowski, Lee 2004, Eq. 1) by a specific ordering, where intervals are ordered, if they are disjoint (non-overlapping).

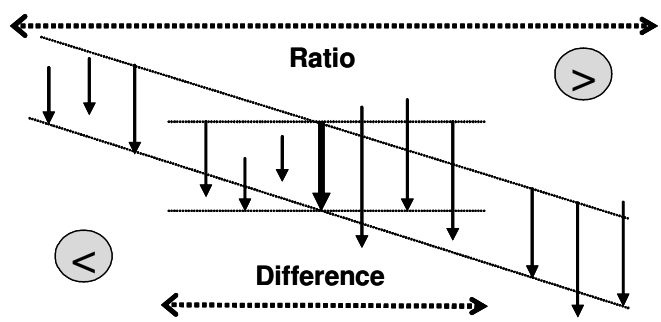

Fig. 1: Changes (indicated by vertical arrows) that can be ordered for differences and ratios. Changes to the left or right of the reference (center, bold) are smaller or larger, respectively, than the reference. By the partial ordering for intervals, in contrast, only the three arrows to the left are higher than the three arrows to the right.

We have recently developed additional partial orderings (see, e.g., Fig. 1). To some extent, differences (e.g., changes over time), are the 'opposite' of intervals, in that they can be unambiguously ordered only if one 'interval' is fully contained in the other, rather than disjoint. Obtaining pair wise orderings of ratios (e.g., changes in concentration) can be obtained under the same conditions, but ratios can also be ordered if the change at the lower level (distance from zero) is at least as large as the change at the higher level.

To better reflect specific problems in genetics and genomics, we will develop more partial orderings, thereby making multivariate u-statistics more widely applicable.

3 U Statistics for Microarrays

3.1 Quality Control

U statistics have various applications in functional genomics. At a low level, quality of data scanned from a microarray (Fig. 2a) can be affected by a plethora of potential confounders, which may act during printing, manufacturing, hybridization, washing, and reading. Given the high probe-to-probe variance and their random allocation on the chip, it is impossible to visually detect all but the starkest artifacts.

As the price for chips drops, a typical experiment now contains several chips, each representing a sample obtained under conditions that were similar except for the experimental factor under investigation. This offers new strategies for testing the effect of the experimental factor, e.g., through 'robust multiarray analysis' (Irizarry, Hobbs 2003). As probes differ in affinity (Naef, Magnasco 2003; Wu, Irizarry 2004), their correlation can also be used to identify small blemishes.

Fig. 2b exhibits a variety of such blemishes. The shadowy circle on the left side, e.g., is clearly an artifact, as are the bright spot in the upper-right corner, and the dark spot in the upper center (Fig. 3). These examples are based on U95 chips with 16 probe pairs per probe set. On U133 chips, with only 11 pairs per set, the effect of artifacts on results is expected to increase.

The choice of the arithmetic mean (average) as the measure of central tendency in linear models relies either on the law of large numbers and the central limit theorem or on the assumption that the distribution of errors at least symmetrical. Here, neither assumption is easily justified. Fig. 4 demonstrates that median filtering, based on u-statistics, causes less 'ghosting' than average filtering, based on the linear model.

A tool to detect blemishes automatically (Fig. 3c) is available at asterion.rockefeller.edu/Harshlight. While the first version (Suárez-Fariñas, Haider 2005) was based on traditional pattern recognition methods, with several parameters to be chosen, we are now working on replacing this linear model approach, too, by a non-parametric approach based on u-statistics.

Having substantially improved quality control of high density oligonucleotide arrays, we will fine-tune the pattern recognition process, to reduce the number of parameters, increase sensitivity for specific types of blemishes, and to allow for more complex, factorial designs.

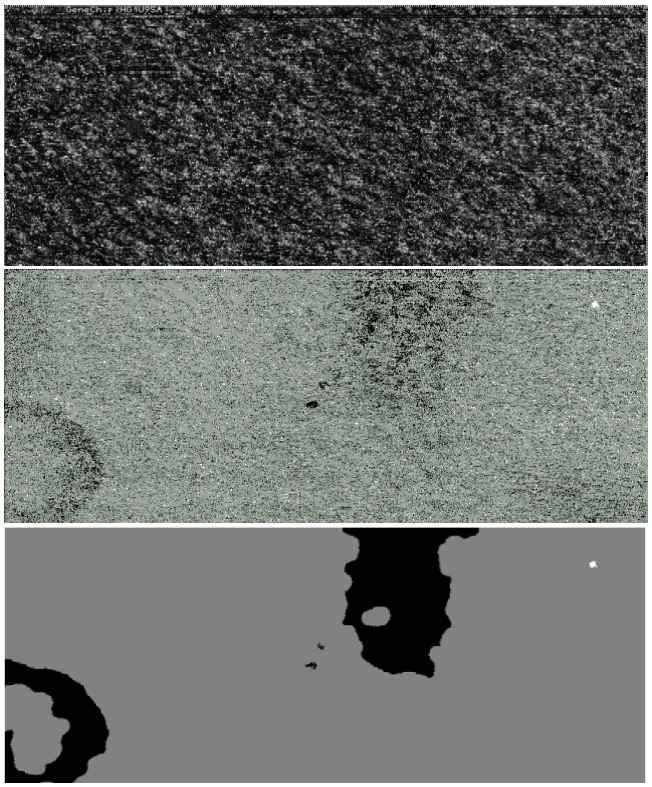

Fig. 2: Top (a): upper 50% pseudo-image of a HuU95av2 chip. Bottom (b): median filtered image (3 chips). (c) HarshLight mask

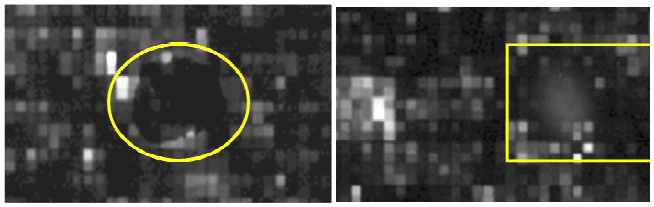

Fig. 3: Left (a): Raw image file detail of the 'dark spot' artifact seen in the center of Fig. 2. Right (b): Raw image file detail of the the 'bright spot' in the top right corner of Fig. 2.

