# Cluster Analysis: A Comparison of Different Similarity Measures for SNP Data

Tina Müller, Silvia Selinski and Katja Ickstadt
SFB 475, Statistics Department, University of Dortmund

**Abstract**

The issue of suitable similarity measures for a particular kind of genetic data - so called SNP data - arises, e.g., from the GENICA (The Interdisciplinary Study Group on Gene Environmental Interactions and Breast Cancer in Germany) case-control study of sporadic breast cancer. The GENICA study aims to investigate the influence and interaction of single nucleotide polymorphic (SNP) loci and exogenous risk factors. It is very unlikely that there exists one main effect, say only one polymorphism, being responsible for such a complex disease as sporadic breast cancer as the role of a single gene within the carcinogenic process is limited (Garte, 2001). Nevertheless, it is assumed that a number of interacting SNPs in combination with certain environmental risk factors increase the individual susceptibility.

The search for SNP patterns in the present data set may be performed by a variety of clustering and classification approaches. Here we consider the problem of adequate similarity measures for variables or subjects as an indispensable basis for a further cluster analysis. The term 'similarity' is still vague for SNP data. A main problem arises by the general structure of such data sets: the proportion of hetero- or homozygous SNPs is rather small compared with the homozygous reference sequence. Thus, the relevant information of combinations of genetic alterations is often masked by a huge amount of common occurrences of homozygous reference types. Therefore, we examine different similarity measures, conventional ones as well as new coefficients which we created especially for SNP data. Furthermore, we compare the resulting partitions with each other adapting the clustering of clustering methods of Rand (1971) for different similarity measures.

Key Words: cluster analysis, clustering clustering methods, GENICA, similarity, single nucleotide polymorphism, sporadic breast cancer

# 1 Introduction

Everybody's DNA is unique. Though we share 99.9% of our DNA there remain about 3 million differences between two individuals. Most of the genetic variation consists of single nucleotide changes. Such a single base exchange - or a deletion or addition of base pairs at one gene locus - is called *single nucleotide polymorphism* (SNP) if it is present in at least 1% of the population.
Some SNPs affect the outcome products coded by the corresponding gene and are therefore considered, in interaction with other SNPs and in combination with further factors, to alter the risk for developing a particular disease.
This paper is based on the case-control study GENICA which investigates genetic and environmental factors with respect to their impact on the risk of developing sporadic breast cancer. The central question we examine is whether we can detect subgroups of SNP loci which seem to interact and whether these groups vary between the cases and controls or not. In other words, we try to divide the variables into groups whose elements are similar to each other.
The term similarity is still vague for SNP data. Usually a SNP occurs rather infrequently compared with the homozygous reference type of a gene locus observing a number of individuals. As a consequence the relevant information for comparing two SNP loci - if two SNPs share a heterozygous or homozygous variant - is often masked by a huge amount of 0-0-matches i.e., a combination of two homozygous reference types.
In this paper we examine different similarity measures (see section 3.2), conventional ones, e.g. Pearson's coefficient of contingency, the Simple Matching Coefficient and Jaccard's Coefficient, as well as new coefficients which we created especially for SNP data (cf. Selinski and Ickstadt, 2005). In particular, we compare the resulting partitions with each other, trying to find groups of measures having the same effect and we evaluate the differences between clusterings of cases and controls.
After short summaries of the genetic background and the study the GENICA data set as well as simulated data are introduced in section 2. The applied methods containing similarity measures, cluster algorithm and clustering clustering methods by Rand are described in section 3. We analyse the performances of the different similarity measures on the real data set, compare them according to our adaption of Rand's method and evaluate the differences between cases and con-

trols in section 4. To confirm our results we cluster the simulated data in section 5. Section 6 gives a final discussion as well as an outlook.

# 2 Background

The search for interactions between different SNP loci plays an important role in cancer research. The reason is quite clear as it is very unlikely that only one SNP or one variant of a gene shows a main effect which can claim responsibility for developing a disease as complex as cancer. This fact is not astonishing because the role of a single gene within the carcinogenic process is limited (Garte, 2001). Furthermore, the effect of a single base pair variation on a metabolic process is usually also restricted depending on its position within a gene or a regulating sequence and, hence, on its impact on the gene product of gene regulation and the role of the associated gene product in the metabolic pathway. For detailed information about the genetic background see Snustad and Simmons (1999).

Still, if one individual carries several gene variations, maybe combined with a certain exposure to critical substances, this combination of factors may change the risk of developing cancer.

For most SNPs it is not clear yet which impact they have on the associated gene products and their function. However, different genes (and the enzymes they code) can be related to different pathways. A pathway is the field in which a specific gene product participates, e.g., the pathway of drug metabolism. So from the assumption that a SNP alters the respective enzyme, it can be deducted which part of the metabolism might be affected.

The data set for the present paper is provided by the German GENICA study on sporadic breast cancer. It aims to investigate the influence of SNPs in combination with epidemiological and clinical factors on the risk of breast cancer in women.

## 2.1 GENICA Study

The GENICA Study is part of the German Human Genome Project (DHGP) and aims to find relevant interactions of potential risk factors which alter the susceptibility for breast cancer. It is an aged-matched population-based case-control study.

The recruitment of test persons was carried out in two phases in the greater Bonn

region between 2000 and 2002. In the second phase it should include the data of 1000 cases (recruited in hospitals and selected by several criteria) as well as of 1000 controls from the same region. Our analysis bases on the first period of recruitment and comprises 1260 women.
All test persons were interviewed as well as genotyped using PCR and MALDI-TOF (Pusch et al., 2002). In this paper, we concentrate on the genetic information.

## 2.2 GENICA Data

The data we use consist of 68 SNP variables measured in 610 cases and 650 controls. They are coded according to the number of occurring polymorphisms, 0 if they show the homozygous reference type, 1 if the variant occurs in one chromosome and 2 if the variant is present in both chromosomes. A special feature of these data is that most observations show the homozygous reference type.
All empirical frequency distributions of the SNPs meet the Hardy-Weinberg equilibrium (HWE) in the control group, at least if the variant categories contain a reasonable number of entries to guarantee that the asymptotic $\chi^2$-test for deviation holds (for details on HWE testing see Hosking et al. (2004)).
Some values are missing, caused, e.g., by detection problems during the laboratory work. We take only women with five or less missings into further account. That means that 1165 women (546 cases and 619 controls) remain in the data set which corresponds to 92.46%.
In this paper, all computations will be done for cases and controls separately. The names of the SNPs are encoded using numbers for the gene they belong to and their specific position on the gene.

