EPISTATIC KINSHIP – A NEW MEASURE FOR THE ASSESSMENT OF GENETIC DIVERSITY IN LIVESTOCK POPULATIONS

Dissertation
for the Doctoral Degree
at the Faculty of Agricultural Sciences,
Georg-August-University Göttingen

presented by
Christine Flury
born in Solothurn (Switzerland)

Göttingen, December 2005

D 7

1. Referee: Prof. Dr. H. Simianer

2. Co-referee: Prof. Dr. G. Thaller

Date of disputation: 2. February, 2006

Acknowledgements

I would like to greatly acknowledge:

Prof. Dr. Henner Simianer for offering me the topic, his great support and for accepting this thesis.

Prof. Dr. George Thaller for taking over the co-reference, motivating feedback and for accepting this thesis.

Prof. Dr. Clemens Wollny acting as third examiner.

The Deutsche Forschungsgemeinschaft (DFG) for the financial support of the project.

Prof. em. Dr. Peter Glodek for the introduction to the history of the Göttingen Minipig and interesting discussions.

Dr. Steffen Weigend and his group, especially Annet Weigend and Anke Flörke for conducting the genotypings, their reliability and friendliness.

Ellegaard Göttingen Minipigs Aps. and its employees for the introduction to their company, providing tissue samples of the two Danish populations and the hospitality of family Ellegaard.

Oskar Lippstreu for his daily engagement for the Göttingen Minipig population in Relliehausen and for his assistance sampling the German population.

Dr. Ralf Fischer, Köllitsch for his support regarding the minipig database.

Frank Bosselmann for sharing the office and not only scientific defeats and highlights, Dr. Sven König for frequent visits to our office, Frederike Köhn for correcting parts of the thesis, Bianca Lind, Janet Schmidtko and Tamina Pinent for the exchange of experiences.

Numerous other colleagues from the Institute of Animal Breeding and Genetics, University of Göttingen for the introduction into German university life.

Dr. Eva Moors for always having a cup of coffee with milk, motivating discussions and dog-sitting several times.

My flatmates Olivia Armbrust, Thomas Conrad, Egbert Griebeling and Yvonne Seidenschwanz for the integration and the comfortable and friendly way we shared the apartment.

Dr. Sabine Reist-Marti for her visits to Göttingen and the nice exchange of experiences.

Sibylle Menet for decisive discussions, long phonecalls and giving a crash course to India.

Leona Jakob for braving the long train trip and her family for always making me welcome in Burgdorf.

My parents Annemarie und Franz Flury for their utmost support, always integrating my plans with theirs and transmitting their fascination of fauna and flora to me.

My brother Stefan Flury for accommodation and his almost always brilliant jokes.

My sister Regula Flury, her partner Boris Leisi and little Timo for their huge encouragement and always having an open door.

And last but not least my friends Martina Dietrich-Meschenmoser, Denise Häfelfinger, Susanne Tschumi and Esther Wyss for several visits to Göttingen, their sympathy for all kind of situations and the long wire back home.

TABLE OF CONTENTS

Zusammenfassung		**8**
Summary		**11**
1st Chapter	**Introduction**	**15**
	Genetic diversity in livestock populations	16
	Assessment of genetic diversity	18
	Kinship coefficient to assess genetic diversity	20
	Scope of the thesis	23
	References	24
2nd Chapter	**Extension of the concept of kinship, relationship, and inbreeding to account for linked epistatic complexes**	**29**
3rd Chapter	**Epistatic kinship a new measure of genetic diversity for short term phylogenetic structures – theoretical investigations**	**53**
4th Chapter	**Epistatic kinship for three subdivided populations of the Goettingen Minipig**	**85**
5th Chapter	**General discussion**	**113**
	Algorithms	114
	Estimation of additive x additive interactions	115
	Assessment of genetic diversity	116
	Analogies with other approaches	124
	References	126

ZUSAMMENFASSUNG

Das Ziel der vorliegenden Dissertation war die Erweiterung der Betrachtungseinheit des Abstammungskoeffizienten von einzelnen Loci auf Chromosomensegmente der Länge x in Morgan. Das neue Maß mit der Bezeichnung epistatische Kinship beschreibt die Wahrscheinlichkeit, dass zwei zufällig gezogene Chromosomensegmente der Länge x in Morgan herkunftsgleich sind. In Anlehnung an Eding und Meuwissens Verwendung des Abstammungskoeffizienten, wurde die epistatische Kinship als neues Maß für genetische Diversität bei landwirtschaftlichen Nutztieren vorgeschlagen.

Im Rahmen der Arbeit wurden Algorithmen für die epistatische Kinship, für den epistatischen Verwandtschaftskoeffizienten und für den epistatischen Inzuchtkoeffizienten hergeleitet. Zusätzlich wurden die Regeln der Tabellenmethode zum direkten Erstellen der Verwandtschaftsmatrix und deren Inverse erweitert. Alle Algorithmen enthalten die Grösse e^{-x} und somit ist die Einzellocusbetrachtung ($x=0$) ein Spezialfall des erweiterten Ansatzes.

In einer Simulationsstudie wurde der Einfluss der Segmentlänge, der Anzahl gezogener Tiere und der Anzahl typisierter Segmente unter Verwendung von Abstammungsinformation untersucht. Die Untersuchung zeigte, dass für verschiedene Generationen nach der Trennung der Populationen verschiedene Segmentlängen den höchsten Informationsgehalt hatten. Im Weiteren wurde ein linearer Effekt der Anzahl typisierter Segmente und ein quadratischer Effekt der Anzahl getesteter Tiere auf die Genauigkeit der epistatischen Kinship als Maß für genetische Diversität gefunden.

Für kleine Rassen und Vergleiche zwischen Rassen ist die Abstammungsinformation oft unvollständig, weshalb die markergestützte Schätzung der epistatischen Kinship vorgeschlagen und in einer zusätzlichen Simulationsstudie theoretisch untersucht wurde. Die Resultate unter der Annahme von bekannten Haplotypen bestätigten das hohe Potential der epistatischen Kinship zur Bestimmung der genetischen Diversität bei kurzen Entwicklungszeiträumen. Weiter zeigte diese Studie, dass die Genauigkeit der markergestützten epistatischen Kinship neben den oben genannten Faktoren

Segmentlänge, Anzahl getesteter Tiere und Anzahl typisierter Segmente auch von der Anzahl Allele pro Locus beeinflusst wird.

Abschließend wurde die markergestützte epistatische Kinship in einer praktischen Anwendung evaluiert. Dazu wurden in drei Unterpopulationen des Göttinger Minischweins Gewebeproben gesammelt. Insgesamt wurde DNA von 167 Vollgeschwisterpaaren für 6 Segmente mit 33 Mikrosatelliten typisiert. Basierend auf der genetischen Karte USDA_MARC_v2 war die durchschnittliche Segmentlänge für die sechs Segmente 0,0665 Morgan. Die Erwartungswerte wurden unter Verwendung des gesamten Pedigrees (2081 Tiere) für die 167 Vollgeschwisterpaare mit den eingangs erwähnten Algorithmen ermittelt.

Für die markergestützte Schätzung der epistatischen Kinship sind Haplotypen relevant. Deshalb wurde eine erweiterte Version des EM-Algorithmus, bei welcher die vollständige Vollgeschwisterinformation berücksichtigt wird, zur Rekonstruktion der Haplotypen verwendet. Alle Marker wurden auf Hardy-Weinberg-Gleichgewicht (HWG) getestet, weil Abweichungen davon bei der Anwendung des EM-Algorithmus zu verzerrten Schätzungen der Haplotypfrequenzen führen könnten. Die Vernachlässigung der Marker, die vom HWG abweichen, hatte allerdings einen beträchtlichen Informationsverlust zur Folge. Deshalb wurden alle Marker, unabhängig vom Ergebnis des HWG-Tests, für die weitere Analyse verwendet.

Die markergestützte epistatische Kinship wurde für die sechs Segmente einzeln zwischen und innerhalb Populationen berechnet. Die Resultate für die einzelnen Segmente variierten. Dennoch war der erwartete Trend zunehmender epistatischer Kinship mit abnehmender Segmentlänge erkennbar. Im Vergleich mit dem Erwartungswert für die durchschnittliche Segmentlänge von 0,0665 Morgan, war die durchschnittliche markergestützte epistatische Kinship höher.

Unter der Annahme, dass alle identischen Haplotypen auch herkunftsgleich sind, wird erwartet, dass der Intercept der Regression von markergestützter epistatischer Kinship auf die pedigreebasierte epistatische Kinship durch Null geht. Basierend darauf wurden Korrekturfaktoren für identische Haplotypen, die jedoch nicht herkunftsgleich sind, ermittelt und angewandt. Die Variabilität der markergestützten epistatischen Kinship zwischen den einzelnen Segmenten wurde unter Anwendung der Korrekturfaktoren geringer.

Zur Beschreibung der genetischen Distanzen wurde ein Distanzmaß hergeleitet. Dieses Maß zeigt einen approximativ linearen Verlauf mit der Anzahl Generationen seit der Auftrennung der Populationen. Die Reihenfolge der Distanzen war für die pedigreebasierten Erwartungswerte gleich wie für die markergestützten Schätzungen. Jedoch waren die Standardfehler für die markergestützten epistatischen Kinship Distanzen hoch.

Verschiedene Gründe für die hohen Standardfehler der markergestützten epistatischen Kinship und den zugehörigen Distanzen wurden diskutiert. Im Vergleich mit den theoretischen Untersuchungen bestätigte die praktische Anwendung das Potential der epistatischen Kinship als Maß für genetische Diversität. Zusätzlich konnten neue Aspekte aufgezeigt werden. Die Korrektur für statusgleiche, jedoch nicht herkunftsgleiche Haplotypen erwies sich als weniger relevant als bei der Einzellocusbetrachtung, dennoch wird dazu angeraten.

