2008

# Phenotyping Genetic Diseases Using an Extension of μ-Scores for Multivariate Data

Jose F. Morales

Tingting Song

Arleen D. Auerbach

Knut M. Wittkowski

*Statistical Applications in Genetics and Molecular Biology*

*Volume* 7, *Issue* 1 2008 *Article* 19

# Phenotyping Genetic Diseases Using an Extension of $\mu$-Scores for Multivariate Data

José F. Morales* Tingting Song†

Arleen D. Auerbach‡ Knut M. Wittkowski**

*The Rockefeller University, moralej@rockefeller.edu

†The Rockefeller University, tsong01@rockefeller.edu

‡The Rockefeller University, auerbac@rockefeller.edu

**The Rockefeller University, kmw@rockefeller.edu

# Phenotyping Genetic Diseases Using an Extension of $\mu$-Scores for Multivariate Data*

José F. Morales, Tingting Song, Arleen D. Auerbach, and Knut M. Wittkowski

## Abstract

As the field of genomics matures, more complex genotypes and phenotypes are being studied. Fanconi anemia (FA), for example, is an inherited chromosome instability syndrome with a complex array of variable disease phenotypes including congenital malformations, hematological manifestations, and cancer. To better understand specific aspects of the genetic etiology of FA and other rare diseases with complex phenotypes, it is often necessary to reduce the dimensions of the disease phenotype information. Towards this end, we extend a novel non-parametric approach to include information about a hierarchical structure among disease phenotypes. The proposed extension increases information content of the phenotype scores obtained and, thereby, the power of genotype-phenotype relationships studies.

**KEYWORDS:** multidimensional, ranking, Fanconi anemia, censoring, genotype, phenotype, non-parametric

---

*This work was partially supported by grants from NIH (R37HL32987) to Arlene D. Auerbach, and by an NIH/Clinical and Translational Science Award (CTSA) grant UL1 RR024143. We would like to acknowledge Orna Levran, Frank Lach, Tom Landers, Rashida Henry for their invaluable help with data. We also thank all the FA family members for their cooperation and the many physicians who have registered their FA patients in the IFAR, without whom this research could not have been done.

## INTRODUCTION

An objective of genomic biology is to elucidate the relationship between genetic variation and phenotype. High throughput technology has led to an expansion of genotypic information including genome sequencing in humans and model organisms. As genotype resources are being annotated, organized, structured, and made accessible, phenotype resources await similar efforts. Thus, the importance of the organization of the human phenome is being recognized (Freimer and Sabatti 2003).

A phenome is a catalog of phenotypes or traits expressed by a cell, tissue, organ, organism, or species that arise from genetic, epigenetic, and environmental interactions. Of significance is the subset of traits associated with disease. Determining genotype-phenotype relationships, essential for the study of genetic disorders, requires well organized disease phenomes (Freimer and Sabatti 2003).

Fanconi anemia (FA) is a disease found in many populations (Auerbach 1993; Offit *et al.* 2003; Savino *et al.* 2003; Yagasaki *et al.* 2003; Kutler and Auerbach 2004; Tamary *et al.* 2004; Callen *et al.* 2005). Mutations in 13 known genes, *FANCA*, *B, C, D1, D2, E, F, G*, *I, J, L, M* and *N* (Meetei *et al.* 2003; Levitus *et al.* 2004; Levran *et al.* 2005; Meetei *et al.* 2005; Reid *et al.* 2007; Smogorzewska *et al.* 2007) are known to cause FA through a defect in a BRCA-related DNA repair pathway (Nakanishi *et al.* 2005; Wang 2007). This defect causes sensitivity to DNA cross-linking agents and chromosomal instability (Auerbach 1993; Meyn 1997; Taylor 2001; Charames and Bapat 2003; De la Torre *et al.* 2003). FA subjects display diverse disease symptoms. Most FA subjects develop hematological abnormalities at a young age, with bone marrow failure resulting in aplastic anemia. Many FA subjects have congenital abnormalities including hyperpigmentation, short stature, and radial ray abnormalities (Giampietro *et al.* 1993; Rosenberg *et al.* 2004). FA subjects have an increased risk of primarily hematological cancers (Alter *et al.* 2003; Kutler *et al.* 2003).

Many diseases have tools for scoring severity or phenotype. In cystic fibrosis, there is a pulmonary disease severity index (Hafen *et al.* 2006), sickle cell anemia researchers have proposed a severity index (Vanscoy *et al.* 2007), and in Huntington's disease there is the Unified Huntington's disease rating scale (Klempir *et al.* 2006). Genetic association studies of complex diseases such as Schizophrenia (John *et al.* 2008), alcoholism (Dick *et al.* 2008), diabetes (Meigs *et al.* 2007) and insulin resistance (An *et al.* 2005), are also using composite and multiple phenotypes.

For FA, scoring systems were developed with different goals. Auerbach et al. (1989) developed a "simplified scoring method" to diagnose FA. This method correlates FA symptoms with increased sensitivity to chromosome breakage in lymphocytes exposed to a DNA cross linking agent. The variables' coefficients

are determined through logistic regression (Giampietro *et al.* 1993). Rosenberg et al. (2004) advanced a system that aggregated FA outcome risk factors into a model for risks assessments with summary scores. Faivre et al. (2000) proposed a scoring system to connect FA outcomes to complementation groups. These tools have uses such as disease state classification, evaluation of disease phenotypes, disease course prediction, and disorder genetics and genomics.

The FA phenotype includes chromosome breaks, hematological manifestations, cancer and congenital malformations. To be able to score such a complex phenotype with respect to a overall severity, we extend a statistical approach (Wittkowski *et al.* 2004) based on u-statistics (Hoeffding 1948) to include information about a hierarchical structure among phenotype variables. The resulting hierarchical μ-scores (for multivariate u-scores) comprise a system for ranking FA subjects according to severity. This system will be called the Fanconi Anemia Phenome Score (FAPS). The FAPS is being developed as a tool for research. One application explores FA genotype-phenotype relationships between *FANCA* mutations and the FAPS.

## MATERIALS AND METHODS

### DATA SOURCE:

This study is based on subject data from the International Fanconi Anemia Registry (IFAR), which was established at The Rockefeller University in 1982 to gather genotype and phenotype information from FA subjects (Auerbach *et al.* 1989). FA subjects are registered in the IFAR when their diagnosis is confirmed in peripheral blood lymphocytes by assessing chromosomal breakage induced through the DNA cross-linking agent diepoxybutane (DEB). The information collected includes cancer incidence, hematological abnormalities, and congenital malformations. Attempts to obtain follow-up data and report results have been made periodically. This study is based on 239 individuals or 19% of the 1200 subjects in the IFAR. To reduce variance and avoid bias due to technical errors we excluded subjects with unique mutations. Approval for these studies was obtained from The Rockefeller University Institutional Review Board. Informed consent was provided according to the Declaration of Helsinki.