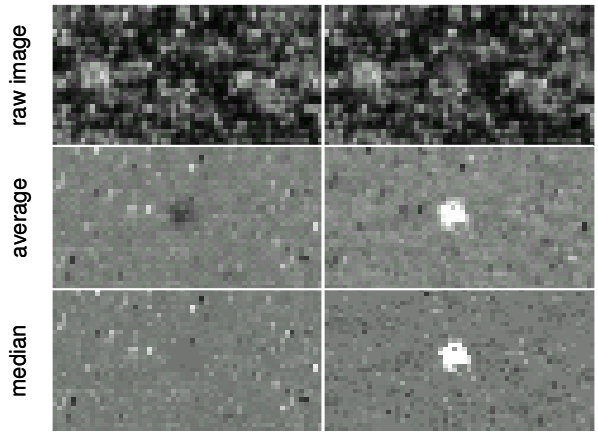

Fig. 4: The artifact of Fig. 3b. Top: raw image from the same area of two chips showing gene expression from the same sample under two experimental conditions. Center: average filtering, bottom: median filtering.

3.2 Signal Value Estimation

To estimate the non-specific portion of the binding on an Affymetrix GeneChip®, each 'perfect match' (PM) is paired with a 'mis-match' (MM), with the middle nucleotide exchanged for its WATSON-CRICK complement. When estimating the signal value for a particular gene from a probe set of pairs of perfect and mismatches, several parametric and semi-parametric ('robust') methods have been proposed.

To allow for a linear model to be based on the logarithms of the differences, it has been suggested (Hubbell, Liu 2002) to artificially decrease x_{MM} of probe pairs where $x_{\mathrm{PM}} < x_{\mathrm{MM}}$ to a heuristically motivated level that ensures that $x_{\mathrm{PM}} - x_{\mathrm{MM}}$ is positive. As a justification for this 'background correction', it is argued that a mismatch should never be higher than a perfect match. For genes that are not expressed in the biological sample, however, one would *expect* 50% of the pairs to have higher mis- than perfect matches. Then, this 'correction' creates a bias, because estimates for genes with expression level zero have signal value estimates as high as genes with low, but positive expression levels, the level of this estimate depending, in part, on the within-probe variance.

Above, the need for various partial orderings has been stressed. When using u statistics, this bias can easily be overcome by employing the following partial ordering for signal value estimation:

$$\{x_k < x_{k'}\} \Leftrightarrow \left\{\left(x_{k,\mathrm{PM}} < x_{k',\mathrm{PM}}\right) \wedge \left(-x_{k,\mathrm{MM}} < -x_{k',\mathrm{MM}}\right)\right\}$$

Within each probe set, one then selects the pair with a score of zero as the most 'typical', or, if necessary, the weighted average among those closest to zero (Fig. 5). As this guarantees 'outliers' to be excluded, the alleged need for taking logarithms has been overcome. Requesting that this estimate be non-negative results in a much smaller bias than decreasing $x_{k,\mathrm{MM}}$ for each pair where $x_{k,\mathrm{PM}} < x_{k,\mathrm{MM}}$ (Haider, Naef 2003).

Fig. 5 shows a spreadsheet implementation of u statistics for probe sets. Fig. 6 and Fig. 7 depict the bias for a chip from the study mentioned earlier and for the 'spike-in' data set (Irizarry, Hobbs 2003), respectively.

> Again, more steps lie ahead. Even after reducing the effect of blemishes, the large differences in affinity between PM probes can render the order of many probe pairs ambiguous. Drawing on the sequence information (Naef, Magnasco 2003), in addition to eliminating blemishes, should further improve the reliability of signal value estimates.

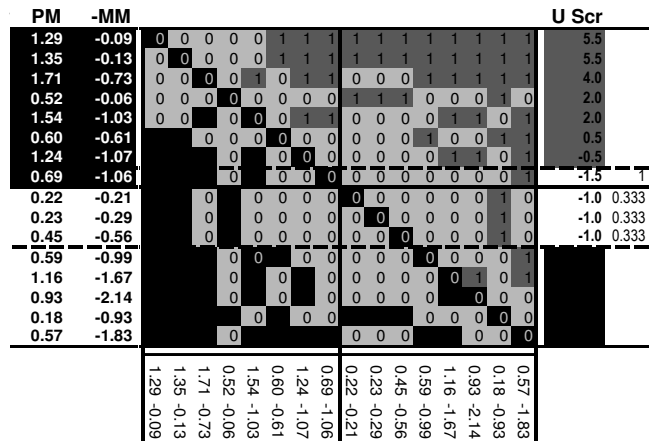

Fig. 5: Signal value estimation with u-statistics. The u scores for probe pairs closest to the median are displayed in white.