Das vorgeschlagene Diversitätmaß ist das erste, welches speziell zur Berücksichtigung von kurzen Differenzierungszeiträumen entwickelt wurde. Dabei werden nicht Drift und Mutation, sondern Rekombination als Hauptgröße zur Entstehung von Unterschieden zwischen Populationen herangezogen. Es wird erwartet, dass dieser Ansatz zu einem besseren Verständnis der genetischen Diversität für kurze Entstehungszeiträume, wie sie bei landwirtschaftlichen Nutztierpopulationen oft gegeben sind, führt.

SUMMARY

The main goal of this thesis was the extension of the single locus concept of the kinship coefficent to chromosomal segments of length x in Morgan. This metric – called epistatic kinship – describes the probability that two randomly drawn segments of length x in Morgan are identical by descent. In analogy to Eding and Meuwissen's application of the kinship coefficient, the epistatic kinship is proposed as a new measure for the assessment of genetic diversity.

Algorithms for the epistatic kinship, the epistatic relationship and the epistatic inbreeding coefficient were derived for a given pedigree. Furthermore the rules to set up the numerator relationship matrix and its inverse were extended for segments of a predefined length in Morgan. The term e^{-x} occurs in all of the proposed algorithms, therefore the single locus consideration i.e. $x=0$ becomes a special case of the extended apporach.

In a simulation study the respective influences of the segment length, of the number of animals sampled and of the number of segments typed on the epistatic kinship for a given pedigree list are examined. One result was, that different generations after fission different segment lengths were most informative. Further it was observed, that the number of segments typed has a linear impact and the number of animals sampled has a squared influence on the resolution of the method.

For a situation without pedigree information, marker based epistatic kinship was investigated in an additional simulation study. The results for the marker estimated epistatic kinship assuming known haplotypes underlined the high potential of the epistatic kinship for short term phylogenies. In addition to the three parameters mentioned above, i.e. the segment length x, the number of animals sampled and the number of segments sampled, the number of alleles per locus was found to influence the accuracy of marker estimated kinship.

Finally the use of marker estimated epistatic kinship was evaluated in a practical application. For this purpose tissue samples were taken in three subdivided populations of the Goettingen minipig. In total 167 fullsibpairs were sampled and genotyped for 6 segments (33 microsatellites). The average segment length for the 6 segments was 0,0665 Morgan based on the genetic map USDA_MARC_v2. The expected values were calculated for the total pedigree (2081 animals) of the sampled animals based on the proposed algorithms.

For the marker based estimation of epistatic kinship haplotypes are requested. Therefore an extended version of the EM-algorithm was applied to fully account for the fullsib information. Hardy-Weinberg-Equilibrium (HWE) testing for all markers was conducted, as the use of markers deviating HWE might lead to biased haplotype frequency estimates applying the EM-algorithm. The negligence of the markers deviating from HWE results in a high information loss. Therefore all initially available markers were kept for further analysis, regardless of being in HWE or not.

The marker estimated epistatic kinship was presented for the six segments within and between populations. The results for the single segments are variable. However, the expected trend of increasing epistatic kinship with decreasing segment length was confirmed. In comparison with the expected value at the average segment length 0,065 Morgan, the average of the marker estimated epistatic kinship for the six segments was on a higher level.

Assuming that all identical haplotypes found are due to identity by descent, the expectation of the intercept from the regression of marker estimated epistatic kinship on pedigree based epistatic kinship would be zero. Based on this assumption a correction factor for identical haplotypes which are not identical by descent is proposed. The variability between the corrected marker estimated epistatic kinship for the six segments decreased, when this correction was applied.

A genetic distance measure for the epistatic kinship and the marker estimated epistatic kinship was derived, which is linear with the number of generations since fission. The genetic distances for the three subpopulations of the Goettingen minipig resulted in the same order for the pedigree based expectations and the marker based epistatic kinship distances. However, standard errors for the latter were found at a remarkable level.

Different reasons for the high standard errors of the marker based epistatic kinship and the corresponding epistatic kinship distances are presented and discussed. The practical application confirmed the potential of the epistatic kinship as diversity measure found in the theoretical investigations and highlighted some additional points. The correction factor for identical haplotypes which are not identical by descent was found to be necessary – but much less important than the correction factor for identical alleles which are not identical by descent for the single locus consideration.

The suggested diversity is the first such measure which was designed for the very purpose of studying short term phylogenies, and which is not using genetic drift and mutation, but recombination as the major force creating population differences. Thus it is expected that the method proposed here has a considerable potential to develop a better understanding of short-term phylogenetic structures in farm animal populations.

1st CHAPTER

INTRODUCTION

INTRODUCTION

Genetic diversity in livestock populations

Genetic diversity is required for populations to cope with future changes. Considering genetic diversity in agricultural populations not only the capacity to evolve with changing production environment (e.g. global warming, changes in disease pressure) but also the capacity to cope with changing market requirements (e.g. other composition fatty acids in animal products) is of high relevance (Simianer, 2005a). Thus genetic diversity is seen as an insurance against future changes (Smith, 1984).

In livestock populations genetic diversity is expressed on the phenotypic level as variability in production traits, exterior traits, reproduction traits, health traits, and other characters. In comparison with natural populations a wide phenotypic diversity is observed within and between livestock populations (Andersson, 2001; Notter, 1999). These phenotypic differences are the result of genetic diversity and environmental differences (Oldenbroek, 1999). Genetic diversity can be assessed between species, breeds, specific lines and within those groups. A breed is defined by the Food and Agricultural Organization of the United Nations (FAO, 1998) as a group of animals which belong to the same population based on certain characteristics.

None of the about 30 livestock species is threatened with extension. But more than a third of the about 6400 documented livestock breeds are under risk of extinction and up to two percent of the breeds go extinct every year (Scherf, 2000). Thus one to two breeds are lost per week. It is estimated that 20% - 50% of the total genetic variation within a species exists between breeds (Hall, 2004; Oldenbroek, 1999). This leads to the assumption that the loss of breeds highly influences the variability within species. However, the small population size of a population at risk causes accelerated erosion of the genetic diversity within this population (Eding et al., 2002). Hence, in terms of total

diversity within species the loss of a highly inbred population is supposed to have a smaller influence.

In the second half of the 20th century the industrialised agricultural production led to a high promotion and selection of some specific breeds (Gandini and Villa, 2003). Today within the commercially most important species (cattle, pig and poultry) about six breeds are globally competitive. Based on biotechnological progress (artificial insemination, embryotransfer, kryoconservation) the global exchange and trade of breeding stock and genetic material became possible. In dairy cattle, sons of limited number of sires and grand sires of the Holstein breed dominate global lists of active sires for artificial insemination (Notter, 1999). Another extreme is the actual market of broilers, layer hens and turkeys, which is dominated by at most 10 multinational breeding companies (Notter, 1999). Breeding companies concentrate their activities on globally tradeable and economically rewarding breeds. The high developing costs are covered with a high market share. Tisdell (2003) concluded, that the extension of markets and economical developments led to a shift from locally adapted multipurpose breeds to highly specialised, global breeds.

This tendency was recognised in the early eighties by some non governmental organisations (e.g. Pro Specie Rara, Switzerland, founded in 1984) who became active in monitoring and supporting local plant and animal genetic resources. On the international level the issue of the ‚Convention on Biological Diversity' (CBD, 1992) of the United Nations stimulated the public awareness for farm animal genetic variation. More than 180 nations ratified this convention which binds the signing countries to develop national strategies, plans or programs for the conservation and the sustainable use of biological diversity'. The national activities are coordinated and monitored by the FAO in Rome.

On the national and international level resources for conservation activities are limited. Therefore not all breeds can be given the same priority for conservation. This means

that potential conservation activities rely on a decision process. The key question is which breeds should be chosen to assure the highest genetic diversity within species for the future. The maximisation of conserved diversity within species is a function of within and between breed diversity (Simianer, 2005a). Seven criteria that can be considered to choose specific breeds for conservation are described by Ruane (1999). The presented thesis deals with the criteria of genetic uniqueness and therefore the further sections concentrate on this.

Assessment of genetic diversity

The genetic composition of a population is usually described in terms of allele frequencies, number of alleles and heterozygosity (Frankham et al., 2002). A wide range of studies for the assessment of genetic diversity in livestock breeds were conducted using genetic distances. For genetic distances the genetic differences between populations are assessed based on differences between allele frequencies at several loci. The wide use of genetic distances is explained with the intuitive appeal of being objective (Ruane, 1999). Additionally the improvement and decreasing costs of DNA-based techniques improved the resolution of genetic distances studies due to their higher per locus heterozygosity (Barker, 1999). Genetic distances based on microsatellites assume an evolutionary timespan since population fission and that no migration occurred between subpopulations. These presumptions often do not hold considering breeds of livestock species.

Initially genetic distances were developed for the description of the differentiation of species. Livestock breeds are domesticated and improved by man, the divergence period between breeds is short from an evolutionary perspective (Nagamine and Higuchi, 2001). Most of the European breeds go back to the 19th or even the beginning of the 20th century (Sambraus, 2001). Those breeds of recent origin were also important for breed development in the New World (Ruane, 1999). Therefore the assumption of an evolutionary time span does not hold for breed specification and the role of mutation of

marker genes in creating genetic differences between breeds is assumed to be small (Nagamine and Higuchi, 2001).

Migration is by definition ignored in the models for genetic distances (Oldenbroek, 1999). However, crossbreeding was commonly practised in livestock 50 - 100 generations ago (Visscher, 2003) and is still a widely used breeding strategy. Thus admixture can not be neglected for livestock breeds. Further the construction of phylogenetic trees for visualisation of genetic distance results based on such data contradicts the principles of phylogenetic reconstruction (Toro and Caballero, 2004).

Weitzman (1992) suggested a concept for decision making in conservation that uses genetic and non genetic information. The current diversity and the expected change in diversity over a certain time horizon is calculated for a set of populations. This approach was applied on livestock breeds by Reist-Marti et al. (2003) and Thaon d'Arnoldi et al. (1998).