## 2.3 Simulated Data

In addition to the real data set from the GENICA study we use simulated data to evaluate the effects of the similarity coefficients. The data are generated by a software program called SNaP (Nothnagel, 2002). It simulates haplotypes first. A haplotype is a gene sequence on one chromosome, whereas genotypes do not contain the information how the SNPs are distributed on the two chromosomes. SNaP simulates haplotypes by employing the idea that there are haplotype blocks with high linkage disequilibrium within the block. A limited number of different

haplotypes per block is generated and stored in a pool. The genotype of a person is derived from two haplotypes per block which are randomly drawn from this pool.
We simulate 35 SNP variables in total from five blocks labelled A to E for 500 cases and 500 controls. The disease status of each person depends on four causative SNPs, each of them belonging to a different block. Three of them contribute in a dominant way and cause the disease as homo- and heterozygous variants. The fourth SNP only impacts the disease status if it shows a homozygous variant genotype. A person carrying all four SNPs in the described states is assigned to the case collective, otherwise the person contributes to the control group.

# 3 Methods

In this paper we compare the performance of similarity measures for cluster analysis introduced by Selinski and Ickstadt (2005). First we will give a short outline of the similarity measures and clustering algorithms in general and then continue introducing the considered measures. Finally we present Rand's method (1971) for the comparisons of clusterings.

## 3.1 Similarity and Distance

The data used in the following context consist of $n$ objects and $m$ variables, given as an $n \times m$ matrix. The rows represent the objects and the columns stand for the variables (Fahrmeir et al., 1996). $V = \{V_1, \dots, V_m\}$ describes the set of variables. To define similarity between elements of $V$, we introduce a function $S : V \times V \to \mathbb{R}$ which meets the first three of the following assumptions:

1. $S(V_k, V_l) > S(V_o, V_l)$, if $V_l$ is more similar to $V_k$ than to $V_o$, $V_k \neq V_o$; $V_k, V_l, V_o \in V$ — comparability
2. $S(V_i, V_j) = S(V_j, V_i)$, $i, j = 1, \dots, m$ — symmetry
3. $S(V_i, V_j) \leq S(V_i, V_i)$, $i, j = 1, \dots, m$ — natural order
4. $S(V_i, V_j) \geq 0$, $i, j = 1, \dots, m$ — positivity
5. $S(V_i, V_i) = 1$, $i = 1, \dots, m$ — normality.

Assumptions 4 and 5 are often useful, though not necessary for $S$ for being a similarity measure.

| | **Variable $V_l$** | | | | |
|---|---|---|---|---|---|
| **Variable $V_k$** | 1 | 2 | $\cdots$ | $q$ | $\sum$ |
| 1 | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1q}$ | $n_{1.}$ |
| 2 | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2q}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $p$ | $n_{p1}$ | $n_{p2}$ | $\cdots$ | $n_{pq}$ | $n_{p.}$ |
| $\sum$ | $n_{.1}$ | $n_{.2}$ | $\cdots$ | $n_{.q}$ | $n$ |

**Table 1:** Contingency table

In practical investigations, distances rather than similarities are of interest. For categorical data, the similarity $S$ is computed first and then transformed into a distance $D$ (Cox and Cox, 2001). Large similarities correspond to small distances and vice versa. Therefore we use the following transformation if $S \in [0, 1]$:

$$D(V_k, V_l) = 1 - S(V_k, V_l), \qquad \text{for all } V_k, V_l \in V. \tag{1}$$

If $S \notin [0, 1]$, we add the absolute value of the minimum to $S$ (if the minimum is negative, otherwise skip this step) and apply

$$D(V_k, V_l) = 1 - \frac{S(V_k, V_l)}{\max S(V_i, V_j)}, \qquad \text{for all } V_k, V_l \in V, \quad i, j = 1, \ldots, m. \tag{2}$$

The returned distance yields $D \in [0, 1]$.

## 3.2 Similarity Measures

The introduction of the conventional similarity measures will be quite short, for details see Anderberg (1973) and Cox and Cox (2001). Variables $V_k$ and $V_l$ with $p$ and $q$ categories, respectively, will be used as examples to explain the principles of the measures (see Table 1). Using the theory of independence analysis, we consider the elements of $V$ as random variables and choose two $\chi^2$-statistics to be the first candidates for meaningful similarity measures. With

$$e_{ab} = \frac{n_{a.} \cdot n_{.b}}{n}, \quad a = 1, \ldots, p \ , \ b = 1, \ldots, q$$

we receive a first estimate for the level of independence between the two variables under investigation by:

$$\chi^2 = \sum_{a=1}^{p} \sum_{b=1}^{q} (n_{ab} - e_{ab})^2 / e_{ab} = n \left( \sum_{a=1}^{p} \sum_{b=1}^{q} \frac{n_{ab}^2}{n_{a.} n_{.b}} - 1 \right).$$

To standardise $\chi^2$, we use $n$ as a normalising factor:

$$\phi^2 = \frac{\chi^2}{n}.$$

For the justification of the $\chi^2$-statistics, the assumptions of random sample, independence and unambiguousness of assignment must be met. The first assumption holds only for the controls as all cases who qualified for the study were included. The second requirement is fulfilled as no relatives were accepted in the study, so we can assume independence for the different genetic profiles. The third assumption is also true. Because the first assumption is not completely met, we have to question the justification of applying the $\chi^2$-statistics.

Other problems arise with $\chi^2$-measures: As for the considered data set it may happen that some variables are treated as constants. This occurs either if all variants of a variable are compared to missing values of another variable and therefore do not contribute to the calculations or if the data set contains monomorphic SNPs. It does not make sense to apply a $\chi^2$-statistics which calculates expected values for a constant.

To avoid this problem, we exclude monomorphic SNPs from the data. For the variables that cause problems due to missing values we draw random samples from their marginal distributions estimated by the relative frequencies and replace the problematic missings.