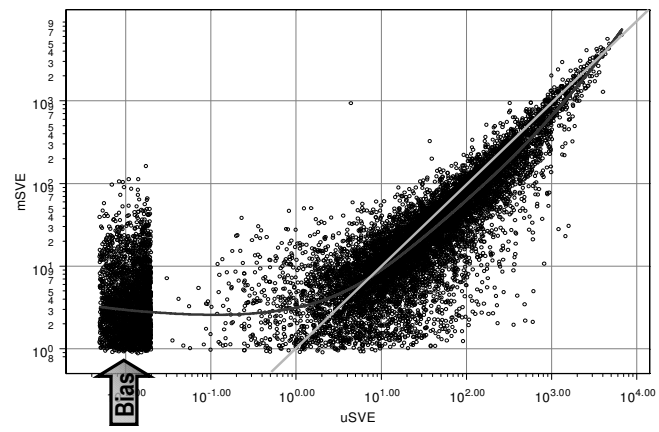

Fig. 6: MAS 5.0 bias for genes with low expression levels. Probes with u statistic signal value estimates ≤0 are scattered around 10^{-2}

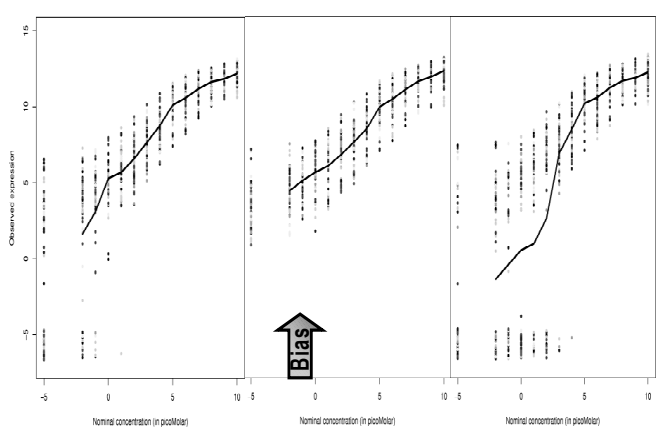

Fig. 7: MAS 4.0, MAS 5.0, and U-statistics – bias vs. variance stabilization, $\log_2$ transformed axes.

3.3 Gene expression profiling

Once a signal value estimate has been obtained for each gene, u statistics for gene expression profiles are conceptually not different from u-statistics for clinical response profiles, although the computational complexity is higher, because neither the subset of relevant genes nor their 'orientation' is known *a-priori*. In one of our first applications (Wittkowski, Lee 2004), we started with scoring clinical response (disease severity) based on several outcomes. As is typical for assessing phenotypes, both the set of variables and their orientation was known. For each outcome (epidermal thickness, K16 histology), it was reasonable to assume that 'more' was 'worse' (more inflammation). In the second step, we then looked for the set of genes most closely related to the effect of the drug given on the phenotype score. As is typical for screeing in functional genomics neither the set of genes involved, nor their orientation (sign of the correlation with the phenotype) was known.

The foundation of a solution sufficiently efficient for genomic screening was laid in 1989 (Wittkowski 1989), when the need for distinguishing between 'exact ties' (complete ordering) and 'inexact ties' (partial ordering) was identified. By applying the distinction between the conditional and the unconditional variance (Wittkowski 1988b) to this special case of a stratified rank test, it was demonstrated that both versions of the sign test (Dixon, Mood 1946; Dixon, Massey 1951) were valid, albeit for different situations, thereby resolving the long-felt discomfort with the treatment of ties in the McNemar test (McNemar 1947) when applied to rounded data. These results have subsequently been independently confirmed (Rayner, Best 1999; Randles 2001; Fong, Kwan 2003) see also (Wittkowski 2004).

Recently, we have made several major advantages in computational efficiency. First, as mentioned above, we moved from the marginal likelihood (MrgL) to u-statistics. While the computational effort of generating all rank permutations rises with the factorial of the sample size ($n!$), the computational effort for u-statistics rises only with the square of the sample size (n^2). This computational simplicity allowed us to increase the number of subjects that could be scored to 32 while using a spreadsheet as the first implementation and a didactical tool (see, e.g., Fig. 5). While the spreadsheet was highly appreciated for visualizing the algorithm underlying u-statistics, manual screening among all possible polarized subsets is impractical.

To overcome these limitations, the next implementation of the method was written in S-Plus (Insightful Corp.). While S-Plus provides for a convenient environment to implement statistical methods at a high level, some operations, notably 'for loops', cannot be represented efficiently. Thus, we developed a library of C subroutines to increase the computational efficiency, while keeping higher level tasks in S-Plus to allow for easy adaptation to various experimental designs and partial orderings. With these optimizations, the approach became sufficiently efficient for a limited number of variables. The software has recently contributed to gaining new insights into the genetic and genomic determinants of atherosclerosis (Dansky, Ono 2001; Smith, James 2003), cancer susceptibility (Banchereau, Palucka 2001; Palucka, Dhodapkar 2003), allergology (King, Jim 2003), psoriasis (Lowes, Lin 2004), and addiction (Spangler, Wittkowski 2004).

While this provided for a proof-of-principle, we soon realized that a single processor system would not be sufficient to fulfill the growing demands of our investigators. To increase the throughput, we parallelized the software and then sought collaboration from the RU Information Technology department to run the software on a multiprocessor cluster. To make this tool more widely available, we then provided access through a Web server (mustat.rockefeller.edu).

Moving from a desktop to a multiprocessor cluster demonstrated the scalability of the problem, but the feedback from our clinical investigators immediately indicated the need for further increases in computational efficiency. As the next step, we have begun to migrate from the 8-processor cluster to a ≈100-processor grid.

Not all analytical tasks are equally easy to distribute across a grid. In some cases, it may turn out that for several task to share intermediate results may provide a computational advantage, so that it would be preferable to run certain tasks on a cluster of processors having joint local access to the same data. We will then separate these tasks from those being distributed throughout the grid and have them routed to the cluster.