Genetic distances describe between population diversity. Eding et al. (2001) argued that considering between population diversity only, highly inbred population tend to have an increased genetic distance to other breeds and are therefore favoured for conservation decisions. The ignoring of the within population diversity is also a widely criticised aspect (Caballero and Toro, 2002; Eding, 2002; Laval et al., 2002) applying the Weitzman approach. However, this negative correlation of the diversity between and within breeds was not confirmed by Pinent et al. (2005) who applied the Weitzman approach on German chicken breeds. Further the use of the expected number of conserved alleles was proposed as diversity metric for the Weitzman method to consider within and between population diversity simultaneously (Simianer, 2005b). Nevertheless, to secure a sustainable conservation of breeds, within population diversity is important to retain the capacity to respond to selection and to protect animals and populations from the adverse effects of inbreeding and random drift.

Kinship coefficient to assess genetic diversity

To overcome the limitations of widely used methods for the assessment of genetic uniqueness, i.e. the assumption of an evolutionary timespan for genetic distances and the ignoring of within breed variability in the Weitzman approach, Eding and Meuwissen (2001) proposed the use of the kinship coefficient for the assessment of genetic diversity.

All measures of relatedness are based on the concept of identity by descent (Lynch and Walsh, 1998). Alleles that are identical by descent are direct descendants of a specific allele in a common ancestor. The kinship coefficient K_{st} describes the probability, that two randomly chosen alleles from the same locus of individuals s and t are identical by descent (Malécot, 1948). The average kinship coefficient is valid for the entire genome and not only for the loci under investigation. The minimisation of the mean kinship coefficient in a set of individuals is supposed to minimise duplicates of alleles descending from the same ancestor (Eding, 2002).

There is an analogy of the kinship coefficient with other important measures of relatedness: The inbreeding coefficient (Wright, 1922) describes the probability that two alleles at one locus in an individual are identical by descent. Thus it is equivalent to the coefficient of kinship of the parents. Another well known measure of relatedness in animal breeding is the relationship coefficient (Wright, 1922), the analogy between the kinship coefficient and the relationship coefficient R_{st} is $R_{st} = 2K_{st}$. Emik and Terrill (1949) proposed a tabular method for the direct set up of the numerator relationship matrix (NRM) for a given pedigree. The well-known rules to set up the inverse of the NRM were first suggested by Henderson (1976) and Quaas (1976). Based on those findings the derivation of the kinship coefficient K_{st} is straightforward if pedigree information is available.

Pedigree information is often missing under poor administration and documentation (which often is the case for local, endangered breeds) or in between breed analysis. Under such circumstances pedigree based kinship coefficients can not be used as measure to assess genetic diversity. To overcome this limitation Eding and Meuwissen (2001) investigated the use of marker estimated kinship coefficients based on similarities of marker alleles. They showed that unbiased estimation of kinship from marker data highly depends on the correction for the probability of alleles being identical by state but not identical by descent. Thus an appropriate estimation of allele frequencies in the founder generation is crucial. Further the authors suggested a core set method for conservation decisions (Eding et al., 2002). The relative contribution of each population to the core set is calculated in such a way, that the average marker estimated kinship is minimised. In an additional publication (Eding and Meuwissen, 2003) the simultaneous estimation of marker estimated kinship and the probabilities of alleles being identical by state was investigated to overcome the problem of negative contributions of breeds and to minimise the errors of marker estimated kinship (Bennewitz and Meuwissen, 2005).

The kinship coefficient and its marker based estimators have some intuitive properties as tool for the assessment of genetic diversity in livestock populations:

- When applying kinship coefficients drift and selection are the only forces generating differences between populations, thus the short developing time for livestock breeds is better accounted for.
- Kinship coefficients can be estimated within and between populations. The consideration of the within population diversity is important for conserving viable populations for the future.
- Kinship coefficients are involved in the variance of quantitative traits, thus the minimisation of the kinship coefficient will lead to conservation of variance of quantitative traits.

However, some aspects remain open. The question arises if the kinship coefficient is powerful enough for the assessment of genetic diversity in short term phylogenies (i.e. 10 – 20 generations since fission). Short developing periods might be of interest where cross breeding was applied 10 - 20 generations ago (e.q. Fleckvieh) or for recently created breeds with a laboratory use in mind (Goettingen Minipig).

Applying marker estimated kinship coefficients a high fraction of identical alleles is due to identity by state. Thus a correction factor is essential. So far, no general applicable rules for the derivation of such a correction factor are given.

Further conserved genomic regions spanning over several cM are reported for different livestock species (Farnir et al., 2000; McRae et al., 2002; Nsengimana et al., 2004; Tenesa et al., 2003). Population bottlenecks can force the creation of so called linkage disequilibrium (LD) (Visscher, 2003). Thus the remaining fraction of conserved haplotypes between populations might be used for the quantification of the number of generations since fission and for the assessment of genetic differences between populations.

Scope of the thesis

The major scope of this thesis was the extension of the single locus consideration of the kinship coefficient to chromosomal segments of length x in Morgan. This measure called epistatic kinship, describes the probability that two chromosomal segments of a predefined length between two individuals are identical by descent. For the segment based epistatic kinship the probability of recombination events is crucial. Thus the epistatic kinship is supposed to lead to a higher resolution for the assessment of genetic diversity assuming short term phylogenies which are given for livestock populations or laboratory populations. In particular this thesis includes:

i) the derivation of algorithms for the calculation of epistatic kinship, epistatic relationship and epistatic inbreeding,
ii) the extension of the rules to set up the epistatic numerator relationship matrix and its inverse directly from a pedigree list,
iii) a simulation study on epistatic effects of linked loci,
iv) theoretical investigations of the marker estimated epistatic kinship as a new measure for diversity studies,
v) the evaluation of the new measure in a practical application to three subdivided populations of the Goettingen Minipig.

The first three issues are presented in chapter 2. In chapter 3 the first issue is given again as introduction to the main part of the chapter which covers the fourth issue. The following chapter 4 contains the fifth issue. The general discussion is held in chapter 5.

References

Andersson, L. 2001. Genetic dissection of phenotypic diversity in farm animals. Nature Genetics 2: 130-138.

Barker, J. S. F. 1999. Conservation of livestock breed diversity. AGRI 1999 25: 33-43.

Bennewitz, J., and Meuwissen, T. 2005. A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance. Genetics Selection Evolution 37: 315-337.

Caballero, A., and Toro, M. 2002. Analysis of genetic diversity for the management of conserved subdivided populations. Conservation Genetics 3: 289-299.

CBD. 1992. Convention on Biological Diversity. Secretariat of the Convention on Biological Biodiversity, St. Jacques Street, H2Y 1N9, Montreal, Canada.

Eding, H., and Meuwissen, T. H. E. 2001. Marker based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118: 141-159.

Eding, J. H. 2002. Conservation of genetic resources. Assessing genetic variation using marker estimated kinships, Wageningen Agricultural University, Wageningen.

Eding, J. H., Crooijmans, R. P. M. A., Groenen, M. A. M., and Meuwissen, T. H. E. 2002. Assessing the contribution of breeds to genetic diversity in conservation schemes. Genetics Selection Evolution 34: 613-633.

Eding, J. H., and Meuwissen, T. H. E. 2003. Linear methods to estimate kinships from genetic marker data for the construction of core sets in genetic conservation schemes. Journal of Animal Breeding and Genetics 120: 289-302.

Emik, L. O., and Terrill, C. R. 1949. Systematic procedures for calculating inbreeding coefficients. Journal of Heredity 40.

FAO. 1998. Primary Guidelines for Development of National Farm Animal Genetic Resources Management Plans, Rome.

Farnir, F., Coppieters, W., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., and Georges, M. 2000. Extensive Genome-wide Linkage Disequilibrium in Cattle. Genome Research 10: 220-227.

Frankham, R., Ballou, J., and Briscoe, D. A. 2002. Introduction to Conservation Genetics. Cambridge University Press.

Gandini, G. C., and Villa, E. 2003. Analysis of the cultural value of local livestock breeds: a methodology. Journal of Animal Breeding and Genetics 120: 1-11.

Hall, J. G. 2004. Livestock biodiversity: genetic resources for the farming of the future. Blackwell Science Ltd.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. Biometrics 32: 69-83.

Laval, G., SanCristobal, M., and Chevalet, C. 2002. Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genetics Selection Evolution 34: 481-507.

Lynch, M., and Walsh, B. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie., Paris.

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Crawford, A. M., and Slate, J. 2002. Linkage Disequilibrium in Domestic Sheep. Genetics 160: 1113-1122.

Nagamine, Y., and Higuchi, M. 2001. Genetic distance and classification of domestic animals using genetic markers. Journal of Animal Breeding and Genetics 118: 101-109.

Notter, D. R. 1999. The importance of genetic diversity in livestock populations of the future. Journal of Animal Science 77: 61-69.

Nsengimana, J., Baret, P., Haley, C. S., and Visscher, P. M. 2004. Linkage Disequilibrium in the Domesticated Pig. Genetics 166: 1395-1404.

Oldenbroek, J. K. 1999. Genebanks and the conservation of farm animals genetic resources. DLO Institute for Animal Science and Health, Lelystad.

Pinent, T., Weigend, S., Tietze, M., and Simianer, H. 2005. Biodiversität zwischen und innerhalb Hühnerpopulationen. In: Vortragstagung der DGfZ / GfT, Berlin

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32: 949.

Reist-Marti, S. B., Simianer, H., Gibson, J., Hanotte, O., and Rege, J. E. O. 2003. Weitzman's Approach and Conservation of Breed Diversity: an Application to African Cattle Breeds. Conservation Biology 17: 1299-1311.

Ruane, J. 1999. A critical review of the value of genetic distance studies in conservation of animal genetic resources. Journal of Animal Breeding and Genetics 116: 317-323.

Sambraus, H. H. 2001. Farbatlas der Nutztierrassen. 6 ed. Verlag Eugen Ulmer, Stuttgart.

Scherf, B. D. (Editor), 2000. World watch list for domestic animal diversity. FAO, Rome.

Simianer, H. 2005a. Decision making in livestock conservation. Ecological Economics 53: 559-572.

Simianer, H. 2005b. Using expected allele number as objective function to design between and within breed conservation of farm animal biodiversity. Journal of Animal Breeding and Genetics 122: 177-187.

Smith, C. 1984. Genetics aspects of conservation in farm livestock. Livestock Production Science 11: 37-48.

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L., and Visscher, P. M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. Journal of Animal Science 81: 617-623.