We consider two $\chi^2$-coefficients for our analysis, Pearson's corrected contingency coefficient $\mathrm{P}_C$ (cf. Sachs (1999)) and Cramèr's C (cf. Anderberg (1973)) given by:

$$\mathrm{P}_C = \sqrt{\frac{\min(p,q)}{\min(p,q)-1}}\left(\frac{\phi^2}{1+\phi^2}\right)^{1/2} \text{ and } \mathrm{C} = \left(\frac{\phi^2}{\min(p-1,q-1)}\right)^{1/2}.$$

For our SNP data, the numbers of categories are $p = q = 3$.

A second approach of defining a similarity measure introduces matching coefficients. They count the number of objects with the same outcome for both variables as well as the number of non-corresponding observations and use different relations and weights to evaluate these counts. As for $\chi^2$-coefficients, the contingency table provides the basis of the comparisons. Table 2 shows a new, more specific labelling of the different cells, in comparison to Table 1. We have already reduced the size of the variable domains to 3 as this corresponds to our data. The categories represent the number of SNPs present and correspond to the coding of

| | | $V_l$ 0 | 1 | 2 | |
|---|---|---|---|---|---|
| | 0 | $a$ | $b$ | $c$ | |
| $V_k$ | 1 | $d$ | $e$ | $f$ | |
| | 2 | $g$ | $h$ | $i$ | |
| | | | | | $n$ |

**Table 2:** $3 \times 3$ - contingency table for matching coefficients

variables described in section 2.2. Furthermore, we define $m^+ := a + e + i$ as the number of matches and $m^- := b+c+d+f+g+h$ as the number of mismatches.

All matching coefficients used in our further analysis can be found in Table 3. $S_D$ was introduced by Dice (1945). $S_K$, $S_{K00}$ $S_J$, and $S_{RT00}$ are taken from Anderberg (1973), $S_{SM}$, $S_{RT}$ and $S_{RR}$ from Cox and Cox (2001) and $S_H$ and $S_{SoSn}$ from Sokal and Sneath (1963). $S_{H00}$ resembles $S_H$, but with no consideration of 0-0-matches.

The treatment of the 0-0-matches (entry $a$ in Table 2) decides on the partition of the coefficients into Groups 1 to 3. The coefficients in Group 1 do not pay special attention to the 0-0-matches. Therefore, SNP variables with a high amount of mutual homozygous reference categories yield higher similarities than comparisons with less matches in total but more common SNPs.

The similarity measures in Group 2 do not count the 0-0-matches at all which seems to settle the shortcoming of the coefficients of Group 1 at first sight. But they have two disadvantages: If all observations of a comparison of two monomorphic SNPs are 0-0-matches, the denominator of the coefficients is 0 and the calculation impossible. On the other hand, variables which are concordant with each other except for a small number of mismatches reach minimal similarity if all matches are 0-0-matches. Even if the information about a common homozygous reference category is less helpful than information about variants, it does not justify this result.

The coefficient of Russell and Rao ignores the 0-0-matches in the nominator, but takes them into account in the denominator. So it shares the bad feature of the measures from Group 2.

As an attempt to avoid the disadvantages of Groups 1 and 2 described above we introduce the newly created coefficients in Group 3. We employed the idea

| Symbol | Name | Coefficient | |
|---|---|---|---|
| $S_{SM}$ | Simple Matching | $\frac{m^+}{m^+ + m^-}$ | Group 1 |
| $S_{SoSn}$ | Sokal & Sneath | $\frac{2m^+}{2m^+ + m^-}$ | Group 1 |
| $S_{RT}$ | Rogers & Tanimoto I | $\frac{m^+}{m^+ + 2m^-}$ | Group 1 |
| $S_K$ | Kulczynski I | $\frac{m^+}{m^-}$ | Group 1 |
| $S_H$ | Hamann I | $\frac{m^+ - m^-}{m^+ + m^-}$ | Group 1 |
| $S_J$ | Jaccard | $\frac{m^+ - a}{(m^+ - a) + m^-}$ | Group 2 |
| $S_D$ | Dice | $\frac{2(m^+ - a)}{2(m^+ - a) + m^-}$ | Group 2 |
| $S_{RT00}$ | Rogers & Tanimoto II | $\frac{m^+ - a}{(m^+ - a) + 2m^-}$ | Group 2 |
| $S_{K00}$ | Kulczynski II | $\frac{m^+ - a}{m^-}$ | Group 2 |
| $S_{H00}$ | Hamann II | $\frac{(m^+ - a) - m^-}{(m^+ - a) + m^-}$ | Group 2 |
| $S_{RR}$ | Russell & Rao | $\frac{m^+ - a}{m^+ + m^-}$ | |
| $S_{QP}$ | Quarterprop | $\frac{m^+ - \frac{3}{4}a}{(m^+ - \frac{3}{4}a) + 2m^-}$ | Group 3 |
| $S_{Prop}$ | Proportions | $\frac{\frac{1}{5}a + 2e + 4i}{\frac{1}{5}a + 2e + 4i + m^-}$ | Group 3 |
| $S_{Mis12}$ | Mismatch12 | $\frac{\frac{1}{4}a + 2e + 4i + \frac{1}{2}(f+h)}{\frac{1}{4}a + 2e + 4i + \frac{1}{2}(f+h) + (b+c+d+g)}$ | Group 3 |
| $S_{Mis01}$ | Mismatch01 | $\frac{\frac{1}{4}a + 2e + 4i + \frac{1}{2}(b+d)}{\frac{1}{4}a + 2e + 4i + \frac{1}{2}(b+d) + (c+f+g+h)}$ | Group 3 |

**Table 3:** Matching coefficients

of weighting the different kinds of matches according to their assumed biological relevance (Selinski and Ickstadt, 2005). As a first step Quarterprop weights the frequent homozygous reference matches, which carry only moderate information, $\frac{1}{4}$. So only a fraction of their amount will be counted which reduces the effect of masking rare information.

Proportions takes this idea one step further. Additionally to scaling down the 0-0-matches by $\frac{1}{5}$, the heterozygous and homozygous variant matches get higher weights (2 and 4, respectively) to stress their value for the comparisons.

In addition, the last two coefficients allow the input of information about dominance or recessiveness of SNPs. Mismatch12 weights a comparison that consists of one heterozygous and one homozygous variant with half its amount as a match (dominance), for at least one chromosome contains a SNP in each of the loci. Mismatch01 works the same way, but with the reference types instead of variance (recessiveness). Which of the two measures should be applied depends on the given data.