### PHENOTYPES

We organized 36 phenotype outcomes in five categories: chromosomal breaks (CB), life span (LS), cancer (CA), hematological manifestations (HM) and congenital malformations (CM) (see Figure 1). The chromosome breakage category is comprised of the proportion of aberrant cells and number of breaks per aberrant

cell for baseline (BL) and DEB-induced breaks. Mosaic status is defined as <50% aberrant cells in the DEB test. In some subjects, mosaicism results from the reversion of pathogenic mutations to wild-type. Twenty outcomes are censored, i.e., observed as two variables: last day negative (LDN, '0' if unknown) and first day positive (FDP, '+∞' if unknown). Thus, the 36 outcomes consist of 56 variables. For instance, LS is defined as time from date of birth to censored date of death. CA data consists of censored data for both solid tumor and leukemia onset. HM data is comprised of censored information on any HM, including platelets, red, and white cells and hematological status at the time of bone marrow transplant. CM data consists of binary data on five minor and eight major malformations as well as continuous data for head circumference and height percentiles.

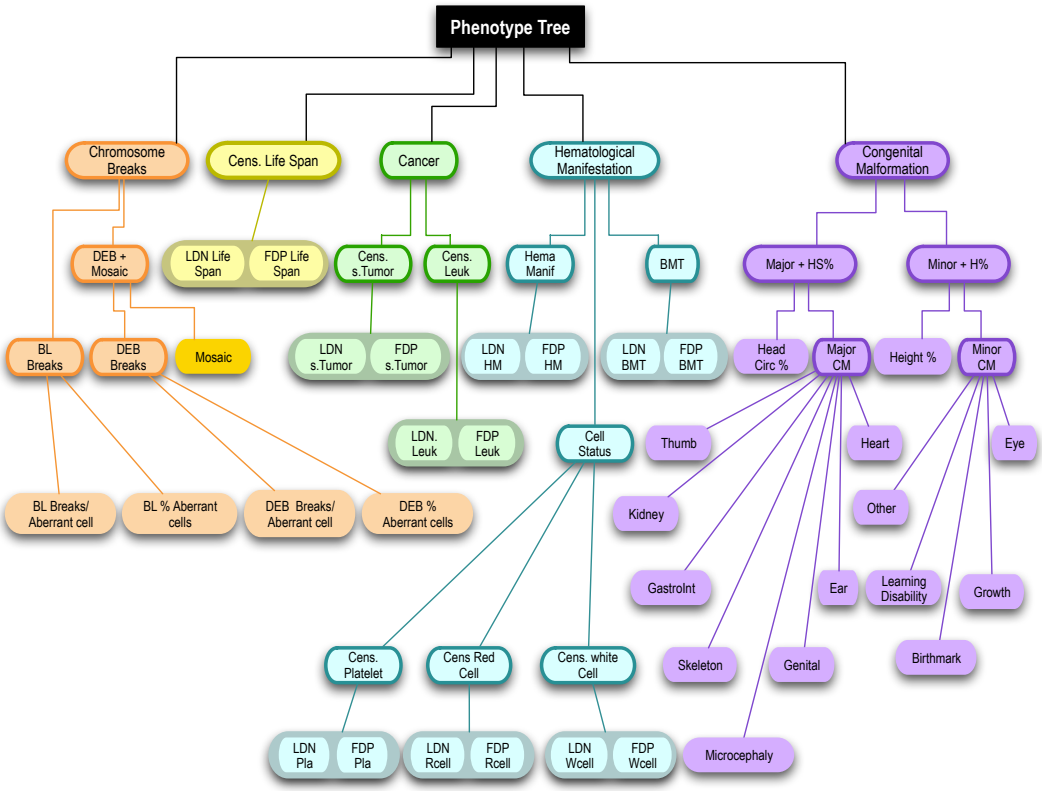


**Figure 1: Fanconi anemia phenotype tree.** The hierarchical organization of the multilayered multi- and univariate FA phenotype data. *Abbreviations* (DEB: Diepoxybutane; BL:Baseline; Cens.: censored; LDN: Last Date Negative; FDP: First Date Positive; Leuk: leukemia; sTumors: solid Tumors; HM: Hematological Manifestation; BMT: Bone Marrow Transplant; CM: congenital malformation; HC%: Head circumference percentile; Height%: Height percentile; GastroInt: gastrointestinal); *Tree node descriptions*: Censoring nodes (nested LDN/FDP pairs, 8 pairs); Univariate terminal nodes ('leaves', no borders, total: 36); Multivariate nodes (higher level information, thick border, total:.20).

## GENOTYPE-PHENOTYPE RELATIONSHIPS

Our study focuses on FANCA subjects, because they constituted the largest group. The protein coded by FANCA's 43 exons is part of a nuclear core complex (Garcia-Higuera *et al.* 1999). The *FANCA* mutational events include amino acid substitutions (26 cases), genomic deletions (133), splicing mutations (2), frameshifts (25), stop codons (10) and unknown mutations (4).

A mutation map was constructed to represent the bi-allelic mutational status in these subjects (Figure 2). For FA to occur, both FANCA alleles need to be mutated, either in the same or in different positions. The mutational events affect the transmission of genetic information into protein. We assume that an amino acid substitution affects the mutation site only, while genomic deletions, frameshifts, splicing mutations and stop codons have effects on protein function that propagate throughout the protein. The mutation status (Figure 2) indicates these assumed changes in the gene product.

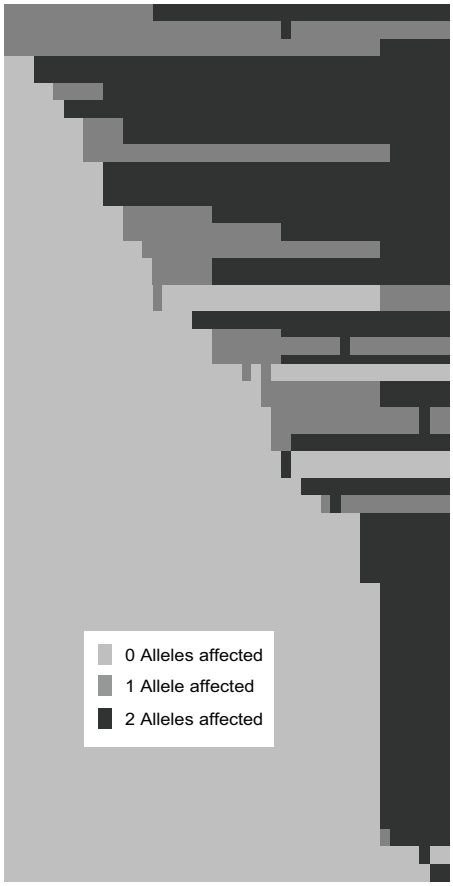


**Figure 2: FANCA mutation map:** This diagram displays the regions (exons and introns) on the x-axis and the FANCA subjects on the y-axis sorted by the first affected region. Each cell indicates the predicted effect on the protein (no, one, or both alleles affected). Certain FANCA mutations (e.g. deletions; splicing; frameshifts; stop codons) occurring in a particular region are assumed to affect all downstream regions.

## MULTIVARIATE U-SCORES

Non-parametric statistical methods have been recommended for analyses in human phenomics (Freimer and Sabatti 2003). Multivariate u-statistics (μ-statistics) (Wittkowski *et al.* 2004) have advantages over multivariate methods based on the (generalized) linear model, where dimensionality is reduced by assuming a linear combination of variables, often after a (supposedly) linearizing transformation. Since the relative importance of the variables, their correlation, and the relationship of each variable with the latent factor ('genetic risk' or 'disease severity') are unknown, model validity cannot be established on theoretical grounds. Trying to resolve this conundrum through empirical

'validation', i.e., by choosing weights and functions that provide a fit with a 'gold standard' is not only conceptually problematic (Popper 1959), but also often impractical. Using a 'Delphi method' approach (Delbecq 1975), where weights and functions are agreed upon by a group of experts, allows for a comparison between studies where the researchers agreed to use the same scoring system. However, the diversity of scoring systems used attests to the subjective nature of this process.