Thaon d'Arnoldi, C., Foulley, J.-L., and Ollivier, L. 1998. An overview of the Weitzman approach to diversity. Genetics Selection Evolution 30: 146-161.

Tisdell, C. 2003. Socioeconomic causes of loss of animal genetic diversity: analysis and assessment. Ecological Economics 45: 365-376.

Toro, M., and Caballero, A. 2004. Characterisation and conservation of genetic diversity between breeds. In: 55th EAAP Annual Meeting, Bled, Slovenia

Visscher, P. M. 2003. Principles of QTL mapping, manual PhD - course Salzburg, Edinburgh.

Weitzman, M. L. 1992. On diversity. Quarterly Journal of Economics 107: 363-405.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56: 330-339.

2nd CHAPTER

EXTENSION OF THE CONCEPT OF KINSHIP, RELATIONSHIP, AND INBREEDING TO ACCOUNT FOR LINKED EPISTATIC COMPLEXES

Christine Flury, Helge Täubert and Henner Simianer

Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen,
Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

Livestock Science (in press)

EXTENSION OF THE CONCEPT OF KINSHIP, RELATIONSHIP, AND INBREEDING TO ACCOUNT FOR LINKED EPISTATIC COMPLEXES

Christine Flury, Helge Täubert and Henner Simianer

Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

Abstract

Although epistatic effects are well defined and, in principle, can be exploited in quantitative-genetic selection theory, they often are ignored or even treated as nuisance parameters in practical applications. Traditionally, epistasis is considered as an interaction between genes at unspecified loci. Inspired by the observation that functional genes are often organised in physical clusters, we developed a model to combine additive effects and additive x additive interactions in linked gene clusters of defined length. Malécot's kinship concept is extended to identity by descent probabilities for chromosome segments of a given length in Morgan units, called epistatic kinship. Using the analogy of Malécot's kinship and Wright's relationship and inbreeding coefficients, epistatic relationship coefficients and epistatic inbreeding coefficients are defined. Simple rules are given to set up the epistatic numerator relationship matrix and its inverse directly from a pedigree list. The well-known single locus parameters and algorithms to set up the additive numerator relationship matrix and its inverse are a special case of the suggested methodology for a chromosome segment length of null Morgan. A proof of concept of the suggested method is given with a small simulation study. Assuming additive, linked epistatic and residual variance components, 100 replicated data sets for 1000 individuals are generated. From these data, residual maximum likelihood estimates of the variance components and of the chromosome segment size are obtained. Potential applications of the methodology are

discussed. Given that a substantial variance component is attributed to this effect, the expected genetic gain can be increased on the short term if selection is on additive and epistatic effects, the latter comprising additive x additive interaction effect of loci in linkage disequilibrium. This extra benefit, however, will diminish through crossing over in subsequent generations. Despite some practical problems yet to be solved, the suggested model and algorithms open new perspectives to use a higher proportion of genetic variability in selection and breeding.

Keywords: epistatic complexes, kinship, relationship, inbreeding

Introduction

Animal breeding traditionally focuses on the improvement of the additive genetic component. Assuming the infinitesimal model (Fisher, 1918), breeding values basically result from a summation of additive effects at discrete, albeit numerous loci. Nevertheless it is suggested by theory and clearly supported by empirical evidence, that both intra-locus interactions (called dominance) and inter-locus interactions (called epistasis) play a fundamental role in the inheritance of traits. Because such interactions are not fully inherited from parent to progeny (Falconer and Mackay, 1996), those factors and the corresponding variance components usually are ignored or even considered as nuisance parameters in animal breeding.

Nevertheless, efforts were made to estimate non-additive genetic components and to predict individual non-additive breeding values (e.g. Du and Hoeschele, 2000; Fuerst and Soelkner, 1994; Hoeschele, 1991; Tempelman and Burnside, 1990; Van Raden and Hoeschele, 1991). In all these approaches, it was attempted to estimate the total dominance or various types of epistatic variances, like additive x additive, additive x dominance, or dominance x dominance etc. Under the infinitesimal model this means, that the respective effects over all loci or pairs of loci are summed to form the respective non-additive values and the corresponding variance components. Mixed model based residual maximum likelihood (REML, Patterson and Thompson, 1971) methodology

was used to estimate these variance components, applying the appropriate non-additive relationship matrices and, where possible, using algorithms to directly set up the inverses, which however is only possible for some components, like the additive (Henderson, 1976), dominance (Hoeschele and VanRaden, 1991), and additive x additive epistatic (Van Raden and Hoeschele, 1991) component. Extensions for the case of inbreeding are available, as e.g. suggested by Quaas (1976) for the inverse additive genetic numerator relationship matrix, however complications arise when inbreeding and dominance are considered (De Boer and Hoeschele, 1993). Du and Hoeschele (2000) have suggested a Gibbs sampler to estimate two-locus based interaction from a finite locus model, avoiding some of the problems encountered under the infinitesimal model. It is straightforward to implement such non-additive genetic components in breeding value estimation using standard mixed model methodology (Henderson, 1973).

In all these studies, epistasis is considered without accounting for the genetic distance of the interacting loci. However, molecular genetic and, increasingly, bioinformatics research has revealed that interacting genes are often organised in physically linked gene clusters, as e.g. the chicken beta-globulin gene cluster (Mason et al., 1995). Given that haplotypes of these clusters tend to be inherited in non-recombined form, some components of the epistatic complex, especially the additive x additive term will be inherited almost in the same form as the additive component. This also means, that selection can to some extent capitalise on this component, allowing additional genetic progress.

This will be demonstrated with a simple example:

Consider two biallelic loci with alleles A, a at the first and B, b at the second locus. The additive effects are α_A, α_a, α_B, and α_b. We assume, that only the additive x additive interactions of the alleles A and B, denoted ε_{AxB}, and of the alleles a and b, denoted ε_{axb}, have a nonzero effect.

A double heterozygous animal with genotype AaBb will have a total genotypic effect

$$G = \alpha_A + \alpha_a + \alpha_B + \alpha_b + \varepsilon_{AxB} + \varepsilon_{axb}$$

If the two loci are unlinked, the animal will produce with equal probability gametes AB, Ab, aB, and ab with the total gametic effects $\alpha_A + \alpha_B + \varepsilon_{AxB}$, $\alpha_A + \alpha_b$, $\alpha_a + \alpha_B$, and $\alpha_a + \alpha_b + \varepsilon_{axb}$, respectively. This means, that the additive x additive epistatic effect is only inherited in 50 per cent of the cases.

If, however, the two loci are linked with recombination rate $0 \le \theta < 0.5$ and, say, the phase is AB‖ab, gametes AB with effect $\alpha_A + \alpha_B + \varepsilon_{AxB}$ and ab with effect $\alpha_a + \alpha_b + \varepsilon_{axb}$ are produced with probability $0.5(1-\theta)$ while the two recombined gametes Ab and aB are produced with probability 0.5θ, respectively. Thus, the additive x additive epistatic component is inherited with a much higher frequency as in the unlinked case. If linkage is very tight ($\theta \approx 0$), inheritance is very similar to a combined 'quasi-gene' with four alleles and combined effects $\alpha_{AB} = \alpha_A + \alpha_B + \varepsilon_{AxB}$, $\alpha_{Ab} = \alpha_A + \alpha_b$, $\alpha_{aB} = \alpha_a + \alpha_B$, and $\alpha_{ab} = \alpha_a + \alpha_b + \varepsilon_{axb}$, respectively. In this case, additive effects are augmented by the epistatic effects, leading to a larger genetic variance that can be used in selection.

In this contribution we will propose a model that takes additive x additive effects within gene clusters of a given genetic length (in Morgan units) fully into account. The theoretical fundament for this is the extension of Malécot's (1948) kinship concept to chromosome segments, leading to a new similarity measure called 'epistatic kinship'. It describes the probability that two randomly drawn chromosome segments of length x in Morgan are identical by descent. The same quantity, named chromosome segment homozygosity was proposed for the estimation of past effective population size (Hayes et al., 2003). It will be shown, that extensions to Wright's (1922) concept of relationship and inbreeding coefficients is straightforward. We suggest simple algorithms to set up the generalised numerator relationship matrix (NRM) and its inverse directly from a pedigree list for populations of any size and with correct accounting for inbreeding. The potential use of this approach will be demonstrated in an application to simulated data sets. Finally, possible applications of the suggested method will be discussed.

Methods

Definition of epistatic kinship, relationship, and inbreeding coefficients

We suggest to extend the concept of kinship introduced by Malécot (1948) for single loci to chromosome segments of a given length x, measured in Morgan (M). At a given chromosome segment length x, an animal S has the two complementary chromosome strands s_1 and s_2. An offspring obtains either entirely s_1 or entirely s_2 or a mixture of both, if at least one crossing over occurs in the meiosis leading to the respective gamete. If we assume that crossing over events follow a Poisson distribution, the probability that an entire strand of length x is inherited without crossing over is e^{-x}. Note that this is only strictly true when Haldane's mapping function (Haldane, 1919) is assumed. However, the main difference between mapping functions is to what extent genetic interference is taken into account (Windemuth et al., 1998), and not so much the probability that a single crossing over event happens in a short chromosome segment, which is not affected by interference. Therefore, the given probability should hold over a variety of mapping functions.

Consider an offspring T of animal S with the two chromosome strands t_1 and t_2 at the considered region. The probability that a randomly chosen strand of T, say t_i where i is either 1 or 2, is identical by descent (i.b.d.) with a randomly chosen strand s_j, j = 1 or 2, of animal S is $0.25e^{-x}$. Note that for $x = 0$ the value of $e^{-x} = 1$ and the probability equals Malécot's kinship coefficient $K_{st} = 0.25$. Due to this analogy, we suggest the term 'epistatic kinship' K_{st}^{x} for the i.b.d. probability of chromosome segments of length x between animal s and t.

The definitions of epistatic kinship, relationship and epistatic inbreeding coefficient are simultaneously derived in a companion paper (Flury et al., 2005). For a better comprehensibility of the algorithms in the following sections the basic definitions are described again here.