### 3.3 Cluster Method

We choose an agglomerative hierarchical cluster algorithm to divide our data into groups, in particular average linkage.

After the description of similarity $S$ and distance $D$ for variables, we introduce s and d as proximity labels for clusters. Suppose every variable is regarded as a cluster with only one element. After applying the coefficient on the data, we obtain a similarity matrix and transform it into a distance matrix. Then the algorithm starts using the following steps:

1. Fuse the two clusters with the smallest distance d.
2. Recompute the distances for the newly formed group to all remaining clusters.
3. Iterate steps 1 and 2 until all variables lie in one big cluster.

For the computation of distances between two groups $G_r$ and $G_t$ with $m_r$ and $m_t$ elements, respectively, we use the average linkage cluster method, i.e.

$$\mathrm{s}(G_t, G_r) = \frac{1}{m_t m_r} \sum_{V_i \in G_t} \sum_{V_j \in G_r} S(V_i, V_j), \qquad \text{with } i = 1, \ldots, m_t, \quad j = 1, \ldots, m_r.$$

For the transformation from s into distances d use Equations (1) and (2) on page 6.

## 3.4 Rand: Clustering Clustering Methods

In most cases, different similarity measures will produce different partitions of the observed variables. With the given statistical means it is not possible to decide which of these partitions is considered best or correct. Even further information from the scientific background will only help to interpret the results, but not to evaluate them mathematically.

Still, similarity of different cluster results can be compared by regarding the partitions as objects to which a certain similarity will be assigned. There exist numerous suggestions how to do it. Rand (1971) proposed an intuitive and simple method for the clustering of clustering. He applied his method for results found by different cluster methods. Because we use the same data and the same clustering algorithm for all computations, the differences between clusterings presented in this paper only depend on the choice of the similarity coefficient. Thus, we use the result of the clustering of clusterings as a direct comparison between the different similarity measures.

As a basis for his computations, Rand counts the number of pairs of variables which are assigned either to the same or to different groups in clustering $C_k$ and $C_l$. For $m$ variables $N = \binom{m}{2}$ different pairs $(V_i, V_j)$, $i, j = 1, \ldots, m$ and $i \neq j$ exist. For further analysis the feature 'Grouping' is introduced. It takes value 1 for each pair $(V_i, V_j)$ which is assigned to the same group and value 2 if the two variables lie in different groups.

The joint distribution of 'Grouping' of two clusterings is presented in a contingency table (cf. Table 4). Rand's idea of quantifying a similarity for this situation uses the Simple Matching Coefficient. All matches ($A$ and $D$) are divided by $N$,

| | | **Grouping $C_l$** 1 | 2 | |
|---|---|---|---|---|
| **Grouping $C_k$** | 1 | $A$ | $B$ | |
| | 2 | $C$ | $D$ | |
| | | | | $N$ |

**Table 4:** Contingency table for the comparison of two clusterings

the total number of possible pairs:

$$R = \frac{A + D}{N}.$$

Thus, we obtain a similarity matrix based on the comparisons between all given clusterings. The next step contains a standard clustering (calculating distances, running the cluster algorithm etc.) on these results including a dendrogram to present the outcome.
We denote by r (instead of d) the distance of two clusterings according to the method of Rand.

### 3.5 Number of Classes

As we use hierarchical methods to cluster the SNP variables, we do not have to specify the number $g$ of clusters in advance, but get partitions for every choice of $g = 1, \ldots, 68$. However, for the comparison of the different measures we need to choose a certain number $g_c$ of clusters as the method by Rand employs the accordance of allocation of variables into groups.
The biological background does not give useful hints for a sensible choice of number of classes. The only given number is the number of different pathways in the study (=10). It is arguable if SNPs belonging to the same pathway are likely to display a similar pattern over many patients. In the absence of further useful indications, we chose 10 as the number of clusters.
The graphical illustration of the clustering results, the dendrogram, yields much more information. If it shows a stepwise structure (cf. Figure 2), meaning that the algorithm adds variables one by one to a single big cluster, then choosing two clusters yields one class containing only one variable and a second containing all the others. Such partitions do not carry much information. Sometimes several subgroups are already visible in the dendrogram (cf. Figure 3), so a meaningful choice of cluster numbers should consider them. On the whole, two clusters picture a very rough division for 68 SNPs, whereas more than 12 classes complicate the interpretation.
If the dendrogram provides insufficient information, we use the following guidelines to choose the number of clusters:

1. Choose $g_c$ as the highest number from the set of clusterings which do not contain clusters with only one element (maximum number = 12).

2. If all partitions contain clusters with one element choose the one with the smallest proportion of single element clusters.

3. If 1 and 2 cannot be applied, choose $g_c = 10$.

# 4 Application to the GENICA - Data

All calculations and figures were done using the software package R 1.8.0. We treat cases and controls separately, i.e., we obtain two clusterings for every similarity measure.
After describing the outcomes of the 17 measures from section 3 in section 4.1, we compare the different clusterings (cf. section 4.2) and the difference between the two collectives (cf. section 4.3) using Rand's method.

## 4.1 Different Clusterings

The dendrograms corresponding to $\mathrm{P}_C$ and C are similar for both coefficients, so we only display the clustering of the cases received by using $\mathrm{P}_C$ (Figure 1).