As a non-parametric method, μ-statistics have the advantage of requiring fewer assumptions to be made about the variables (Friedman 1937). The only assumption regarding the relationships between variables and latent factors is that each variable has an orientation (is 'directional') – if all other variables are held constant, an increase is either always 'good' or always 'bad'.

μ-Scores are based on the concept of partial orderings, where some of the pairwise orderings can be ambiguous, i.e., one can find any set of monotonous transformations applied to or any set of non-zero weights assigned to the variables that change the pairwise ordering. As μ-scores are not affected by such transformations or weights, one does not need to select and justify them.

When variables can be assumed to be correlated with a single latent factor (e.g., disease severity), a partial ordering among the subjects can be defined (Wittkowski 1992). If one of two subjects has values at least as high among all variables, but higher in at least one variable than the other subject, it will be called 'superior'. A μ-score is assigned to each subject by counting the number of inferior subjects and subtracting the number of superior subjects.

A downside of this approach is that the number of ambiguous pairwise orderings 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. When additional information is available, some of these ambiguities can be resolved. In this paper we utilize information about a structure among the variables. If the order between subjects A and B is ambiguous with respect to variables related to one factor (e.g., HM), unambiguous results with respect to another factor (e.g., CA) can 'resolve' this ambiguity.

In Wittkowski *et al.* (2004), the square matrix describing the partial ordering based on ordinal data was derived by comparing data profiles. Here, we separate this process into two steps, as outlined in Figure 3. Let $j, j' = 1, \ldots, n$ denote the subjects (rows) and $i$ the variables (columns). The univariate pair-wise orderings for a set of subjects are represented as the $n \times n$ matrices $U^{(i)} = \left( \left( u_{jj}^{(i)} \right) \right)$ (middle row) with $u_{jj'}^{(i)} = ?$ if $x_j^{(i)}$ or $x_{j'}^{(i)}$ are missing and $u_{jj'}^{(i)} = \mathrm{I}\left( x_{j'}^{(i)} < x_j^{(i)} \right) - \mathrm{I}\left( x_{j'}^{(i)} > x_j^{(i)} \right)$ otherwise (Wittkowski *et al.* 2004). This extends the matrix introduced by Deuchler (1914) to missing data.

The matrix $U$ obtained by the 'AND' operation

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

is the same as one would obtain by applying the partial ordering defined in (Wittkowski *et al.* 2004) and, thus, the scores obtained from $U = \bigoplus_i U^{(i)}$ (bottom of Figure 3) are the non-hierarchical μ-scores.

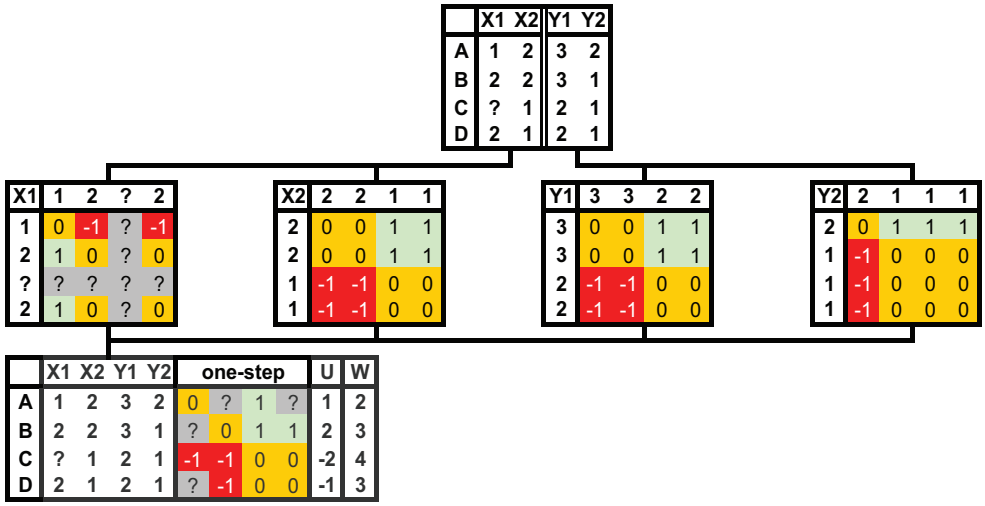


**Figure 3: Ambiguity caused by discordant pairwise orderings across variables.** Hypothetical example with four variables *X1*, *X2*, *Y1*, and *Y2*, observed in four subjects A, B, C, and D. The node on top shows the data, the center row the four univariate partial (X1) and complete (X2, Y1, Y2) orderings, The node at the bottom shows the data, the multivariate partial ordering, the μ-scores (U) and their information content (W). Each partial ordering shows whether the row element is smaller (-1), identical (0), or larger (1) than the column element. Ambiguous pairwise orderings are indicated as '?'. On the one hand, ambiguities (such as those in X1) can be resolved if the corresponding pairwise orderings in the other orderings are unambiguous. On the other hand, ambiguities can arise (such as in the upper right and lower left corner) if some pairwise orderings are negative, while others are positive.

If the variables X1 and X2 are related to the same 'factor' (e.g., CB), while the variables Y1 and Y2 are related to another 'factor' (e.g., CM), one can replace $U_{\text{NH}} = \bigoplus_i U^{(i)}$ by $U_{\text{H}} = \bigoplus \left\{\bigoplus_{\{i:\,\text{X1, X2}\}} U^{(i)}, \bigoplus_{\{i:\,\text{Y1, Y2}\}} U^{(i)}\right\}$. The advantage of creating the matrices for the univariate orderings first and combining them in a separate step before compute the scores, is that incorporating knowledge about the

sub-factor hierarchy by hierarchically combining the matrices can reduce loss of information content (number of unambiguous pairwise orderings contributing to a score). Figure 4 demonstrates how reflecting the structure increases information content by resolving the ambiguity related to comparing A vs. D.