Finally, we anticipate that a multiprocessor architecture may be not optimal for all tasks, but that some task would benefit from dedicating specific hardware. For such tasks, we are currently exploring the possibility to equip dedicated servers with field programmable gate arrays (FPGA).

4 U Statistics in Genetics

Associations between genetic risk factors and clinically relevant phenotypes were first sought by means of simple χ^2 statistics comparing allele prevalence between cases and controls. To better deal with confounding through population admixture, an analytic approach was suggested, termed 'transmission disequilibrium test' (TDT) (Spielman, McGinnis 1993) that would allow to correct for confounding by obtaining the genotypes from cases and their parents, rather than from cases and controls. Over 10 years, this landmark paper has been cited more than 1500 times, making the TDT one of the most frequently used analytic approach in the field of genetics. Still, the TDT was never presented with the formal rigor that had evolved in the field of statistics, although it was thought to be a sign test (Dixon, Massey 1951) or McNemar (1947) test for data with exact ties (Wittkowski 1998; Wittkowski 2004). Instead, the TDT was heuristically motivated using genetic terminology, which led to the belief that independence of transmission events (genetics) implied independence of the observations (statistics). In 2002 (Wittkowski, Liu), we separated the statistical theory from the genetic application, thereby demonstrating that children should be stratified according to the parental mating type using standard statistical methodology for stratified non-parametric tests (Wittkowski 1988b). The resulting stratified McNemar (SMN) test proved to be superior to the TDT in many ways and the discussion shed light on the heuristics underlying the TDT (Wittkowski, Liu 2004).

In particular, the new insight into the nature of this approach now allows for generalizing the family of TDT-like tests. For instance, the power of the TDT and, albeit to a lesser extend, of the SMN, is unsatisfactory for alleles with either both low prevalence (frequency) and low dominance ($r(aA) \approx r(aa)$) or with both high prevalence and high dominance ($r(aA) \approx r(AA)$) (Goddard, Wittkowski 2003).

To further improve statistical tests for association based on trios, we will compare different variants of the SMN through simulation studies to find tests with either uniformly higher power across different levels of dominance or, alternatively, with power characteristics optimized to detect alleles with either low or high dominance. Then, we will extend the methodology from binary to ordinal outcomes (as in the qTDT) and then, as a direct consequence, to multivariate ordinal outcomes (as a novel "uTDT").

As yet another direction for development, our investigators requested a more flexible tool, where information from neighboring marker loci is integrated to identify a disease locus. For inbred populations, all possible diplotypes coincide with the haplotypes, so that 'marker intervals' can be easily ordered (Fig. 8a). For outbred populations, however, the partial ordering can be more complicated (Fig. 8b). The need for a special partial ordering arises from the specific meaning of the term 'interval' in this context.

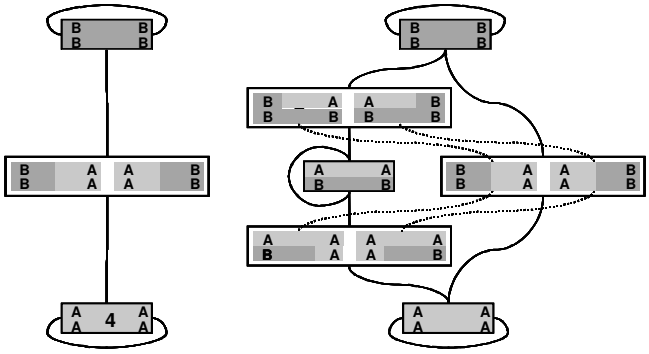

Fig. 8: Partial orderings of genetic evidence for an interval between two markers to contain a disease gene G, left: inbred strains, right: outbread strains. Nodes within boxes are comparable only with nodes connected through a dashed line or through the lines connecting the box, but not among each other.

In one of our first applications (Sehayek, Yu 2004), this new approach let to identifying synteny of a human locus with a 'high plant sterol allele' on chromosome 20 to the mouse locus Plast2b on distal chromosome 2 (Sehayek, Duncan 2002). As the number of SNPs available for analysis increases, several adjacent SNPs may be in linkage disequilibrium with a diseases locus. Moreover, a phenotype may be associated with an 'epistatic set' of diplotypes being several markers apart, or even on different chromosomes (Gambis, Sehayek 2003). Zwei weitere aktuelle Anwendungsgebiete sind die Identifikation genetischer Faktoren bei olfaktorischen Störungen (Leslie Voshall) und bei polycystic kidney diseases (Rogosin Institute).

From our experience, the next steps are (a) to extend the partial ordering for marker intervals to a partial ordering for diplotypes consisting of more than two neighboring SNPs and (b) to extend the default partial ordering of epistatic SNPs to one that allows for epistatic diplotypes.

As u statistics can easily integrate SNP and gene expression data in a single analysis, the above extensions will allow complex interactions between genetic and functional genomic to be addressed.

5 Data Models

A suitable data model is crucial for the ease by which biostatistics and bioinformatics tools can interact with users and among themselves. Making a tool for the analysis of complex designs available for a large group of users using different data base and analysis systems while having limited knowledge in statistics and/or informatics requires data models that are rich enough to represent necessary information, yet simple enough to be easily communicated.

For small data sets and simple statistical methods, it is often sufficient to store data in a single table, e.g., as a spreadsheet. In functional genomics, however, data is frequently associated to different objects. Some variables (e.g., SNPs) are measured once per subject, while others (e.g., gene expression) are measured repeatedly, under different conditions. Storing data in a universal relation (single table), is convenient for data retrieval and inspection, but repeating genetic information for every gene expression profile poses problems for statistical analyses (Wittkowski 1988d), as the number of genetic observations could be seen a inflated by the number of genomic conditions. Of course, this dilemma could be avoided by requiring data to be transmitted in third normal form (3NF), where redundancy is avoided (Codd 1970). This, however, would require that the user decomposes the familiar universal relation into a set of 3NF relations, the number of such relations depending on the particular study design. Still, not all relevant design characteristics are easily represented in the relational data model. To guide the selection of appropriate statistical methods, for instance, it is important to distinguish completely randomized designs, where the first patient in group A bears no similarity to the first patient in group B, from a matched pair design, where the first patient in group A is the sibling of the first patient in group B.