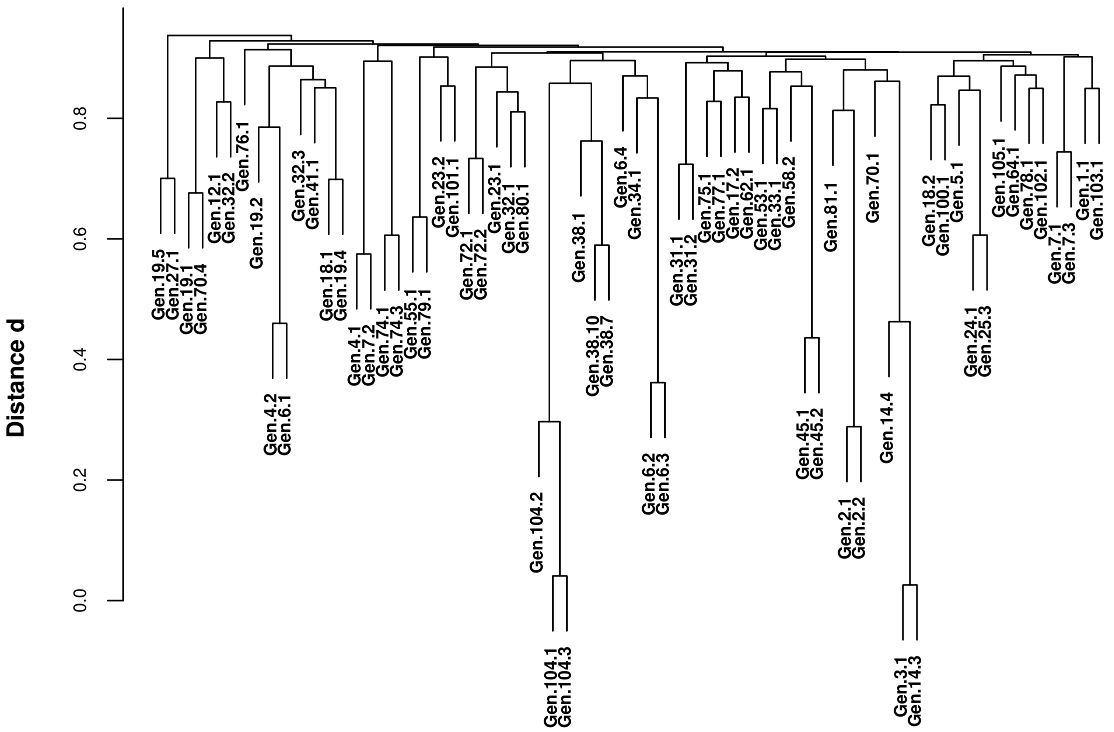


**Figure 1:** Clustering of cases obtained by Pearson's $\mathrm{P}_C$

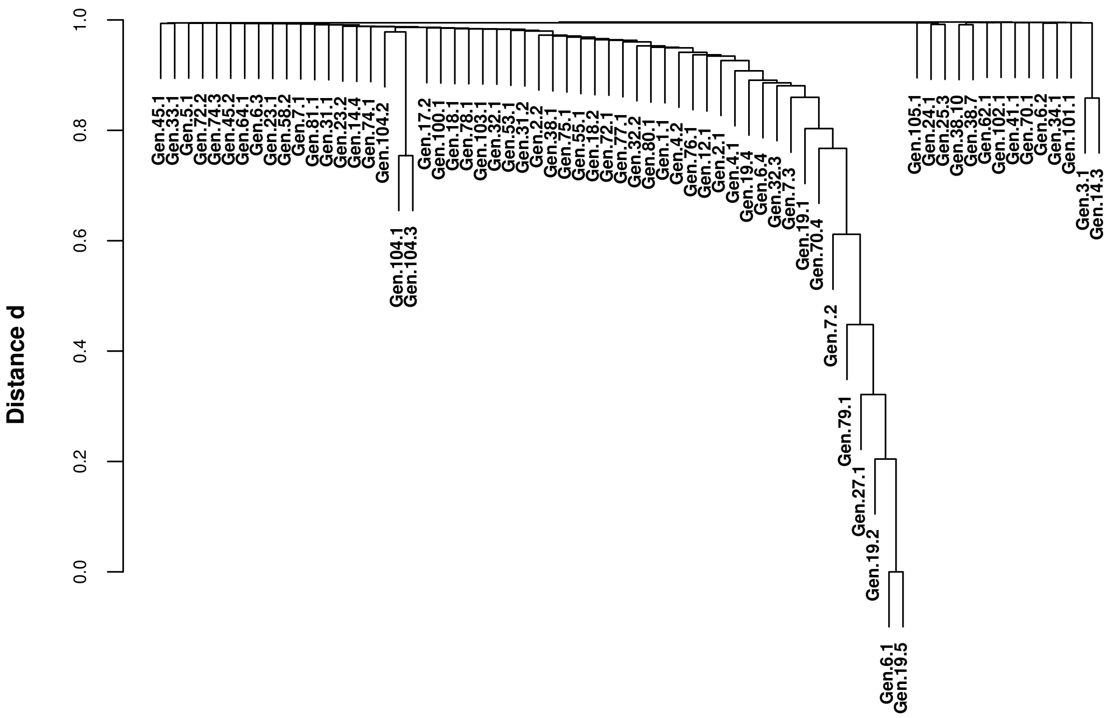


**Figure 2:** Clustering of cases obtained by $S_K$ (Kulczynski)

The dendrogram shows a couple of similar pairs and triples, e.g. Gen.104.1 - Gen.104.3, Gen.2.1 and Gen.2.2 or Gen.3.1, Gen.14.3 and Gen.14.4, which merge into smaller subgroups. The revealed structure among the variables displays lots of SNPs that belong to one gene in one cluster, so the interpretation of several subgroups of dependent SNP loci seems reasonable. Following the algorithm of section 3.5, we chose 12 as the number of classes.

All coefficients from Group 1 (in Table 3 on page 9) show a similar dendrogram structure: Two big clusters are visible, the bigger one consisting of a stepwise structure which means that the algorithm assigns all variables one by one to the same cluster. Thus, interacting subgroups of SNPs cannot be found. Additionally to a similar structure, the order in which the variables join the clusters closely resemble each other for all the coefficients of this group. The most similar pair consists of Gen.6.1 and Gen.19.5. $S_K$ by Kulczynski gives the worst clustering results of this group of measures (see Figure 2). Lots of variables are joint not only to the same cluster but at a similar level of distance. The number of classes chosen for the coefficients of this group lies between 3 and 5.