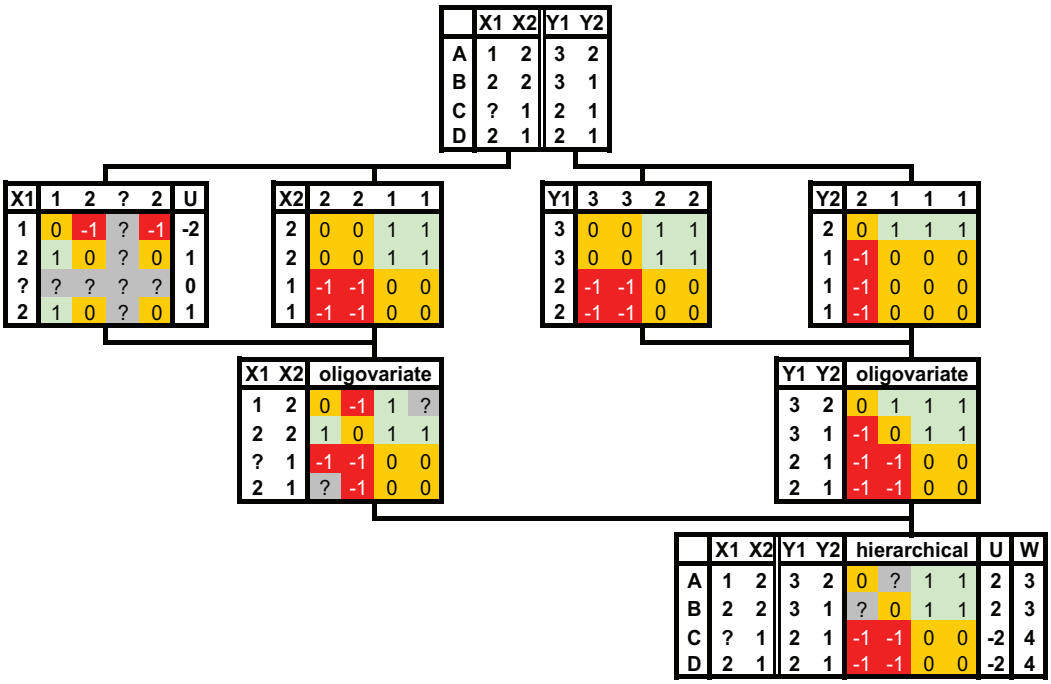


**Figure 4: Resolving ambiguity from discordant pairwise orderings:** Using hierarchical structure in the example of Figure 3. Note that adding the intermediate step resolves the ambiguity in the lower left and upper right corner of the matrices representing the partial orderings.

Figure 4, 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 pairwise orderings 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.

μ-Score can easily handle censored (including interval-censored) variables such as LS, CA and HM, 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 (see Figure 5). 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. The censored information includes date of birth and date of death for life span, dates for solid tumor and leukemia onset for CA and dates of onset for any HM, abnormal platelets, red, and white cells and hematological status on the bone marrow transplant date.

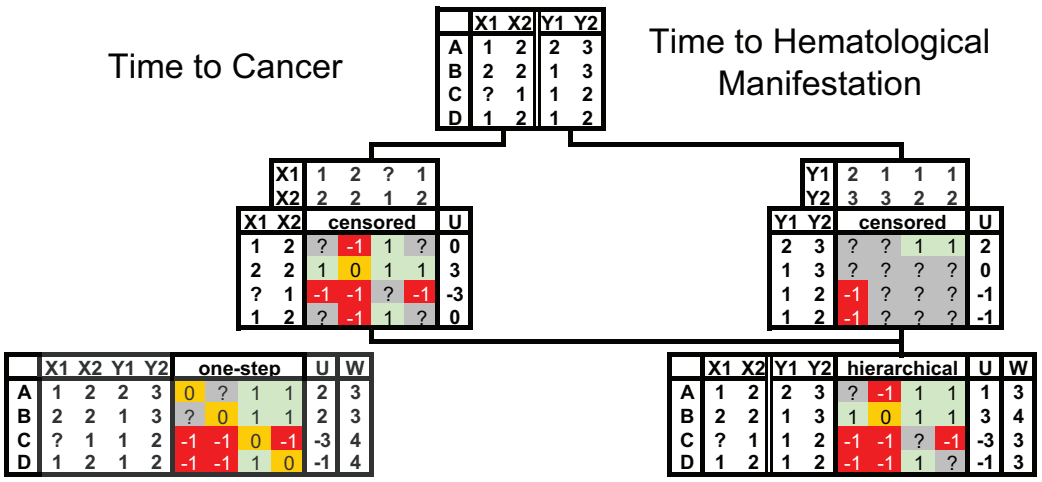


**Figure 5: Reflecting censoring** during the creation of matrices of pairwise orderings. Subjects A and B can be ordered if X1(A)>X2(B) or Y2(A)<Y1(B). unless one observation is uncensored, in which case "≥" or "≤" suffice.

By adjusting the transformation (censored or non-censored) and the aggregation of subsets (hierarchical or overlapping), μ-scores can be used for a range of problems. The 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.

In the muStat package for S and R, `mu.PwO` generates Deuchler's pairwise (univariate) orderings, `mu.AND` combines them into a partial pairwise ordering, and `mu.Sums` computes scores and weights from a partial pairwise ordering. The information about the variables is provided as a formula, where parentheses indicate the hierarchy and variables separated by colon indicate the bounds of interval censored observations (see the Appendix for details).

## RESULTS

### CALCULATING THE FANCONI ANEMIA PHENOME SCORE

μ-scores were computed for each of the 56 branches of Figure 1 with respect to the structure among the 36 FA variables. In Figure 6, variables were assigned a 'polarity' of +1 or −1 to ensure a common direction with 'severity' (high are 'good' for LS, but 'bad' for CB). This system for ranking FA subjects according to severity comprises the Fanconi Anemia Phenome Score (FAPS).