The increasing volume of data and the diversity of statistical methods create new challenges in data management. As part of the '**PA**rametric and **NO**nparametric **S**tatistics' (PANOS) project (Wittkowski 1985), a data model was presented that allowed for data to be represented in the familiar form of a universal relation, while includingthe first representation of meta data ('knowledge') sufficient to guarantee semantically meaningful statistical analyses. This knowledge was then structured into different layers, among them the DESIGN and MODEL layers (Elliman, Wittkowski 1987; Wittkowski 1987; 1992b; 1993). The former specifies the data structure, i.e., the logical relation between different observations, the latter describes the variables' known characteristics, that are relevant for choosing statistical methods, in general, and statistical graphics (Wittkowski 1983), in particular: scale level (nominal < ordinal < interval < absolute), accuracy (exact, rounded), granularity (2 … ∞), and causality (stratum<intervention<observation).

When we developed the current server for '**m**ultivariate **u** **Stat**istics' (muStat.rockefeller.edu), we faced a different problem. The PANOS data model had focused on multifactorial designs with univariate observations. The muStat data model, in contrast, was initially restricted to elementary statistical methods: product moment correlation (ordinal phenotype), 2-sample t-test (binary phenotype). With the focus on multivariate data in genetics and genomics, the model incorporates two additional features. First, the location of the SNPs is formally described to allow the statistical methods to build upon our understanding of the correlation between adjacent loci (linkage). Second, as it is often not known whether a particular allele and gene increases or decreases risk, the polarity of a variable could be declared as unknown ('0'), in addition to '1' or '-1'. For genetic information, one might want to separate chromosomes further into exons or even genes. For genomic (or proteomic) information, one might want to group genes into known functional pathways.

	Nam	SCL	IntrVntn	Pat_Name	BetaDose	Chr1Loc1	Chr1Loc2	Chr1Loc3	Chr2Loc1	Sex	BodyWght	00001_AT	00002_AT	00003_AT	ErgoMetr	ErgoWght
DESIGN	OBJ		A	AB	C	AB	AB	AB	AB	AB	AB	ABC	ABC	ABC	ABC	ABC
	N		2	10	4	10	10	10	10	10	10	40	40	40	40	40
hierarchical variable structure	Chr	Nom				1	1	1	2							
	Pos	Ord				1	2	3	1							
	Pwy	Nom										1	2	3		
	Pht	Nom													1	2
MODEL	SCL		Nom	Nom	Abs	Ord	Ord	Ord	Ord	Nom	Abs	Ord	Ord	Ord	Abs	Abs
	Pol		-	-	0	0	0	0	1	-	1	0	1	1	0	0
DATA	DAT		A	MICKEY	10.0	1	2	0	0	M	74	245	172	2.3	172	2.32
	DAT		A	MICKEY	20.0	1	2	0	0	M	74	74	190	2.6	190	2.57
	DAT		A	MICKEY	50.0	1	2	0	0	M	74	354	195	2.6	195	2.64
	DAT		A	MICKEY	100.0	1	2	0	0	M	74	5	180	2.4	180	2.43
	DAT		A	MINNIE	10.0	0	0	0	0	F	76	22	156	2.1	156	2.05

Fig. 9: Draft proposal for a unification of the PANOS and the muStat data model. The rows in the block 'hierarchical variable structure' provide information necessary to generate a 3NF relation by adding additional keys/factors and 'stacking' variables.

To overcome the deficiencies of traditional data models when representing relevant relationships between variables, we propose to define a 'factorial' structure not only for the subjects, but also for the variables. This structure should be hierarchical, to allow for step-by-step refinement of topological structures (chromosome, intron, gene, SNP) or logical structures (ontology group, pathway, gene).

6 Knowledge Acquisition

The PANOS system (Wittkowski 1985) focused on how best to choose statistical methods, but shared a feature of many previous 'knowledge based' systems: It did not yet include a sufficiently user-friendly (DESIGN) knowledge acquisition module. Thus, when the results were commercialized by SAS, Inc. in 1989 as JMP, only a small portion of DESIGN knowledge was incorporated.

As new technologies become available we are now working on a prototype of a graphical user interface, where clinical trials, genetic studies, or gene expression experiments can be described interactively in sufficient detail to extract DESIGN knowledge. Fig. 10 shows an example based on one of our ongoing studies on cardiovascular and metabolic diseases (with the RU Laboratory of Biochemical Genetics and Metabolism).