If the 0-0-matches are excluded completely from the analysis (using measures from Group 2 in Table 3), the results get even worse. Except for some variables that

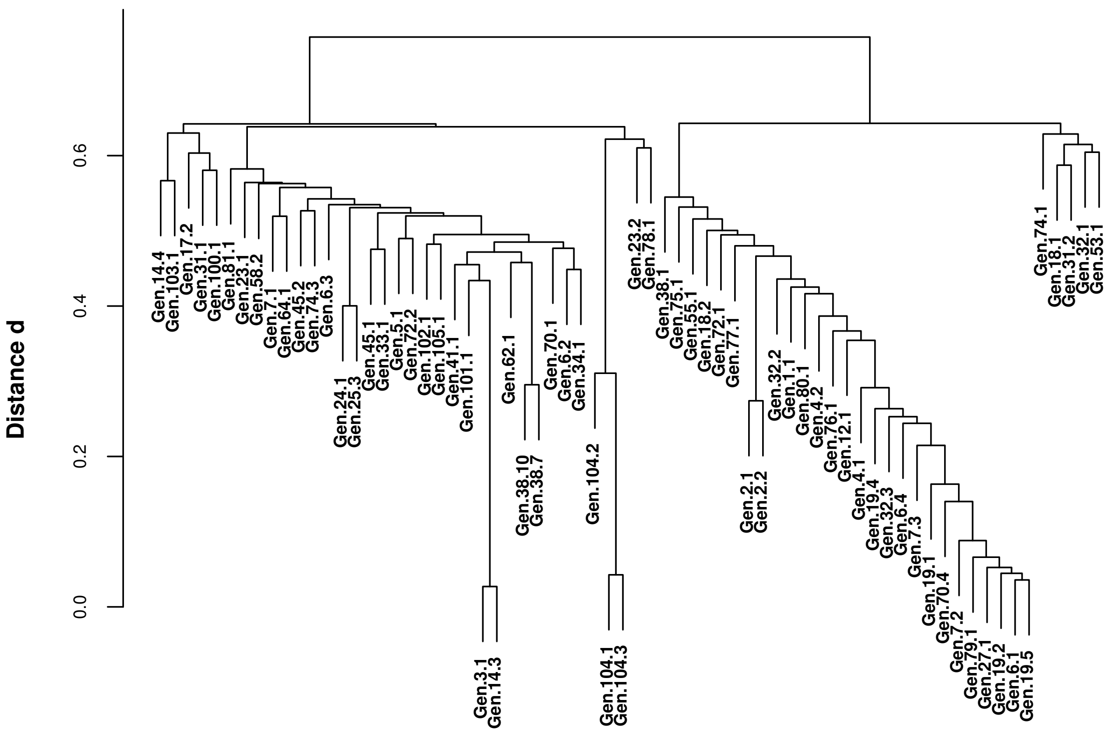


**Figure 3:** Clustering of cases obtained by $S_{Prop}$ (Proportions)

form pairs at first (e.g. Gen.3.1 and Gen.14.3, Gen.104.1 and Gen.104.3 as well as Gen.2.1 and Gen.2.2), all features are fused to one big cluster. The measure by Russell and Rao shows similar characteristics as the measures of Group 2. The number of classes chosen ranges from 7 to 11 exceeding the number of classes for measures of Group 1.

The newly created coefficients perform better, see, e.g., the dendrogram obtained by $S_{Prop}$ (Proportions) in Figure 3. Several subgroups are visible, and even though one cluster contains the same stepwise structure as the previous dendrograms it is restricted to only a part of the data and does not affect all variables. Gen.3.1 and Gen.14.3, Gen.104.1 and Gen.104.3, Gen.2.1 and Gen.2.2 as well as Gen.6.1 and Gen.19.5 can be found among the most similar pairs. Unlike in Figure 2, the range of distance at which a fusion takes place is wide and not similar for most pairs. We chose 5 as the number of classes for $S_{Prop}$. For the other new coefficients, the number of classes ranges from 5 to 10.

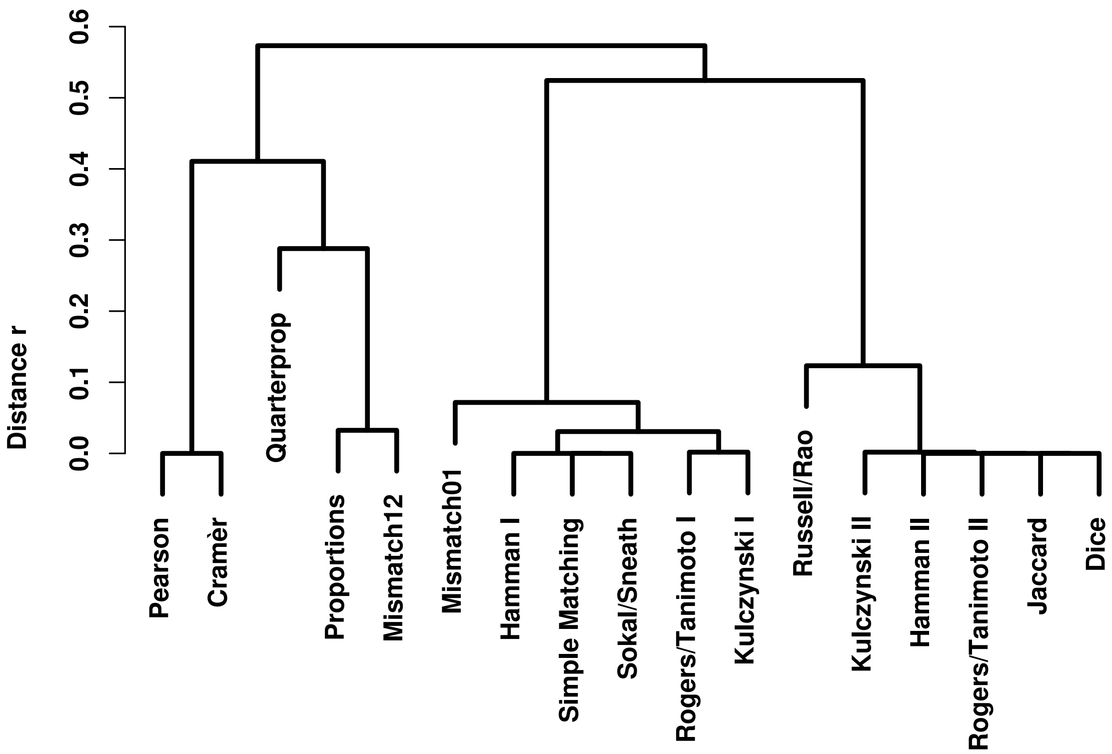


**Figure 4:** Clustering of similarity measures

## 4.2 Comparison

As indicated by the grouping of the different coefficients in Table 3, the results of measures within a group resemble each other quite strongly. This can be shown by applying Rand's method to the clusterings computed above (cf. Figure 4). The results for cases and controls agree.