The analogy of Malécot's kinship coefficient K_{st} and Wright's (1922) relationship coefficient $R_{st} = 2K_{st}$ is extended to epistatic kinship and epistatic relationship, i.e. $R_{st}^{x} = 2A_{st}^{x}$.

There is also an analogy to the usual inbreeding coefficient F_j as defined by Wright (1922). Consider animal J with sire S and dam D. The kinship of individual J with itself , K_j, is the probability, that two randomly sampled alleles at one locus of this animal are i.b.d. If we denote the two alleles of J as s and d (reflecting the paternal and maternal origin), the sampled pairs (with replacement), are, with equal probability 0.25, $\{s,s\}$, $\{s,d\}$, $\{d,s\}$, or $\{d,d\}$, respectively. In half of the cases, $\{s,s\}$ and $\{d,d\}$, the two sampled alleles are clearly i.b.d. because the same alleles of animal J were sampled. If a paternal and a maternal allele are sampled, i.e. $\{s,d\}$ or $\{d,s\}$, the probability that the two alleles are i.b.d. is by definition the kinship of the parents K_{sd}. So, the kinship of individual J with itself is

$$K_j = 0.5 \times 1 + 0.5 \times K_{sd} = 0.5 \times (1 + K_{sd}) .$$

Note that

$$2 \times K_j = 1 + K_{sd} = 1 + F_j$$

since Wright's inbreeding coefficient is defined as half the relationship of the parents

$$F_j = 0.5 \times R_{sd} = K_{sd}$$

If the same concept is extended to consider chromosome segments, we have to account for crossing over events in the formation of the parental gametes. Considering the sampled pairs $\{s,d\}$ and $\{d,s\}$, the chromosome segments are only entirely i.b.d. if they were already i.b.d. in the parents, of which the probability is K_{sd}^{x}, and if they are both inherited without crossing over. Hence, for a chromosome segment of length x,

$$K_j^{x} = 0.5 \times 1 + 0.5 \times K_{sd}^{x} \times (e^{-x})^2 = 0.5 \times (1 + e^{-2x} K_{sd}^{x})$$

Using this result,

$$2 \times K_j^x = 1 + e^{-2x} K_{sd}^x = 1 + F_j^x$$

which leads to the definition of the epistatic inbreeding coefficient

$$F_j^x = e^{-2x} K_{sd}^x = 0.5 e^{-2x} R_{sd}^x$$

A tabular method to set up the epistatic numerator relationship matrix

The epistatic NRM $\mathbf{A}^x$ for N individuals is a matrix of dimension $N \times N$ where element

$A_{ij}^x = R_{ij}^x$ for $i \neq j$, and

$A_{ii}^x = 1 + F_i^x$

Note that for $x = 0$ the epistatic NRM becomes the well-known numerator relationship matrix.

Analogously to the tabular method to set up the NRM (Emik and Terrill, 1949), the following algorithm is suggested:

The animals are numbered by age from 1 to N such that the oldest animal is number 1. A pedigree list is defined giving for each animal the sire and dam number. All animals appearing as sires and dams also have to have an animal number between 1 and N. Unknown parents are denoted by a ‘0’.

Using this pedigree list, the following algorithm is performed:

1. Set $i = 1$ and $A^x_{11} = 1$
2. Set $i = i + 1$, read sire s and dam d of animal i from the pedigree list.
3. Set $A^x_{ii} = 1 + 0.5e^{-2x} A^x_{sd}$ if s and d are $\neq 0$, otherwise set $A^x_{ii} = 1$
4. Let j go from 1 to $i-1$, set $A^x_{ji} = 0.5e^{-x}(A^x_{js} + A^x_{jd})$.

   If $s = 0$ ($d = 0$) use $A^x_{js} = 0$ ($A^x_{jd} = 0$). Finally set $A^x_{ij} = A^x_{ji}$.
5. If $i < N$ continue with step 2.

After going through these steps for all animals, the epistatic NRM is complete.

A direct method to set up the inverse epistatic numerator relationship matrix

Henderson (1973) suggested the mixed model equations to estimate random genetic effects and variance components. In this system, the inverse dispersion matrix of the random effects is required. It was observed (Henderson, 1976; Quaas, 1976) that the inverse NRM, which is the dispersion matrix of the additive genetic breeding values, has some special properties, c.f. that it is extremely sparse and that simple rules can be used to derive the non-zero elements from a pedigree list. Similar observations were made for the inverse dominance and additive x additive relationship matrices (Hoeschele and VanRaden, 1991; Van Raden and Hoeschele, 1991).

To derive the inverse epistatic NRM $(\mathbf{A}^x)^{-1}$, we need to augment the pedigree list with the epistatic inbreeding coefficient for each animal. This parameter can be derived by extracting for each animal i a complete list of direct ancestors (parents, grandparents ...) from the pedigree list and computing the epistatic NRM for this subset, leading to a value for F_i^x.

Having for each animal an epistatic inbreeding coefficient, the inverse epistatic NRM can be derived by the following algorithm (we denote element i, j of $(\mathbf{A}^x)^{-1}$ as A^{ij}):

1. Preset all elements of $(\mathbf{A}^x)^{-1}$ with zero.

2. Go through all elements i=1, ..., N and add the following elements:

   Case 1: parents unknown ($s = d = 0$)

   add to element A^{ii} the value 1.0

   Case 2: one parent j known ($s = j$ and $d = 0$ or $s = 0$ and $d = j$)

   add to element A^{ii} the value $\dfrac{4}{4-(1+F_j^x)e^{-2x}}$

   add to elements A^{ij} and A^{ji} the value $-\dfrac{2e^{-x}}{4-(1+F_j^x)e^{-2x}}$

   add to element A^{jj} the value $\dfrac{e^{-2x}}{4-(1+F_j^x)e^{-2x}}$

   Case 3: both parents $s = j$ and $d = k$ known

   add to element A^{ii} the value $\dfrac{4}{4-(2+F_j^x+F_k^x)e^{-2x}}$

   add to elements A^{ij}, A^{ji}, A^{ik}, A^{ki} the value $-\dfrac{2e^{-x}}{4-(2+F_j^x+F_k^x)e^{-2x}}$

   add to elements A^{jj}, A^{kk}, A^{jk}, A^{kj} the value $\dfrac{e^{-2x}}{4-(2+F_j^x+F_k^x)e^{-2x}}$

Note that the well-known rules to set up the inverse NRM as first suggested by Henderson (1976) and Quaas (1976) are a special case of this algorithm and result for $x = 0$.

Illustration of the method

We will illustrate the suggested method with an application to the pedigree displayed in Figure 1. The corresponding pedigree list is given in Table I. The (epistatic) inbreeding coefficients in column 4 and 5 are not known a priori and are a result of the construction of the (epistatic) NRM, to be used in the construction of the inverse (epistatic) NRM.

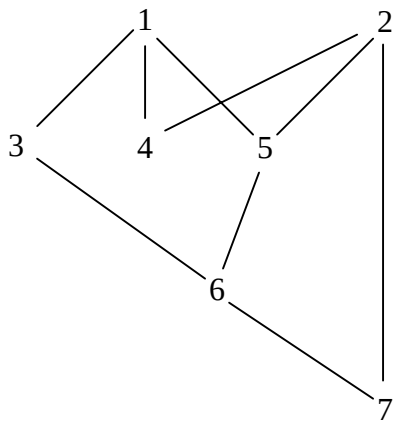


Figure 1. Pedigree for the example used as illustration.

Table I. Pedigree list for the example used as illustration, augmented by the conventional F_i and the epistatic inbreeding coefficient $F_i^{0.05}$ for a chromosome segment length of $x = 0.05$ Morgan.

Animal	Sire	Dam	F_i	$F_i^{0.05}$
1	0	0	0	0
2	0	0	0	0
3	1	0	0	0
4	1	2	0	0
5	1	2	0	0
6	3	5	0.125	0.102
7	6	2	0.125	0.102

The NRM and its inverse for this example are:

$$\mathbf{A} = \begin{bmatrix} 1. & 0. & 0.5 & 0.5 & 0.5 & 0.5 & 0.25 \\ & 1. & 0. & 0.5 & 0.5 & 0.25 & 0.625 \\ & & 1. & 0.25 & 0.25 & 0.625 & 0.3125 \\ & & & 1. & 0.5 & 0.375 & 0.4375 \\ & & & & 1. & 0.625 & 0.5625 \\ & & \textit{sym.} & & & 1.125 & 0.6875 \\ & & & & & & 1.125 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 2.333 & 1. & -.667 & -1. & -1. & 0. & 0. \\ & 2.533 & 0. & -1. & -1. & 0.533 & -1.067 \\ & & 1.833 & 0. & 0.5 & -1. & 0. \\ & & & 2. & 0. & 0. & 0. \\ & & & & 2.5 & -1. & 0. \\ & & \textit{sym.} & & & 2.533 & -1.067 \\ & & & & & & 2.133 \end{bmatrix}$$

Note that in the inverse only offdiagonal elements pertaining to parents and progeny and to mating partners are non-zero, while e.g. full- or halfsibs have zero offdiagonal elements.

Assuming a chromosome segment length $x = 0.05M$, the corresponding epistatic relationship matrix, rounded to three decimals, is:

$$\mathbf{A}^{0.05} = \begin{bmatrix} 1. & 0. & 0.476 & 0.476 & 0.476 & 0.452 & 0.215 \\ & 1. & 0. & 0.476 & 0.476 & 0.226 & 0.583 \\ & & 1. & 0.226 & 0.226 & 0.583 & 0.277 \\ & & & 1. & 0.452 & 0.323 & 0.380 \\ & & & & 1. & 0.583 & 0.504 \\ & & \textit{sym.} & & & 1.102 & 0.632 \\ & & & & & & 1.102 \end{bmatrix}$$

With segment length $x = 0.05M$ the probability that no crossing over occurs in an inherited chromosome segment is $e^{-0.05} = 0.951$. Note that for indivdual 7 with parents 6 and 2 the diagonal element is

$$A_{77}^{0,05} = 1 + 0.5e^{-2\times 0,05} A_{26}^{0,05} = 1 + 0.5 \times 0.905 \times 0.226 = 1.102 .$$

Similarly, the off-diagonal elements linking animal 7 to all ancestors are $A_{j7}^{0,05} = 0.5e^{-0,05}(A_{j2}^{0,05} + A_{j6}^{0,05})$ for all $j = 1,...,6$. For element $A_{67}^{0,05}$ this gives

$$A_{67}^{0,05} = 0.5e^{-0,05}(A_{62}^{0,05} + A_{66}^{0,05}) = 0.5 \times 0.951 \times (0.226 + 1.102) = 0.632 .$$

The following observations, which hold in general, can be made:

- zero elements in $\mathbf{A}$ are also zero in $\mathbf{A}^x$, and non-zero elements in $\mathbf{A}$ are also non-zero in $\mathbf{A}^x$
- $A_{ij}^x < A_{ij}^y$ if $i \neq j$ and $x > y$, i.e. offdiagonal elements decrease with increasing segment length
- for inbred animals, $F_i^x < F_i^y$ if $x > y$, i.e. the probability of i.b.d. chromosome segments is smaller when larger segments are considered.