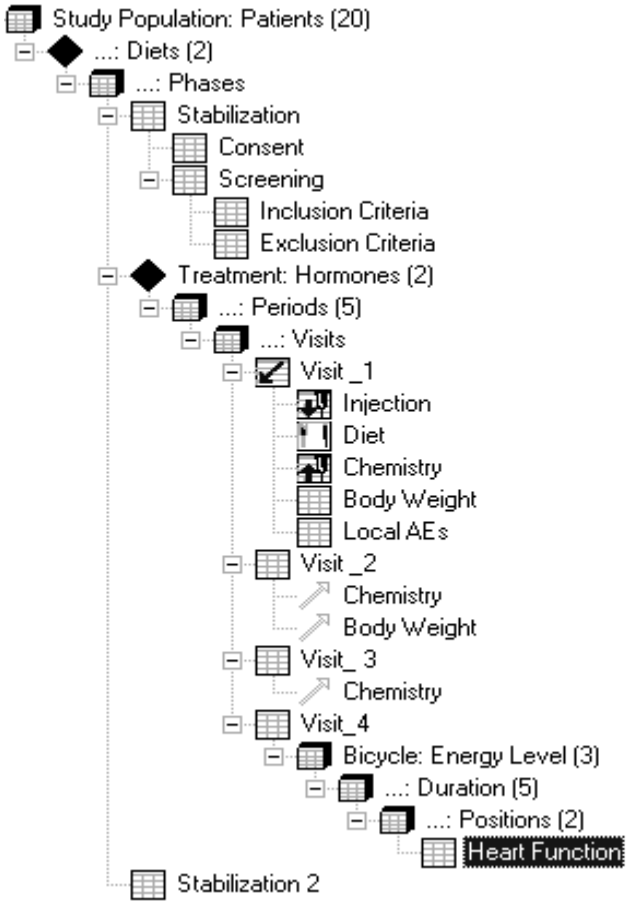

Fig. 10: Screen shot from the **W**eb-based **I**nteractive **S**tudy **D**esign, **O**peration, & **M**anagement (WISDOM) module: Patients are first randomized to two diets (low carb vs low fat) and then, within each diet to two hormones (leptin vs placebo). During Visit_4, Heart function is evaluated in a factorial design.

Allowing for information on genetic epistasis, genomic pathways, and proteomic profiles to be combined for statistical analysis requires even more meta data to be added to data models (see Section 5. Data Model). The analysis system needs to be able to differentiate between different partial orderings (see Section 2. U Statistics for Changes in Function), to employ appropriate screening strategies (see Section 3.3. U Statistics for Microarrays: Gene expression profiling), to consider topological and functional relations between variables (see Section 4, U Statistics in Genetics).

Having a sufficiently rich formal description of experimental designs available has a number of advantages, each of which typically draws on a different view of the knowledge base. In particular, the DESIGN knowledge base can be used to facilitate

database creation, by automatically setting up a (relational or object-attribute-value) database,

sample size calculations, by allowing to simulate power for complex designs,

protocol writing, e.g., by generating pre-populated protocol templates for NIH's ProtoType system,

security, by facilitating the use of centralized servers, thereby reducing the need for decentralized storage, e.g., as spreadsheets on various PCs,

study management, by generating "pathways" providing nurses with workflow information,

data entry, by generating case report forms, either on paper or as Web interfaces,

monitoring, by allowing software to automatically screen for adverse event profiles and alerting data and safety monitoring boards based on preset criteria, if necessary,

inspection, by giving investigators real-time information while automatically protecting them from becoming accidentally unblinded,

review, by providing the Institutional Review Board with progress information.

analysis, by storing primary and secondary objectives and automatically initiating the appropriate analyses according the protocol,

data sharing, by providing standardized dictionaries, allowing data to be related across studies.

Thus, to facilitate use of the new statistical methods and analytical tools in genetics, functional genomics, and proteomics, we will continue to work with investigators, computer scientists, and statisticians on integrating knowledge acquisition modules and knowledge based analysis tools.

7 Summary

The traditional statistical methods implemented in most of the bioinformatics tools currently used in the novel field of genetics and functional genomics are based on the linear model and, thus, have shortcomings when applied to nonlinear biological systems. The previous work on partially ordered data (Wittkowski 1988b; 1992a), when combined with theoretical results (Hoeffding 1948) and computational strategies (Deuchler 1914) has opened a new field of nonparametric statistics. With grid technology, new tools are now feasible when screening for interactions between genetics (Wittkowski, Liu 2002) and functional genomics (Wittkowski, Lee 2004). We will continue the ongoing collaborative efforts in broadening the spectrum of partial orderings to meet the demands of more complex study designs and to make these bioinformatics/biostatistics tools available over the Web.

Having more complex study designs and more specific methods available increases the demand for decision support when selecting appropriate bioinformatics tools. With the advent of rapid prototyping systems for Web based database application, we have recently begun to complement my previous work on knowledge based systems (Wittkowski 1985; 1988c; 1988a; 1991; 1993) with graphical Web-based tools for acquisition of DESIGN and MODEL knowledge. To make the biostatistics tools more widely available and more easily accessible, we will continue to improving data base design and knowledge acquisition tools for clinical trials, in general, and for functional genomics and genetics, in particular.

References

BANCHEREAU J, PALUCKA AK, DHODAPKAR M, KURKEHOLDER S, TAQUET N, ROLLAND A, TAQUET S, COQUERY S, WITTKOWSKI KM, BHARDWJ N, PINEIRO L, STEINMAN R, FAY J (2001) Immune and clinical responses after vaccination of patients with metastatic melanoma with CD34+ hematopoietic progenitor-derived dendritic cells. *Cancer Research* **61**: 6451–8

CODD EF (1970) A relational model of dta for large shared data bases. *Comm ACM* **13**: 377–87

CRONBACH LJ, MEEHL PE (1955) Construct validity in psychological tests. *Psychol Bull* **52**: 281–302

DANSKY HM, ONO JG, DANSKY AH, WITTKOWSKI KM, BRESLOW JL (2001) A novel method to detect gene interactions determining atherosclerosis susceptibility in apoE-deficient mice. *Circulation* **104**: 1069

DEUCHLER G (1914) Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. *Z pädagog Psychol* **15**: 114–31 45–59 229–42

DIXON WJ, MASSEY FJJ (1951) *An Introduction to Statistical Analysis*. New York NY: McGraw-Hill