The choice of the number of classes has a big influence on the result; if two clustering methods give the same partitions, but two different numbers of classes are chosen, Rand's coefficient will not regard them as identical. Therefore, careful consideration is necessary for the final choice, e.g., similar numbers of classes should be taken for outcomes which showed many common characteristics.

Group 2 forms a cluster together with the measure by Russell and Rao, whereas Mismatch01 completes the class built up by Group 1. The rest of the new coefficients forms a separate cluster, which is most similar to the cluster containing the $\chi^2$-coefficients.

These results confirm the conclusions already drawn from construction and performance of the different similarity measures. Besides Mismatch01, the new coefficients do find different structures within the variables. Additionally, they resemble

| Coefficient | Rand's Measure | Coefficient | Rand's Measure |
|---|---|---|---|
| $P_C$ | $R = 0.8429$ | $S_{RR}$ | $R = 0.9491$ |
| C | $R = 0.8481$ | $S_{H00}$ | $R = 0.9491$ |
| $S_{K00}$ | $R = 0.8481$ | $S_{SoSn}$ | $R = 0.9605$ |
| $S_{Prop}$ | $R = 0.9241$ | $S_{SM}$ | $R = 0.9605$ |
| $S_{Mis01}$ | $R = 0.9245$ | $S_H$ | $R = 0.9605$ |
| $S_{RT00}$ | $R = 0.9249$ | $S_{QP}$ | $R = 0.9640$ |
| $S_{Mis12}$ | $R = 0.9320$ | $S_K$ | $R = 0.9886$ |
| $S_J$ | $R = 0.9491$ | $S_{RT}$ | $R = 0.9886$ |
| $S_D$ | $R = 0.9491$ | | |

**Table 5:** Similarity between clusterings of cases and controls

the results of the $\chi^2$-coefficients, which show a good structure as well, more closely than the results of the other matching coefficients.

## 4.3 Differences between Cases and Controls

We use Rand's method not only for comparing the different similarity coefficients, but also for identifying the amount of difference between the case group and the control group. A large deviation (meaning small values of R) indicates different genetic profiles in the two subgroups. The results are given in Table 5. The two $\chi^2$-coefficients find most differences of all measures. $S_{K00}$ seems to show differences as well, but as $S_K$ it gives the poorest clustering result in its group. The other coefficients reveal less differences. Nevertheless, the two new coefficients $S_{Prop}$ and $S_{Mis01}$ perform slightly better than the rest of the other measures. As, additionally, their clustering results show a very reasonable structure, they form a good compromise for both aims.

Two possible reasons why the differences between the cases and controls are rather small might be that either the underlying impact of SNP interactions cannot be found with these methods, or, more likely, many SNPs chosen for the analysis do not help to distinguish between the disease status.

To evaluate our results of section 4, we analyse the simulated data.

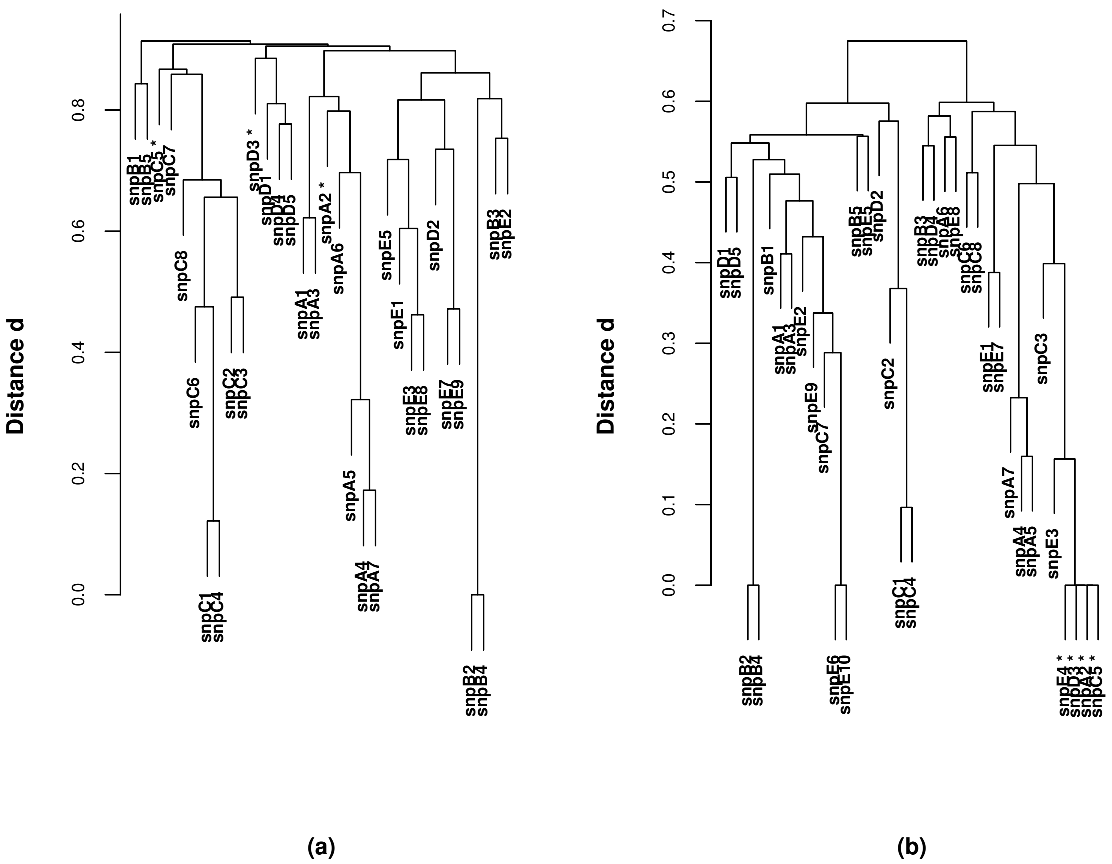


**Figure 5:** Clustering of simulated case data obtained by Pearson's $P_C$ (a) and by $S_{Mis12}$ (Mismatch12) (b)

# 5 Application to the Simulated Data

Cluster analysis of the simulated data should find both the block structure and, in the case group, the four causative SNPs, indicated by an asterisk in Figure 5. By chance, the simulated data set contains two SNPs which are monomorphic for the reference type in the case group (snpE6 and snpE10). Additionally, the recessive causative SNP snpE4 shows the homozygous variant for all cases by definition. These three SNPs are excluded for the analysis based on $\chi^2$-measures in the case group because both coefficients cannot deal with constants. Thus, they cannot be compared to the other coefficients by Rand's method. The dendrograms show that both C an $P_C$ find the block structure best of all coefficients. For the case group, only two SNPs are assigned to clusters containing SNPs from different blocks, the other clusters are homogenous (cf. Figure 5 (a)). Thus, the $\chi^2$-measures are efficient for the detection of linkage disequilibrium. However, C and $P_C$ do not find the four causative SNPs. For searching for interactions which