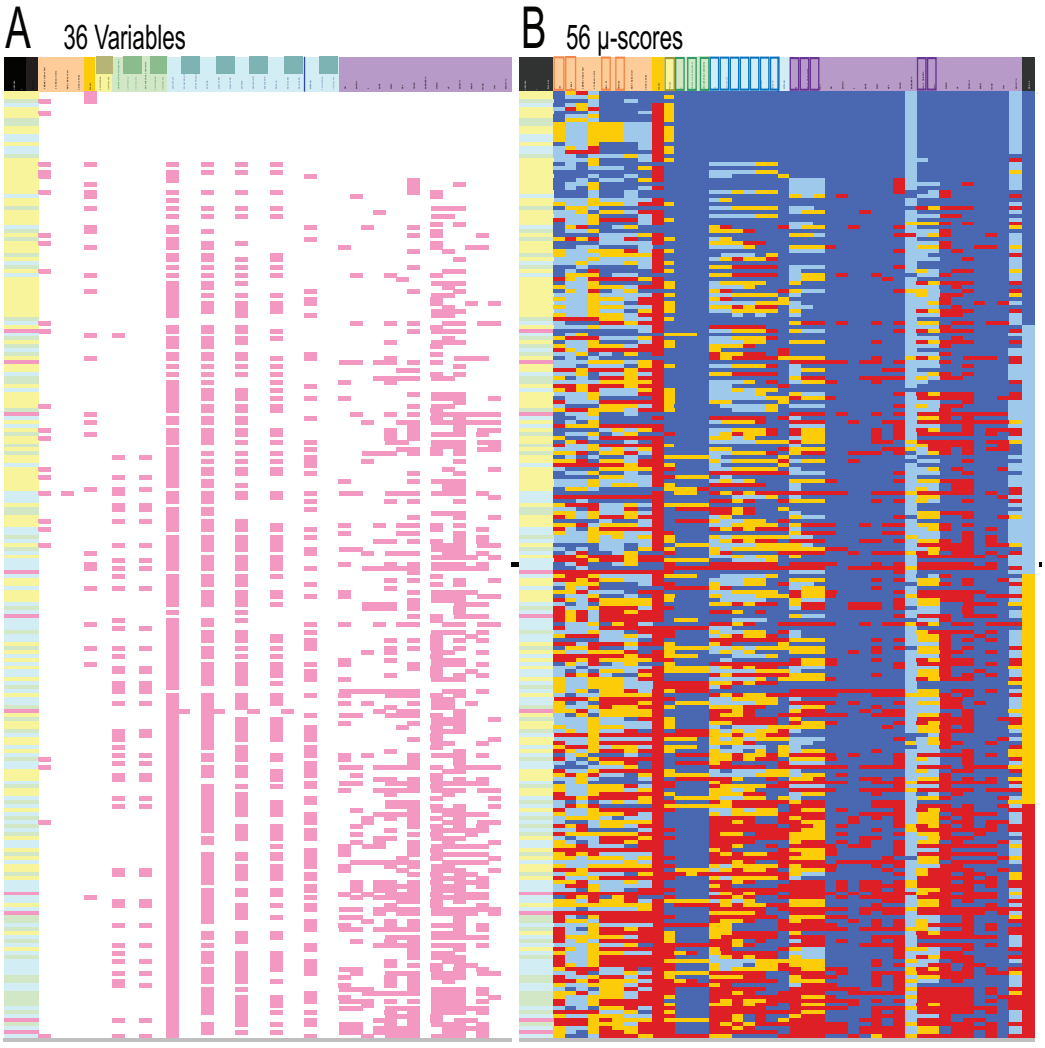


**Figure 6: FAPS Input/Output:** The column header colors indicate the FA phenotypes (orange: CB, yellow: LS, green: CA, blue: HM, purple: CM) *Part A:* Input data (36 variables by 239 subjects) with binary scores (white: 0, pink: 1) for the CA, HM, CM variables. Censored variables have dark rectangles in their column headers. *Part B*: µ-Scores for the 36 'leaves' and the 20 multivariate phenotypes (dark column header borders) (total: 56), including the global FAPS (black column header). Rows sorted by the global FAPS score (right most column) and colored according to quartiles of FA severity (red: severe, orange: less severe, light blue: mild, Blue: very mild).

To verify the hierarchical structure, multidimensional scaling (MDS) was performed (Torgerson 1952; Shepard 1962a, b; Kruskal 1964). MDS generates a map where the distances between variables are fitted to represent their correlations (Figure 7), which reconstitutes the structure of the FA phenotypes (Figure 1). How to calculate the hierarchical (UH) and the non-hierarchical (UNH) µ-scores is demonstrated in the Appendix.

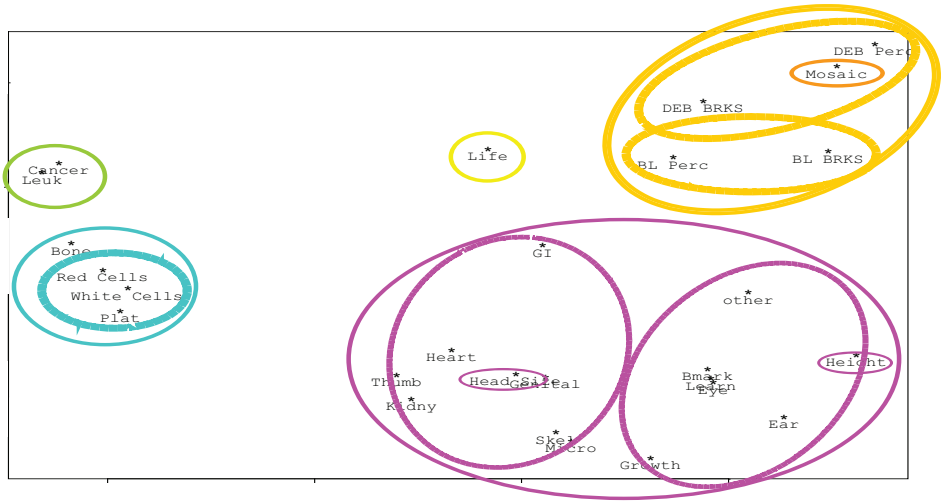


**Figure 7: Two-factor multidimensional scaling:** The correlation among the 27 univariate/interval μ-scores (Figure 6) with $\sqrt{1-r^2}$ as a dissimilarity measure

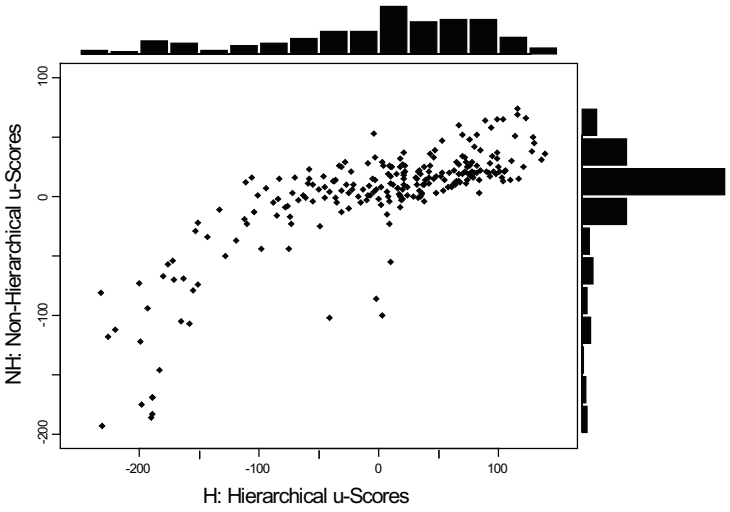


**Figure 8: Information content of hierarchical vs non-hierarchical μ-scores** Each point represents a subject. Hierarchical scores (H) have a higher information content (larger spread) and are more evenly distributed than non-hierarchical scores (NH)

Both μ-scores are intrinsically valid, but increasing information content by incorporating hierarchical information makes the distribution of scores based on partially ordered data more similar to the wider, uniform distribution of totally ordered data (Figure 8). The non-hierarchical (NH) scores are concentrated in a narrow range (0–40), while the hierarchical (H) scores are spread from −100 to 150.

The four 'outliers' in Figure 8 have very low NH scores (≈−100), while their H scores are closer to the median (≈0). These subjects all have CA, three of them died and the other had HM. Thus, few subjects are worse with respect to CA/HM. On the other hand, they lack any CM, so that no subject is better with respect to CM. H scores appropriately balance CA/HM vs CM. With NH scores, however, subjects with any CM are excluded from the comparisons against these four subjects, unless they are also higher with respect to CA/HM.

*Genotype-Phenotype Relationships*

Hierarchical μ-scores utilize knowledge of relationships between phenotypes. We examined the FA genotype-phenotype relationships by linking the FAPS to the *FANCA* mutation map. Among the FANCA subjects, the mutational events cover the whole gene (Figure 2); many subjects had mutations affecting the last seven regions (exons, introns).

For both the NH and the H FAPS, Figure 9 shows curves of the average FAPS among subjects by genotype at each gene region. The severity at each region frequently varies in all genotypes, producing distinct curves.