DIXON WJ, MOOD AM (1946) The statistical sign test. *J Am Statist Assoc* **41**: 557–66

EINSELE H, EHNINGER G, HEBART H, WITTKOWSKI KM, SCHULER U, JAHN G, MACKES P, HERTER M, KLINGEBIEL T, LÖFFLER J, ET AL. (1995) Polymerase chain reaction monitoring reduces the incidence of cytomegalovirus disease and the duration and side effects of antiviral therapy after bone marrow transplantation. *Blood* **86**: 2815–20

ELLIMAN AD, WITTKOWSKI KM (1987) The impact of expert systems on statistical database management. *Stat Softw Newsl* **13**: 14–27

FONG DYT, KWAN CW, LAM KF, LAM KSL (2003) Use of the sign test for the median in the presence of ties. *Am Statist* **57**: 237–40

GAMBIS A, SEHAYEK E, WITTKOWSKI KM (2003) It's not Broadway - Visualizing Epistatic Interaction. *TIGR's 15th Annual Genome Sequencing and Analysis Conference (GSAC XV)* 2003–08–03..07. Savannah GA

GEHAN EA (1965a) A generalised two-sample Wilcoxon test for doubly censored samples. *Biometrika* **52**: 650–3

GEHAN EA (1965b) A generalised Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* **52**: 203–23

GODDARD NL, WITTKOWSKI KM (2003) A class of stratified McNemar tests to replace the TDT. *Joint Statistical Meetings*. San Francisco CA

HAIDER A, NAEF F, WITTKOWSKI KM (2003) A non-parametric method for signal value calculation (SVC) from pairs of perfect and mis-matches. *CAMDA: Critical Assessment of Microarray Data Analysis* 2003–11–12..14. Durham NC

HOEFFDING W (1948) A class of statistics with asymptotically normal distribution. *Ann Math Statist* **19**: 293–325

HUBBELL E, LIU W-M, MEI R (2002) Robust estimators for expression analysis. *Bioinformatics* **18**: 1585–92

IRIZARRY RA, HOBBS B, COLLIN F, BEAZER-BARCLAY YD, ANTONELLIS KJ, SCHERF U, SPEED TP (2003) Exploration normalization and summaries of high density oligonucleotide array probe level data. *Biostat* **4**: 249–64

KING TP, JIM SY, WITTKOWSKI KM (2003) Inflammatory role of two venom components of yellow jackets (Vespula vulgaris): a mast cell degranulating peptide mastoparan and phospholipase A1. *International Archives of Allergy and Immunology* **131**: 25–32

LOWES MA, LIN S-L, LEE E, KIKUCHI T, GILLEAUDEAU P, SULLIVAN-WHALEN M, CARDINALE I, KHATCHERIAN A, NOVITSKAYA I, WITTKOWSKI KM, KRUEGER JG (2004) Alefacept reduces infiltrating T cells activated dendritic cells and inflammatory genes in psoriasis vulgaris. *Proceedings of the National Academie of Sciences*

MANN HB, WHITNEY DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist* **18**: 50–60

MCNEMAR Q (1947) Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrica* **12**: 153–7

NAEF F, MAGNASCO MO (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys* **68**: 011906. Epub 2003 Jul 16.

PALUCKA AK, DHODAPKAR MV, PACZESNY S, BURKEHOLDER S, WITTKOWSKI KM, STEINMAN RM, FAY J, BANCHEREAU J (2003) Single injection of CD34+ progenitor-derived dendritic cell vaccine can lead to induction of T-cell immunity in patients with stage IV melanoma. *J Immunother* **26**: 432–9.

POPPER KR (1937) *Logik der Forschung*. Wien: Julius Springer

RANDLES HR (2001) On neutral responses (zeros) in the sign test and ties in the Wilcoxon-Mann-Whitney Test. *Am Statist* **55**: 96–101

RAYNER JCW, BEST DJ (1999) Modelling Ties in the Sign Test. *Biometrics* **55**: 663–5

SAS INSTITUTE INC. (2002) *JMP Version 5*. Cary NC: SAS Publishing 1776

SCHEMPER M (1983) A nonparametric k-sample test for data defined by intervals. *Statistica Neerlandica* **37**: 69–71

SEHAYEK E, DUNCAN EM, LUTJOHANN D, VON BERGMANN K, ONO JG, BATTA AK, SALEN G, BRESLOW JL (2002) Loci on chromosomes 14 and 2 distinct from ABCG5/ABCG8 regulate plasma plant sterol levels in a C57BL/6J x CASA/Rk intercross. *Proc Nat Acad Sci* **99**: 16215–9

SEHAYEK E, YU HJ, VON BERGMANN K, LUTJOHANN D, WITTKOWSKI KM, LEVENSTIEN MA, GORDON D, STOFFEL M, GRARCIA-NAVEDA L, SALIT J, BLUNDELL ML, FRIEDMAN JM, BRESLOW JL (2004) Genetics of cholesterol absorption and plasma plant sterol levels on the Pacific island of Kosrae. *Circulation*: (in press)

SMITH JD, JAMES D, DANSKY HM, WITTKOWSKI KM, MOORE KJ, BRESLOW JL (2003) In silico quantitative trait locus map for atherosclerosis susceptibility in apolipoprotein e-deficient mice. *Arterioscler Thromb Vasc Biol* **23**: 117–22. (22411808)

SPANGLER R, WITTKOWSKI KM, GODDARD NL, AVENA NM, HOEBEL BG, LEIBOWITZ SF (2004) Opiate-like Effects of Sugar on Gene Expression in Reward Areas of the Rat Brain. *Molecular Brain Research* **124**: 134–42

SPIELMAN RS, MCGINNIS RE, EWENS WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506–16.