alter the disease risk they do not seem to be suitable.
The performance of the other coefficients resembles the one of the real data set. All measures find the three dominant causative SNPs, but the new coefficients except Mismatch01 structure the data best. The result of $S_{Mis12}$ is displayed in Figure 5 (b). $S_{Mis12}$ is the only coefficient which groups all four causative SNPs together.
In the comparison by Rand, Group 1 and Group 2 form a cluster each with $S_{RR}$ belonging to Group 2. Proportions and Mismatch12 give similar performances and are joint later on to the cluster of Group 1. Mismatch01 gives a result most different from all other coefficients.
The difference between the case and the control group is not large as only 4 out of 35 SNPs determine the disease status. The values of Rand's similarities for the comparison of both groups lie between 0.75 and 0.87. They are lower than in the real data set. This strengthens the idea that lots of the SNPs chosen for the GENICA data do not contribute to a genetic profile suitable to identify women with a specific breast cancer susceptibility.

# 6 Results and Discussion

The conventional coefficients of Group 1 and, respectively, of Group 2 yield similar results. If the 0-0-matches are treated normally or are left out completely, the weighting of matches and mismatches does not effect the outcome. All these clusterings do not reveal an underlying structure within the real data set. In the simulated data in which a strong structure is present, the measures did find it, but not as distinctly as the new coefficients.
The performance of the two $\chi^2$-measures is similar to each other. They both display a better structure of subgroups of SNPs and they find dependent structures within genes, but in many studies it is questionable if random sampling of cases can be assumed. Their performance on simulated data shows that they can detect blocks of high linkage disequilibrium very well, but fail to find the interactions causing the susceptibility.
The new coefficients of Group 3, on the other hand, show a much better structure for both the real and simulated data. Especially Proportions and Mismatch12 yield good results. They outperform $\mathrm{P}_C$ and C in the detection of causative SNPs in the simulated data set. Furthermore, they do not need to fulfil assumptions as

there is no underlying model implied. The $\chi^2$-coefficients demand preprocessing of the data in order to avoid constant variables. This problem does not affect the new measures.

It is also possible to compare cluster results for cases and controls by the method of Rand. We detected fairly small differences in the real data set. The differences in the simulated data set consist of four out of 35 SNPs and, indeed, Rand's method yields smaller similarity values than in the real data set.

If only SNP data are to be analysed, Proportions and Mismatch12 are sensible choices for similarity measures. If the analysis involves clinical, epidemiological or environmental variables as well, these measures usually cannot be employed anymore since they require comparable categories for all features. Thus, for mixed variables, $\mathrm{P}_C$ and C represent a better choice. Another approach uses mixed similarity measures which differentiate between comparisons (cf. Selinski and Ickstadt (2005)). If SNP variables are compared, the new similarity measures are chosen. If a SNP variable is compared to a different feature, the $\chi^2$-coefficients, for example, can be applied. For the comparison of, say, continuous clinical variables, metrical measure can be employed.

Cluster analysis of SNP data is only a first step on the way of identifying risk factors for complex genetic diseases like cancer. Further methods, e.g., classification methods like CART or Support Vector Machines (Schwender et al., 2004) can use the results of a cluster analysis to concentrate on the meaningful clusters found by the suitable coefficients and to obtain better results by disregarding uninformative variables.

## Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged. The authors thank all partners within the GENICA research network for their cooperation. The GENICA Network are The Interdisciplinary Study Group on Gene Environmental Interactions and Breast Cancer in Germany, represented by C. Justenhoven (Stuttgart), H. Brauch (Stuttgart), S. Rabstein (Bochum), B. Pesch (Bochum), V. Harth (Bonn/Bochum), U. Hamann (Heidelberg), T. Brüning (Bochum), Y. Ko (Bonn).

# References


Anderberg, M. (1973): *Cluster Analysis for Applications*. New York: Academic Press.

Cox, T. and Cox, M. (2001): *Multidimensional Scaling*. London: Chapman and Hall, 2nd edition.

Dice, L. (1945): Measures of the amount of ecologic association between species. *Ecology*, 26, 297–302.

Fahrmeir, L., Hamerle, A., and Tutz, G. (1996): *Multivariate statistische Verfahren*. Berlin: Walter de Gruyter, 6th edition.

Garte, S. (2001): Metabolic susceptibility genes as cancer risk factors: Time for a reassessment? *Cancer Epidemiology, Biomarkers & Prevention*, 10, 1233–1237.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., and Xu, C.-F. (2004): Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*, 12, 395–399.

Nothnagel, M. (2002): Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. *Am J Hum Genet*. 71, (Suppl.)(4): A2363.

Pusch, W., Wurmbach, J.-H., Thiele, H., and Kostrzewa, M. (2002): MALDI-TOF mass spectrometry-based SNP genotyping. *Pharmacogenomics*, 3 (4), 537–548.

Rand, W. (1971): Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.

Sachs, L. (1999): *Angewandte Statistik*. Berlin: Springer, 9th edition.

Schwender, H., Zucknick, M., Ickstadt, K., and Bolt, H. (2004): A pilot study on the application of statistical classification procedures to molecular epidemiological data. *Toxicology Letters*, 151, 291–299.

Selinski, S. and Ickstadt, K. (2005): Similarity Measures for Clustering SNP and Epidemiological Data. *Technical Report*, Statistics Department University of Dortmund, Germany.

Snustad, D. and Simmons, M. (1999): *Principles of Genetics*. New York: Wiley, 2nd edition.

Sokal, R. and Sneath, P. (1963): *Principles of Numerical Taxonomie*. San Francisco: Freeman.