The epistatic relationship between e.g. sire 1 and offspring 3 is $R_{13}^{0.05} = 0.5e^{-0.05} = 0.476$. Note, however, that the epistatic relationship between fullsibs 4 and 5 is $R_{45}^{0.05} = 0.452$, which is less than the parent-offspring epistatic relationship. This is due to the fact that there is only one meiosis between parent and offspring, while fullsibs are linked by two meioses. Therefore, the probability that in at least one of the gametes no crossing over appears shared by fullsibs is $\left(e^{-0.05}\right)^2 = 0.904$, and the resulting fullsib epistatic relationship is $R_{45}^{0.05} = 0.5\left(e^{-0.05}\right)^2 = 0.452$.

The corresponding inverse epistatic NRM for the example data set is:

$$\left(\mathbf{A}^{0.05}\right)^{-1} = \begin{bmatrix} 2.119 & 0.826 & -.614 & -.869 & -.869 & 0. & 0. \\ 0.826 & 2.258 & 0. & -.869 & -.869 & 0.431 & -.907 \\ -.614 & 0. & 1.705 & 0. & 0.413 & -.869 & 0. \\ -.869 & -.869 & 0. & 1.826 & 0. & 0. & 0. \\ -.869 & -.869 & 0.413 & 0. & 2.239 & -.869 & 0. \\ 0. & 0.431 & -.869 & 0. & -.869 & 2.258 & -.907 \\ 0. & -.907 & 0. & 0. & 0. & -.907 & 1.907 \end{bmatrix}$$

The function of the suggested algorithm can be illustrated by showing the inverse epistatic NRM after including animals 1 to 6. This matrix, indicated by $\left(\mathbf{A}^{0.05}\right)_6^{-1}$, is

$$\left(\mathbf{A}^{0.05}\right)_6^{-1} = \begin{bmatrix} 2.119 & 0.826 & -.614 & -.869 & -.869 & 0. & 0. \\ 0.826 & 1.827 & 0. & -.869 & -.869 & 0 & 0. \\ -.614 & 0. & 1.705 & 0. & 0.413 & -.869 & 0. \\ -.869 & -.869 & 0. & 1.826 & 0. & 0. & 0. \\ -.869 & -.869 & 0.413 & 0. & 2.239 & -.869 & 0. \\ 0. & 0 & -.869 & 0. & -.869 & 1.827 & 0. \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. \end{bmatrix}$$

In the next step, the elements pertaining to animal 7 with sire 6 and dam 2 are added. Using the rules suggested above, we have to add

- to the diagonal element of animal 7, A^{77}, the value $\dfrac{4}{4-(2+F_2^{0.05}+F_6^{0.05})e^{-2\times 0.05}} = \dfrac{4}{4-(2+0+0.102)\times 0.904} = 1.907$
- to the offdiagonal elements, A^{72}, A^{76}, A^{27}, A^{67} linking animal 7 to its parents the value $-\dfrac{2e^{-0.05}}{4-(2+F_2^{0.05}+F_6^{0.05})e^{-2\times 0.05}} = -\dfrac{2\times 0.951}{4-(2+0+0.102)\times 0.904} = -0.907$

- to the diagonal elements A^{22} and A^{66} pertaining to the parents and the corresponding offdiagonals A^{26} and A^{62} the value

$$\frac{e^{-2\times 0.05}}{4-(2+F_2^{0.05}+F_6^{0.05})e^{-2\times 0.05}} = \frac{0.904}{4-(2+0+0.102)\times 0.904} = 0.431$$

Adding these values to the respective matrix $\left(\mathbf{A}^{0.05}\right)_6^{-1}$ results in $\left(\mathbf{A}^{0.05}\right)^{-1}$. It should be noted that $(\mathbf{A}^{0.05})^{-1}$ and $\mathbf{A}^{-1}$ are structurally very similar, in that the positions of zero and non-zero elements are identical. Also, matrix elements in both cases are only affected by inbreeding if the parents are inbred, regardless whether or not the resulting offspring is also inbred. This becomes obvious with animal 7 in the example pedigree, which is an offspring of the inbred sire 6. This sire's inbreeding coefficient is the reason, why the diagonal element A^{77} is different from the diagonal element A^{44}, even though the amount of information (both parents known, no offspring) is identical for individual 4 and 7.

Proof of concept

The potential usefulness of the suggested methodology will be demonstrated in a simulation study. We simulated data using the following genetic model: on a chromosome segment of length 0.1 M two biallelic loci with alleles P, p and Q, q formed an epistatic complex. Both loci had neither additive nor dominance effects, but the epistatic combinations $P-Q$ and $p-q$ had the epistatic effect +1 and $P-q$ and $p-Q$ had the epistatic effect –1. This results in total genotypic effects of +4 for animals with combined genotype $PPQQ$ or $ppqq$ and in an effect of –4 for animals with combined genotype $PPqq$ or $ppQQ$, respectively, while all other combined genotypes, containing at least one heterozygous single locus genotype, have the effect 0. With an allele frequency of 0.5 for all alleles the genetic variance for such an epistatic complex is 4.

We simulated a base population with 50 male and 50 female unrelated individuals. Each individual had ten independently segregating epistatic complexes (conceptually located on ten different chromosomes) of the described type. So, the genetic variance due to these epistatic complexes is $\sigma_x^2 = 40$. In addition, a polygenic additive component was simulated with variance $\sigma_a^2 = 40$ and the residual variance was assumed to be $\sigma_e^2 = 80$. Starting from this base population, nine subsequent generations of equal size and sex ratio were generated at random. For the epistatic complexes, linked mendelian inheritance was assumed and the recombination rate between the two loci was generated assuming a Poisson distribution of crossing over events. Each animal had a phenotype, made up of the additive and the total epistatic effect and the error term, apart from an overall mean μ no fixed effects were assumed. The whole simulation procedure was repeated 100 times.

From the resulting population of 1000 individuals for each replicate, variance components were estimated under a mixed model of the type

$$y = \mathbf{1}\mu + \mathbf{I}a + \mathbf{I}a_x + \mathbf{I}e$$

where

y is the vector of observations

μ is the overall mean

a is a vector of random additive breeding values

a_x is a vector of random epistatic (linked additive x additive) effects

e is a random error term

$\mathbf{1}, \mathbf{I}$ are a column vector of ones and the identity matrix used as incidence matrices pertaining to μ and both a and a_x.

The observation vector has the multivariate normal distribution

$$y \sim MVN(\mathbf{1}\mu, \mathbf{A}\sigma_a^2 + \mathbf{A}^x\sigma_x^2 + \mathbf{I}\sigma_e^2)$$

where $\mathbf{A}$ and $\mathbf{A}^x$ are the additive and epistatic NRM and σ_a^2, σ_x^2, and σ_e^2 are the variance components pertaining to additive, epistatic (in the sense defined above), and residual random terms.

Under this model residual maximum likelihood (REML, Patterson and Thompson, 1971) estimates of variance components were estimated using the program DFREML (Meyer, 1998). This, however, is only possible conditional on a defined segment length x, since the dispersion matrix or, more accurately, its inverse $(\mathbf{A}^x)^{-1}$ need to be provided externally to the program. We therefore calculated six different inverses $(\mathbf{A}^x)^{-1}$ for $x = 0.$ to $x = 0.15$ in steps of 0.025, for $x = 0.$ the model is equivalent to a purely additive model. For each such matrix, a full DFREML estimation of the variance components, conditional on the assumed value of x, was conducted.

The most likely value of x was identified with a grid search over the predefined values of x. However, we observed that the final log-likelihood value provided by the DFREML program consistently grew with increasing values of x. This is caused by the fact, that DFREML considers the log-determinants of the dispersion matrices as constant and therefore the likelihood is not comparable between runs using different dispersion matrices (Meyer, 1991). Therefore, we calculated the full log-likelihood of the data, using the converged variance components and the estimate of μ taken from the DFREML solutions as

$$\log L(y) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\hat{\mathbf{V}}\right| - \frac{1}{2}(y - \mathbf{1}\hat{\mu})'\hat{\mathbf{V}}^{-1}(y - \mathbf{1}\hat{\mu})$$

where

n is the number of individuals in the sample

$\hat{\mu}$ is the estimate of the mean

and

$$\hat{\mathbf{V}} = \mathbf{A}\hat{\sigma}_a^2 + \mathbf{A}^x\hat{\sigma}_x^2 + \mathbf{I}\hat{\sigma}_e^2$$

is the estimated variance-covariance matrix based on the REML estimates of the variance components. We accepted the value of x giving the highest log-likelihood as the best estimate and used the corresponding estimates of the variance components for the final evaluation of the 100 replicates.