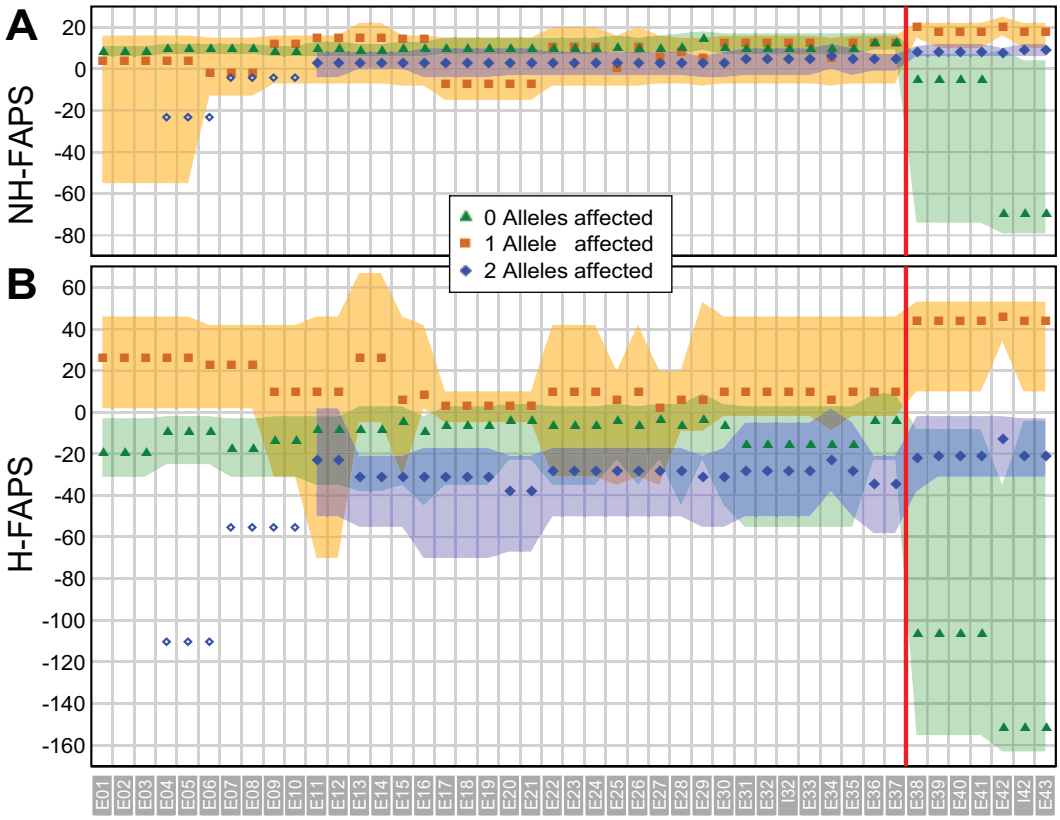


**Figure 9: FANCA severity profiles.** For each region (exon/intron) the severity scores for the 239 subjects were grouped by mutation status (0: no allele affected, 1: one allele affected, 2: both alleles affected; see Figure 2). The lines depict the median hierarchical (A) and non-hierarchical (B) FAPS for each mutation status. The shaded areas indicate the 67% confidence intervals (the non-parametric analogue to "±SEM" in parametric models). For estimates based on less then 6 subjects (dotted lines), confidence intervals are omitted. Note: the baseline indicates the median score among all FA subjects.

When applied to the FA data, Figure 9 shows that both methods extract the same features. With H-FAPS, however, the difference among the curves is more pro-

nounced, the '1' curve is higher than the '0' or '2' curves, the '0' curve is higher than the '2' curve until its drop in exon 38, where the '2' and '0' curve are more steeply in- and decreasing, respectively.

## DISCUSSION

### SCORING SYSTEM

We present an FA phenotype scoring system based on a multivariate, non-parametric statistical method (μ-scores) that can score subjects characterized by several ordinal variables. Here, the term 'ordinal' is used in its literal meaning, namely that the order among the outcomes is known (Wittkowski 1991). Ordered variables can be both discrete (ties are exact) and continuous (the order within ties is ambiguous) (Wittkowski 1998). As in linear model regression, discrete nominal variables can be included if split into binary 'dummy' variables.

With this extension, μ-scores can integrate knowledge about structure among the variables, several of which may be (interval) censored. This knowledge can be based on expert information on correlations between variables, using statistical methods, such as MDS. These scores are used to study genotype-phenotype relationships by relating *FANCA* mutations to 36 FA phenotype variables. Depending on the degree of detail sought, scores with varying levels of integration are obtained, ranging from 36 univariate u-scores to a single global μ-score (Figure 6). This approach offers several advantages over existing methods.

For example, μ-scores are 'intrinsically valid' by construction. Among two subjects, the subject with higher disease severity will always be assigned a higher μ-score. Thus, μ-scores can score phenotypes for diseases lacking a 'gold standard' to which conventional linear weight scores would need to be fit as an empirical 'justification' for the selection of a particular set of weights and transformations.

The proposed extension of μ-scores reduces the tendency of partial orderings to have increasing numbers of ambiguous pairwise orderings as the number of variables increases. To counter the resulting decrease in information content, the extension utilizes knowledge about the structure of the variables. Our results demonstrate the importance of retaining as much information content as possible. Hierarchical μ-scores discriminate better, thereby increasing the sensitivity for genotype-phenotype studies (see Figure 9). The difference between the curves is more pronounced with the H-FAPS than the NH-FAPS. Also, the H-FAPS differentiates better between the areas in the *FANCA* gene (i.e. 3' protein binding sites). This better separation of curves results from the higher information content of the H-FAPS.

Additionally, since the FAPS can be created 'ad hoc' for each study population under consideration and is currently not intended to be used with uncharacter-

ized patients, the training and target population are identical, so that there is no need to show that μ-score is valid when used with other populations, making them particularly useful in the field of 'personalized medicine' (Wittkowski 2003).

With an earlier intrinsically valid approach based on the marginal likelihood (MrgL) principle (Wittkowski 1992; Susser *et al.* 1998), one needs to generate all rank permutations compatible with the partial ordering to compute the average across the compatible rankings (up to a scale transformation). μ-Scores, being the average of the smallest and the largest rank across these rankings, converge against the MrgL scores, yet are computationally feasible (growing with the square of the number of subjects only, see Figure 3), while MrgL scores are np-hard (growing with the factorial of the number of subjects).

## GENOTYPE-PHENOTYPE RELATIONSHIPS

We present genotype-phenotype relationship data linking FAPS with *FANCA* mutations (Figure 2). Our data provide support for the method used to generate the FAPS and about the relationship of *FANCA* regions to the FAPS.

In Figure 9, the severity at each *FANCA* region varies by mutation status (0, 1, 2 affected alleles), suggesting that domains of the FANCA protein relate differently to severity. At exon 38, for instance, severity increases steeply in the curve for the single affected allele and decreases steeply in the curve for unaffected alleles, pointing to a gene region with potential protein binding sites and motifs (Garcia-Higuera *et al.* 1999; Otsuki *et al.* 1999; Huber *et al.* 2000; Otsuki *et al.* 2001; Otsuki *et al.* 2002; Ferrer *et al.* 2005; Yang *et al.* 2005; Medhurst *et al.* 2006; Oda *et al.* 2007). Our results based on the FAPS are consistent with previous findings based on experimental work. This suggests that μ-scores can help with detecting relevant phenomena.