STUDENT (1908) On the probable error of a mean. *Biometrika* **6**: 1–25

SUÁREZ-FARIÑAS M, HAIDER A, WITTKOWSKI KM (2005) "Harshlighting" small blemishes on microarrays. *BMC Bioinformatics* **6**: 65

SUSSER E, DESVARIEUX M, WITTKOWSKI KM (1998) Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. *Am J Public Health* **88**: 671–4

TALAAT M, WITTKOWSKI KM, HUSEIN MH, BARAKAT R (1998) A new procedure to access individual risk of exposure to cercariae from multivariate questionnaire data. In: BARLOW R, BROWN JW (eds) *Reproductive health and infectious diseases in the Middle East.* Aldershot U.K.: Ashgate 167–74

WILCOXON F (1954) Individual Comparisons by Ranking Methods. *Biometrics* **1**: 80–3

WITTKOWSKI KM (1980) *Ein nichtparametrischer Test im Stufenblockplan [A nonparametric test for the step-down design]*. Göttingen D: Georg-August-Universität Diplomarbeit [M.S. thesis] 87

WITTKOWSKI KM (1983) The use of data structures for statistical graphics in clinical trials. In: FOKKENS O, AL. E (eds) *Medinfo-83 Seminars.* 235

WITTKOWSKI KM (1984) Semiquantitative Merkmale in der nichtparametrischen Statistik. In: KÖHLER CO, WAGNER E, TAUTU P (eds) *Der Beitrag der Informationsverarbeitung zum Fortschritt der Medizin.* (*Medizinische Informatik und Statistik* **55**) Berlin D: Springer 100–5

WITTKOWSKI KM (1985) *Ein Expertensystem zur Datenhaltung und Methodenauswahl für statistische Anwendungen [An expert system for data management and method selection for statistical applications]*. Stuttgart D: Technische Universität Dissertation [Ph.D. dissertation] 386

WITTKOWSKI KM (1987) An expert system approach for generating and testing statistical hypotheses. In: PHELPS B (ed) *Interactions in artificial intelligence and statistical methods.* Aldershot GB: Unicom 45–59

WITTKOWSKI KM (1988a) Building a statistical expert system with knowledge bases of different levels of abstraction. In: EDWARDS D, RAUNE NE (eds) *Compstat 1988 - Proceedings in Computational Statistics.* Heidelberg D: Physica 129–34

WITTKOWSKI KM (1988b) Friedman-type statistics and consistent multiple comparisons for unbalanced designs. *J Am Statist Assoc* **83**: 1163–70

WITTKOWSKI KM (1988c) Intelligente Benutzerschnittstellen für statistische Auswertungen. In: FAULBAUM F, UEHLINGER HM (eds) *Fortschritte der Statistik-Software.* (**1**) Stuttgart D: Fischer

WITTKOWSKI KM (1988d) Knowledge based support for the management of statistical databases. In: RAFANELLI M, KLENSIN JC, SVENSSON P (eds) *Statistical and Scientific Database Management.* (*Lecture Notes in Computer Science* **339**) Berlin D: Springer 62–71

WITTKOWSKI KM (1988e) Small sample properties of rank tests for incomplete unbalanced designs. *Biometrical Journal* **30**: 799–808

WITTKOWSKI KM (1989) An asymptotic UMP sign test for discretized data. *The Statistician* **38**: 93–6

WITTKOWSKI KM (1991) A structured visual language for a knowledge-based front-end to statistical analysis systems in biomedical research. *Comput Methods Programs Biomed* **35**: 59–67

WITTKOWSKI KM (1992a) An extension to Wittkowski. *J Am Statist Assoc* **87**: 258

WITTKOWSKI KM (1992b) A knowledge-based interface to statistical analysis systems — report on the current status of the MS-DOS implementation. In: FAULBAUM F (ed) *SoftStat '91.* (*Advances in Statistical Software* **3**) Stuttgart D: Fischer 57–64

WITTKOWSKI KM (1993) Eine wissensbasierte Schnittstelle zu statistischen Auswertungssystemen. In: BÖCKER HD, GLATTHAAR W, STROTHOTTE T (eds) *Mensch-Computer-Kommunikation: Benutzergerechte Systeme auf dem Weg in die industrielle Gesellschaft.* Berlin D: Springer 166–78

WITTKOWSKI KM (1998) Versions of the sign test in the presence of ties. *Biometrics* **54**: 789–91

WITTKOWSKI KM (2003) Novel Methods for Multivariate Ordinal Data applied to Genetic Diplotypes Genomic Pathways Risk Profiles and Pattern Similarity. *Computing Science and Statistics* **35**: 626–46

WITTKOWSKI KM (2004) Effects and non-effects of paired identical observations in comparing proportions with binary matched-paris data. *Stat Med* **23**: 2317

WITTKOWSKI KM, LEE E, NUSSBAUM R, CHAMIAN FN, KRUEGER JG (2004) Combining several ordinal measures in clinical studies. *Stat Med* **23**: 1579–92

WITTKOWSKI KM, LIU X (2002) A statistically valid alternative to the TDT. *Hum Hered* **54**: 157–64. (22513746)

WITTKOWSKI KM, LIU X (2004) Beyond the TDT: Rejoinder to Ewens and Spielman. *Hum Hered* **58**: 60–1

WITTKOWSKI KM, SUSSER E, DIETZ K (1998) The protective effect of condoms and nonoxynol-9 against HIV infection. *Am J Public Health* **88**: 590–6 972

WU Z, IRIZARRY RA, GENTLEMAN R, MARTINEZ MURILLO F, SPENCER F (2004) A model based background adjustment for oligonucleotide expression arrays. *J Am Statist Assoc*: (in press)