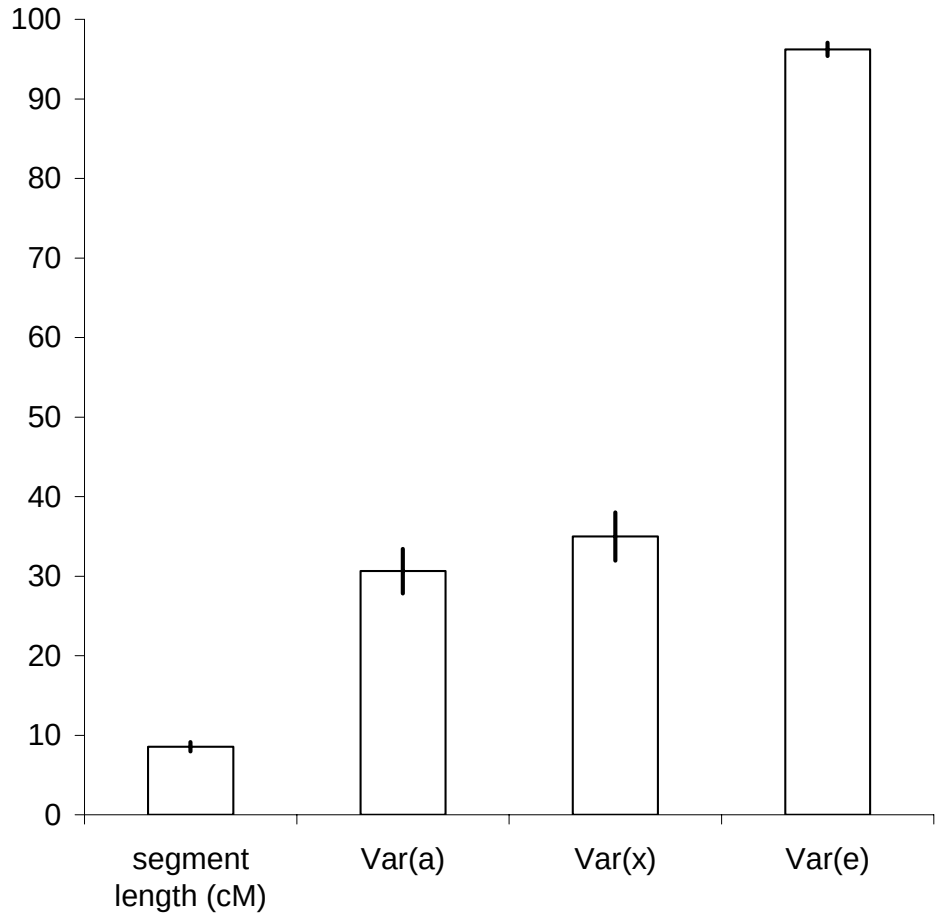


Figure 2. Means ± standard errors of means of the maximum likelihood estimates for the chromosome segment length and the additive (Var(a)), the epistatic (Var(x)), and the residual (Var(e)) variance component obtained from the simulation study.

Figure 2 shows the means of the so obtained maximum likelihood estimates for the chromosome segment length and the additive, the epistatic, and the residual variance component, together with the standard error of the means. Although the segment length is slightly underestimated ($\hat{x} = 0{,}086 \pm 0{,}0058$), the value is nicely within the expected range. Both genetic variance components are underestimated (the simulated values were 40), while the residual variance is upward biased, so that the total variance of 160 is accurately estimated.

On average, the mean estimate ± the standard error of the mean for the additive and the epistatic variance are $\hat{\sigma}_a^2 = 30{,}6 \pm 2{,}8$ and $\hat{\sigma}_x^2 = 35{,}0 \pm 3{,}0$, respectively. For both variance components, a large variation of estimates between replicates was observed with estimates ranging between close to 0 and 76 for $\hat{\sigma}_a^2$ and 0 and 93 for $\hat{\sigma}_x^2$, respectively. At the same time, a strong negative correlation between the estimates of these two genetic components was found, illustrating the difficulty to correctly dissect the two sources of variability. Although the likelihood profile was observed to be rather flat in most cases, twice the difference between the highest log-likelihood and the log-likelihood obtained from the purely additive model was up to 5.58 in some replicates, which corresponds to a significance of the epistatic model vs. the purely additive model with an error probability $\alpha < 0.02$ in the likelihood ratio test.

This small simulation study is primarily meant as a proof of concept, and thus a detailed analysis of the sources of bias and a derivation of the necessary sample sizes and data structures to obtain sufficiently accurate estimates is clearly beyond the scope of it. Nevertheless these results show, that the suggested model is applicable to farm animal data structures and has the potential to dissect the total genetic variance according to the suggested model.

Discussion

This study suggests to use the i.b.d. probability of chromosome segments of a given genetic length as a similarity index between individuals. Hayes et al. (2003) suggested the estimation of past effective population size based on observed disequilibrium of linked markers called chromosome segment homozygosity. In a companion paper (Flury et al., 2005) we suggest the average epistatic kinship within and between populations as a new measure for genetic similarity of populations and show that the phylogenetic resolution is higher compared to traditional single-locus similarity measures. Both studies, however, are based on the neutrality assumption of the considered chromosome segments, while in the present study a genetic effect of the considered chromosome segment is assumed.

The approach suggested above is a generalisation of the usual quantitative genetic model. It is especially attractive that the former model is a special case of the epistatic model with $x = 0$. This is also true for the described algorithms to set up the epistatic NRM and its inverse directly, which, for $x = 0$, simplify to the well-known algorithms to set up the NRM and its inverse.

The suggested algorithm to set up $(\mathbf{A}^x)^{-1}$ is linear in N, given that epistatic inbreeding coefficients are available. It would have been possible to do all computations in one recursive algorithm, comparable to the one suggested by Quaas (1976) to set up the inverse NRM, however the computations then are proportional to N^2, which might be prohibitive for large pedigrees. It is also possible to set up the inverse epistatic NRM implicitly in an 'iteration on the data' type of algorithm (Schaeffer and Kennedy, 1985) to estimate epistatic breeding values.

Using epistatic relationship in a mixed model based selection has the potential to pick up some of the non-additive genetic components, which are ignored in purely additive models (Falconer and Mackay, 1996). More precisely, the approach will account for additive x additive effects of genes which are in linkage disequilibrium. This extra genetic gain holds for few generations and slowly but continuously with rate e^{-x} erodes due to crossing over. However, under a short to medium term perspective (one or few

generations), breeding programs may benefit from thise extra genetic gain, even though it is not fully sustainable in the long term.

The suggested model is different from other models in this area, in that it combines epistatic effects with the inherent stability of linkage groups. Other models accounting for epistatic effects (Du and Hoeschele, 2000; Fuerst and Soelkner, 1994; Hoeschele, 1991; Hoeschele and VanRaden, 1991; Tempelman and Burnside, 1990; Van Raden and Hoeschele, 1991; VanRaden et al., 1992) ignored the possible linkage of interacting genes. Models considering the effects of linkage were very specific to certain genes or genomic regions, requiring gene or marker information pertaining to a specific chromosomal region. The epistatic model suggested here accounts for the entity of unspecified and non localised gene complexes of a given segment length and sums the respective effects over the whole genome.

The reported results of the simulation study show, that the suggested concept is applicable to farm animal data sets and allows, in principle, to disentangle the epistatic from the additive genetic variance component. It was observed, though, that with the sample sizes underlying the reported simulation study the power to separate these two variance components σ_a^2 and σ_x^2, is limited. The primary reason is, that with small values of x the matrices $\mathbf{A}$ and $\mathbf{A}^x$ are rather similar, which makes the corresponding variance components almost exchangeable. This corresponds with the situation where a mixed model contains both the NRM and an i.b.d.-probability matrix as dispersion matrix of a QTL conditional on marker information, as originally suggested by Fernando and Grossman (1989). With a typically low proportion of genotyped animals and eventually markers of limited information content and linkage to the QTL position, these two dispersion matrices often will tend to be very similar, creating problems to statistically disentangle the corresponding variance components (Simianer, 1994). In the mixed additive and epistatic model, the power to estimate additive and epistatic variance components depends primarily on the size and structure of the data and the magnitude of the true effect. Using the suggested algorithms, applications to much larger real data sets are possible, which may help to avoid some of the discussed limitations.

Despite these practical problems, the suggested model provides a novel perspective to genetic analyses and might be an option to use a larger share of the total genetic variation in selection programs.

Acknowledgements

This study was conducted with financial support through the Deutsche Forschungsgemeinschaft (DFG) which is gratefully acknowledged.

References

De Boer, I. J. M., and Hoeschele, I. 1993. Genetic evaluation methods for populations with dominance and inbreeding. Theoretical and Applied Genetics 86: 245-258.

Du, F.-X., and Hoeschele, I. 2000. Estimation of additive, dominance and epistatic variance components using finite locus models implemented with a single-site Gibbs and a descent graph sampler. Genetical Research 76: 187-198.

Emik, L. O., and Terrill, C. R. 1949. Systematic procedures for calculating inbreeding coefficients. Journal of Heredity 40: 51-55.

Falconer, D. S., and Mackay, T. F. C. 1996. Introduction to Quantitative Genetics. 4. ed. Longman Group Ltd., Essex.

Fernando, R. L., and Grossman, M. 1989. Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution 21: 467-477.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. R. Soc. (Edinburgh), Trans. 52: 321-341.

Flury, C., Tietze, M., and Simianer, H. 2005. Epistatic Kinship a new measure of genetic diversity for short term phylogenetic structures -theoretical investigations. Journal of Animal Breeding and Genetics (in press).

Fuerst, C., and Soelkner, J. 1994. Additive and non additive genetic variance of milk yield, fertility, and life time performance traits of dairy cattle. J. Dairy Sci. 77: 1114-1125.

Haldane, J. B. S. 1919. The combination of linkage values and the combination of distance between the loci of linkage factors. J. Genet. 8: 299-309.

Hayes, B. J., Visscher, P. M., McPartlan, H., and Goddard, M. E. 2003. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research 13: 635-643.

Henderson, C. R. 1973. Sire evaluation and genetic trends. p 10-41.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. Biometrics 32: 69-83.

Hoeschele, I. 1991. Additive and non additive genetic variance in female fertility of Holsteins. Journal of Dairy Science 74: 1743-1752.

Hoeschele, I., and VanRaden, P. M. 1991. Rapid Inversion of Dominance Relationship Matrices for Noninbred Populations by Including Sire by Dam Subclass Effects. Journal of Dairy Science 74: 557-569.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie., Paris.