Subjects with mutations in different regions (the '1' curve in Figure 9) have a higher FAPS and, among the 40 compound heterozygotes, the 25% (10) with two different predicted defective proteins have the highest FAPS. A reason for this seeming contradiction with reported results that homozygous mutations are more severe (Faivre *et al.* 2000) could be that univariate approaches tend to overlook the overall severity conferred by two different defective proteins causing different phenotypes. Molecular data can also be correlated with phenotype scores. For example, SNP or gene expression array data can be correlated to the FAPS to implicate loci or genes in disease severity or specific phenotypes (Song 2007). These correlations can also be used to generate hypotheses about the mechanisms of phenotype formation.

Resources for calculating μ-scores and related statistics are available from The Rockefeller University (see Appendix). To facilitate screening studies a Web service (http://muStat.rockefeller.edu) to a grid resource. provides capacities that promise to match computational challenges (Jorge Andrade 2006).

## Appendix

In the muStat package the function for R[1] and S-Plus.[2] `mu.PwO` generates Deuchler's pairwise (univariate) orderings, `mu.AND` combines them into a partial pairwise ordering, and `mu.Sums` computes scores and weights from a partial pairwise ordering. The pseudo-code below demonstrates calculation of the hierarchical (UH) and the non-hierarchical (UNH) μ-scores of Figure 7. The parentheses indicate the hierarchy and variables separated by colon indicate the bounds of interval censored observations.

```
frml <- "(
   ((BLnB,BLpC),((DEBnB,DEBpC),Mosaic)),
   LS0:LS1,
   (CA0:CA1,LK0:LK1),
   (HM0:HM1,(P0:P1,R0:R1,W0:W1),BMT0:BMT1),
   ((HZ%,(EAR,…)),(HT%,(BMRK,…)))
)"

x    <- importData(…)
PO   <- mu.PwO(x, frml)      # creates univariate pw orderings
                             # using censoring info from frml only

POH  <- mu.AND(PO,frml)      # creates part. ordering  w/hierarchy
PONH <- mu.AND(PO     )      # creates part. ordering wo/hierarchy

UH   <- mu.Sums(POH )$score  # creates scores  w/hierarchy
UNH  <- mu.Sums(PONH)$score  # creates scores wo/hierarchy
```

The statements

```
mu.Score(x,frml) and
mu.Sums( mu.AND(mu.PwO(x,frml), frml) )$score
```

are equivalent to:

```
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
```

[1] http://cran.r-project.org
[2] http://csan.insightful.com/PackageDetails.aspx?Package=muStat

## References

Alter, B.P., Greene M.H., Velazquez I., Rosenberg P.S. (2003). Cancer in Fanconi anemia. Blood 101(5): 2072.

An, P., Freedman B.I., Hanis C.L., Chen Y.D., Weder A.B., Schork N.J., Boerwinkle E., Province M.A., Hsiung C.A., Wu X.*et al.* (2005). Genome-wide linkage scans for fasting glucose, insulin, and insulin resistance in the National Heart, Lung, and Blood Institute Family Blood Pressure Program: evidence of linkages to chromosome 7q36 and 19q13 from meta-analysis. Diabetes 54(3): 909-14.

Auerbach, A.D. (1993). Fanconi anemia diagnosis and the diepoxybutane (DEB) test. Exp Hematol 21(6): 731-3.

Auerbach, A.D., Rogatko A., Schroeder-Kurth T.M. (1989). International Fanconi Anemia Registry: relation of clinical symptoms to diepoxybutane sensitivity. Blood 73(2): 391-6.

Callen, E., Casado J.A., Tischkowitz M.D., Bueren J.A., Creus A., Marcos R., Dasi A., Estella J.M., Munoz A., Ortega J.J.*et al.* (2005). A common founder mutation in FANCA underlies the world's highest prevalence of Fanconi anemia in Gypsy families from Spain. Blood 105(5): 1946-9.

Charames, G.S., Bapat B. (2003). Genomic instability and cancer. Curr Mol Med 3(7): 589-96.

De la Torre, C., Pincheira J., Lopez-Saez J.F. (2003). Human syndromes with genomic instability and multiprotein machines that repair DNA double-strand breaks. Histol Histopathol 18(1): 225-43.

Delbecq, A. (1975). Group techniques for program planning. . Glenview, IL: , Scott Foresman.

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

Dick, D.M., Aliev F., Wang J.C., Grucza R.A., Schuckit M., Kuperman S., Kramer J., Hinrichs A., Bertelsen S., Budde J.P.*et al.* (2008). Using dimensional models of externalizing psychopathology to aid in gene identification. Arch Gen Psychiatry 65(3): 310-8.

Faivre, L., Guardiola P., Lewis C., Dokal I., Ebell W., Zatterale A., Altay C., Poole J., Stones D., Kwee M.L.*et al.* (2000). Association of complementation group and mutation type with clinical outcome in fanconi anemia. European Fanconi Anemia Research Group. Blood 96(13): 4064-70.

Ferrer, M., Rodriguez J.A., Spierings E.A., de Winter J.P., Giaccone G., Kruyt F.A. (2005). Identification of multiple nuclear export sequences in Fanconi anemia group A protein that contribute to CRM1-dependent nuclear export. Hum Mol Genet 14(10): 1271-81.

Freimer, N., Sabatti C. (2003). The human phenome project. Nat Genet 34(1): 15-21.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200): 675-701.

Garcia-Higuera, I., Kuang Y., Naf D., Wasik J., D'Andrea A.D. (1999). Fanconi anemia proteins FANCA, FANCC, and FANCG/XRCC9 interact in a functional nuclear complex. Mol Cell Biol 19(7): 4866-73.

Giampietro, P.F., Adler-Brecher B., Verlander P.C., Pavlakis S.G., Davis J.G., Auerbach A.D. (1993). The need for more accurate and timely diagnosis in Fanconi anemia: a report from the International Fanconi Anemia Registry. Pediatrics 91(6): 1116-20.

Hafen, G.M., Ranganathan S.C., Robertson C.F., Robinson P.J. (2006). Clinical scoring systems in cystic fibrosis. Pediatr Pulmonol 41(7): 602-17.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19: 293-325.

Huber, P.A., Medhurst A.L., Youssoufian H., Mathew C.G. (2000). Investigation of Fanconi anemia protein interactions by yeast two-hybrid analysis. Biochem Biophys Res Commun 268(1): 73-7.

John, J.P., Arunachalam V., Ratnam B., Isaac M.K. (2008). Expanding the schizophrenia phenotype: a composite evaluation of neurodevelopmental markers. Compr Psychiatry 49(1): 78-86.

Jorge Andrade, L.B., Mathias Uhlén, Jacob Odeberg (2006). Using Grid technology for computationally intensive applied bioinformatics analyses. In Silico Biology 6(0046).

Klempir, J., Klempirova O., Spackova N., Zidovska J., Roth J. (2006). Unified Huntington's disease rating scale: clinical practice and a critical approach. Funct Neurol 21(4): 217-21.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29(1): 1-27.

Kutler, D.I., Auerbach A.D. (2004). Fanconi anemia in Ashkenazi Jews. Fam Cancer 3(3-4): 241-8.

Kutler, D.I., Singh B., Satagopan J., Batish S.D., Berwick M., Giampietro P.F., Hanenberg H., Auerbach A.D. (2003). A 20-year perspective on the International Fanconi Anemia Registry (IFAR). Blood 101(4): 1249-56.

Levitus, M., Rooimans M.A., Steltenpool J., Cool N.F., Oostra A.B., Mathew C.G., Hoatlin M.E., Waisfisz Q., Arwert F., de Winter J.P.*et al.* (2004). Heterogeneity in Fanconi anemia: evidence for 2 new genetic subtypes. Blood 103(7): 2498-503.

Levran, O., Attwooll C., Henry R.T., Milton K.L., Neveling K., Rio P., Batish S.D., Kalb R., Velleuer E., Barral S.*et al.* (2005). The BRCA1-interacting helicase BRIP1 is deficient in Fanconi anemia. Nat Genet 37(9): 931-3.

Medhurst, A.L., Laghmani E.H., Steltenpool J., Ferrer M., Fontaine C., de Groot J., Rooimans M.A., Scheper R.J., Meetei A.R., Wang W.*et al.* (2006). Evidence for subcomplexes in the Fanconi anaemia pathway. Blood.

Meetei, A.R., Sechi S., Wallisch M., Yang D., Young M.K., Joenje H., Hoatlin M.E., Wang W. (2003). A multiprotein nuclear complex connects Fanconi anemia and Bloom syndrome. Mol Cell Biol 23(10): 3417-26.

Meetei, A.R., Medhurst A.L., Ling C., Xue Y., Singh T.R., Bier P., Steltenpool J., Stone S., Dokal I., Mathew C.G.*et al.* (2005). A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. Nat Genet 37(9): 958-63.

Meigs, J.B., Manning A.K., Fox C.S., Florez J.C., Liu C., Cupples L.A., Dupuis J. (2007). Genome-wide association with diabetes-related traits in the Framingham Heart Study. BMC Med Genet 8 Suppl 1: S16.

Meyn, M.S. (1997). Chromosome instability syndromes: lessons for carcinogenesis. Curr Top Microbiol Immunol 221: 71-148.

Nakanishi, K., Yang Y.G., Pierce A.J., Taniguchi T., Digweed M., D'Andrea A.D., Wang Z.Q., Jasin M. (2005). Human Fanconi anemia monoubiquitination pathway promotes homologous DNA repair. Proc Natl Acad Sci U S A 102(4): 1110-5.

Oda, T., Hayano T., Miyaso H., Takahashi N., Yamashita T. (2007). Hsp90 regulates the Fanconi anemia DNA damage response pathway. Blood.

Offit, K., Levran O., Mullaney B., Mah K., Nafa K., Batish S.D., Diotti R., Schneider H., Deffenbaugh A., Scholl T.*et al.* (2003). Shared genetic susceptibility to breast cancer, brain tumors, and Fanconi anemia. J Natl Cancer Inst 95(20): 1548-51.

Otsuki, T., Kajigaya S., Ozawa K., Liu J.M. (1999). SNX5, a new member of the sorting nexin family, binds to the Fanconi anemia complementation group A protein. Biochem Biophys Res Commun 265(3): 630-5.

Otsuki, T., Nagashima T., Komatsu N., Kirito K., Furukawa Y., Kobayashi Si S., Liu J.M., Ozawa K. (2002). Phosphorylation of Fanconi anemia protein, FANCA, is regulated by Akt kinase. Biochem Biophys Res Commun 291(3): 628-34.

Otsuki, T., Furukawa Y., Ikeda K., Endo H., Yamashita T., Shinohara A., Iwamatsu A., Ozawa K., Liu J.M. (2001). Fanconi anemia protein, FANCA, associates with BRG1, a component of the human SWI/SNF complex. Hum Mol Genet 10(23): 2651-60.

Popper, K.R. (1959). The Logic of Scientific Discovery. New York, Basic Books.

Reid, S., Schindler D., Hanenberg H., Barker K., Hanks S., Kalb R., Neveling K., Kelly P., Seal S., Freund M.*et al.* (2007). Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nat Genet 39(2): 162-4.

Rosenberg, P.S., Huang Y., Alter B.P. (2004). Individualized risks of first adverse events in patients with Fanconi anemia. Blood 104(2): 350-5.

Savino, M., Borriello A., D'Apolito M., Criscuolo M., Del Vecchio M., Bianco A.M., Di Perna M., Calzone R., Nobili B., Zatterale A.*et al.* (2003). Spectrum of FANCA mutations in Italian Fanconi anemia patients: identification of six novel alleles and phenotypic characterization of the S858R variant. Hum Mutat 22(4): 338-9.

Shepard, R.N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. Psychometrika 27(2): 125-140.

Shepard, R.N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. Psychometrika 27(3): 219-246.

Smogorzewska, A., Matsuoka S., Vinciguerra P., McDonald E.R., 3rd, Hurov K.E., Luo J., Ballif B.A., Gygi S.P., Hofmann K., D'Andrea A.D.*et al.* (2007). Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. Cell 129(2): 289-301.

Song, T.C., C; Wittkowski, KM (2007). Screening for gene expression profiles and epistasis between diplotypes with S-Plus on a grid. Statistical Computing and Graphics 18: 20-5.

Susser, E., Desvarieux M., Wittkowski K.M. (1998). Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. American Journal of Public Health 88(4): 671-674.

Tamary, H., Dgany O., Toledano H., Shalev Z., Krasnov T., Shalmon L., Schechter T., Bercovich D., Attias D., Laor R.*et al.* (2004). Molecular characterization of three novel Fanconi anemia mutations in Israeli Arabs. Eur J Haematol 72(5): 330-5.

Taylor, A.M. (2001). Chromosome instability syndromes. Best Pract Res Clin Haematol 14(3): 631-44.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. Psychometrika 17(4): 401-419.

Vanscoy, L.L., Blackman S.M., Collaco J.M., Bowers A., Lai T., Naughton K., Algire M., McWilliams R., Beck S., Hoover-Fong J.*et al.* (2007). Heritability of lung disease severity in cystic fibrosis. Am J Respir Crit Care Med 175(10): 1036-43.

Wang, W. (2007). Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. Nat Rev Genet 8(10): 735-48.

Wittkowski, K.M. (1991). A structured visual language for a knowledge-based front-end to statistical analysis systems in biomedical research. Computer Methods and Programs in Biomedicine 35(1): 59-67.

Wittkowski, K.M. (1992). An extension to Wittkowski. Journal of the American Statistical Association 87: 258.

Wittkowski, K.M. (1998). Versions of the sign test in the presence of ties. Biometrics 54: 789–791.

Wittkowski, K.M. (2003). Novel Methods for Multivariate Ordinal Data applied to Genetic Diplotypes, Genomic Pathways, Risk Profiles, and Pattern Similarity. Computing Science and Statistics 35: 626-646.

Wittkowski, K.M., Lee E., Nussbaum R., Chamian F.N., Krueger J.G. (2004). Combining several ordinal measures in clinical studies. Stat Med 23(10): 1579-92.

Yagasaki, H., Oda T., Adachi D., Nakajima T., Nakahata T., Asano S., Yamashita T. (2003). Two common founder mutations of the fanconi anemia group G gene FANCG/XRCC9 in the Japanese population. Hum Mutat 21(5): 555.

Yang, Y.G., Herceg Z., Nakanishi K., Demuth I., Piccoli C., Michelon J., Hildebrand G., Jasin M., Digweed M., Wang Z.Q. (2005). The Fanconi anemia group A protein modulates homologous repair of DNA double-strand breaks in mammalian cells. Carcinogenesis 26(10): 1731-40.