Mason, M., Lee, E., Westphal, H., and Reitman, M. 1995. Expression of the chicken beta-globin gene cluster in mice: correct developmental expression and distributed control. Mol. Cell. Biol. 15: 407-414.

Meyer, K. 1991. Estimating variances and covariances for multitrait animals models by restricted maximum likelihood. Genetics Selection Evolution 23: 67-83.

Meyer, K. 1998. DFREML- User Notes.

Patterson, H. D., and Thompson, R. 1971. Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika 58: 545-554.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32: 949.

Schaeffer, L. R., and Kennedy, B. W. 1985. Computing strategies for solving mixed model equations. Journal of Dairy Science 69: 575.

Simianer, H. 1994. Derivation of single locus relationship coefficients conditional on marker information. Theoretical and Applied Genetics 88: 548-556.

Tempelman, R. J., and Burnside, E. J. 1990. Additive and nonadditive genetic variation for production traits in Canadian Holsteins. Journal of Dairy Science 73: 2206-2213.

Van Raden, P. M., and Hoeschele, I. 1991. Rapid inversion of additive by additive relationship matrices by including sire-dam combination effects. Journal of Dairy Science 74: 570-579.

VanRaden, P. M., Lawlor, T. J., Short, T. H., and Hoeschele, I. 1992. Use of Reproductive Technology to Estimate Variances and Predict Effects of Gene Interactions. Journal of Dairy Science 75: 2892-2901.

Windemuth, C., Simianer, H., and Lien, S. 1998. Fitting genetic mapping functions based on sperm typing: Results for three chromosomal segments in cattle. Animal Genetics 29: 425-434.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56: 330-339.

3rd CHAPTER

EPISTATIC KINSHIP A NEW MEASURE OF GENETIC DIVERSITY FOR SHORT TERM PHYLOGENETIC STRUCTURES – THEORETICAL INVESTIGATIONS

Christine Flury, Manfred Tietze and Henner Simianer

Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen,
Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

Journal of Animal Breeding and Genetics (in press)

EPISTATIC KINSHIP A NEW MEASURE OF GENETIC DIVERSITY FOR SHORT TERM PHYLOGENETIC STRUCTURES - THEORETICAL INVESTIGATIONS

Christine Flury, Manfred Tietze and Henner Simianer

Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany

Abstract

The epistatic kinship describes the probability that chromosomal segments of length x in Morgan are identical by descent. It is an extension from the single locus consideration of the kinship coefficient to chromosomal segments. The parameter reflects the number of meioses separating individuals or populations. Hence it is suggested as measure to quantify the genetic distance of sub-populations that have been separated only few generations ago. Algorithms for the epistatic kinship and the extension of the rules to set up the rectangular relationship matrix are presented. The properties of the epistatic kinship based on pedigree information were investigated theoretically. Pedigree data is often missing for small livestock populations. Therefore, an approach to estimate epistatic kinship based on molecular marker data is suggested. For the epistatic kinship based on marker information haplotypes are relevant. An easy and fast method that derives haplotypes and the respective frequencies without pedigree information was derived based on sampled fullsib pairs. Different parameters of the sampling scheme were tested in a simulation study. The power of the method decreases with increasing segment length x and with increasing number of segments genotyped. Further it is shown, that the efficiency of the approach is influenced by the number of animals genotyped and the polymorphism of the markers. It is discussed, that the suggested method has a considerable potential to allow a phylogenetic differentiation between close populations, where small sample size can be balanced by the number, the length, and the degree of polymorphism of the chromosome segments considered.

Introduction

Phenotypic selection since domestication has created a wide diversity of breeds of domestic animal that are adapted to different climatic conditions and purposes (Andersson, 2001). Today more than 20% of the roughly 6400 documented breeds are at risk of extinction (Scherf, 2000). Due to limited financial and human resources, not all breeds can be given the same priority for conservation (Oldenbroek, 1999). One – but not the only – important criterion (Ruane, 1999) is the uniqueness of breeds. Genetic distance studies are based on evolutionary models which often do not hold for the development of livestock breeds. Most of the approaches were developed for the description of the evolutionary differentiation between species, while for livestock the differentiation occurred within species (Ruane, 1999; Simianer, 2002).

The formation of today's breeds goes back to the 19th or even the beginning of the 20th century (Sambraus, 2001). Thus the assumption of an evolutionary time span does not hold for breed differentiation. Based on the reduced divergence time the role of mutation in creating differences between breeds is expected to be small (Takezaki and Nei, 1996; Toro and Caballero, 2004).

Toro and Caballero (2004) summarized further problems of conservation decisions based on phylogenetic diversity like the complete ignorance of genetic variance within population, the failure of principles of phylogeny reconstruction to account for population admixture, the problems arising from varying distances among the markers used and the impact of the demographic history of a population. Also, markers used for genetic distances are assumed to represent neutral loci.

Ignoring the genetic variance within population often leads to the conservation of the most inbred population (Eding, 2002). To overcome this weakness of genetic distances the authors proposed mean coefficients of kinship between and within populations as tool to assess genetic similarity in livestock populations. The coefficient of kinship K_{st} is defined as the probability that two randomly sampled alleles from the same locus in two individuals S and T are identical by descent (*ibd*) (Malécot, 1948). Another concept for the estimation of genetic similarity between individuals is the coefficient of relationship R_{st} specified by (Wright, 1922). The link between the two parameters is

$R_{st} = 2K_{st}$. Kinship coefficients can be calculated based on pedigree information (Cockerham, 1967). As pedigree data is often not available for small livestock populations, some authors suggested the estimation of kinship coefficients based on marker information (Caballero and Toro, 2000; Eding and Meuwissen, 2001). Having non-unique founder alleles the correction for alleles identical by state, but not identical by descent is crucial. Lynch (1988) proposed a similarity index to overcome this problem for single loci. Eding and Meuwissen (2001) showed, that marker based estimates of kinship yielded higher correlations with pedigree-based kinships than genetic distance measures.

Coefficients of kinship refer to the *ibd* probability for a randomly chosen single locus or an average over all loci (Simianer, 1994). This presumes independently segregating loci. For the genetic control of important traits the formation of gene complexes over multiple loci and epistatic interactions is important (Brockmann et al., 2000). Various studies investigate the properties of conserved haplotypes around a functional polymorphism. Haplotype sharing is important in the context of *ibd*-mapping of QTLs (Meuwissen and Goddard, 2000; Nezer et al., 2003). The length of conserved haplotypes depends on the timespan since separation or rather the number of recombination events. Visscher (2003) suggests, that linkage disequlibrium (*LD*) created by crossbreeding may still persist in many of todays livestock populations, because crossbreeding was commonly practised 50 to 100 generations ago. Coppieters et al. (1999) and Farnir et al. (2000) found strong evidence for long range *LD* for all autosomes of the Holstein Friesian population, with *LD* extending over regions greater than 20 cM. Beside other factors they explain the disequilibrium particularly with drift, due to the small effective population size of the Holstein Friesian population.

In this study we assume the existence of *LD* for small livestock populations and propose a diversity measure based on shared haplotypes within and between populations. Therefore the coefficient of kinship will be extended from single loci to chromosomal segments of length x in Morgan. This leads to a new similarity index called epistatic kinship, which describes the probability of chromosomal segments being identical by

descent. A similar measure was proposed by Hayes et al. (2003) as chromosome segment homozygosity for the estimation of past effective population size.

In the method section this parameter will be defined and algorithms to calculate epistatic kinship, epistatic relationship coefficient, epistatic inbreeding and the epistatic kinship matrix will be presented. An extension from the average homozygosity (Falconer and Mackay, 1996) to average expected epistatic kinship is derived. The properties of the average epistatic kinship as a tool for the analysis of short term phylogenetic structures are investigated for a known simulated pedigree structure in the first results section. In the second results section of the results the epistatic kinship will be estimated based on marker information. Typing of animals results in genotypes, thus a method to derive haplotypes from genotyping information is needed. Different algorithms to infer haplotypes exist and are discussed by Niu (2004). For some algorithms pedigree information is a prerequisite, others who run without pedigree information are often complex and computing intensive (Windig and Meuwissen, 2004). An easy and fast method to derive haplotypes without pedigree information or in simple standard pedigrees (e.g. only fullsib pairs are available) is suggested. The efficiency of the differentiation of close populations based on average epistatic kinship was compared for reconstructed vs. true haplotypes.

Methods

Epistatic kinship, epistatic relationship and epistatic inbreeding

We define K_{st} as Malécot's (1948) kinship coefficient between individual S and T, reflecting the probability that a randomly chosen allele at a given locus of individual S is *ibd* with a randomly chosen allele at the same locus in animal T, Consider now a randomly chosen chromosome segment of length x Morgan. We chose at random one of the two homologous strands of this chromosome segment in individual S and T, respectively. We define K_{st}^{x} as the probability, that these two strands are *ibd* and call this parameter 'epistatic kinship'. This name is derived from the use of the same parameter to estimate epistatic effects in gene clusters which is described in a companion paper (Flury et al., 2005).

The extension from single locus to chromosomal segments requires a correction for the probability that crossing over occurs. Under the assumption that crossing over events follow a Poisson distribution, the probability that an entire chromosome strand of length x is inherited without crossing over is e^{-x}. Consider an offspring T of animal S with the two strands t_1 and t_2 at the considered region. The probability that a randomly chosen strand of T, say t_i where i is either 1 or 2, is identical by descent with a randomly chosen strand s_j, j = 1 or 2, of animal S is $K_{st}^x = K_{st} \times e^{-x}$ thus $0.25e^{-x}$. Note that for $x = 0$ the value of $e^{-x} = 1$ and the probability equals the kinship coefficient $K_{st} = 0.25$, hence Malécot's kinship coefficient is a special case of the epistatic kinship coefficient for $x = 0$.

It is straightforward to extend the analogy of Malécot's kinship coefficient K_{st} and Wright's (1922) relationship coefficient $R_{st} = 2K_{st}$ to epistatic kinship and epistatic relationship, i.e. $R_{st}^x = 2K_{st}^x$.

There is also an analogy to the usual inbreeding coefficient F_j as defined by Wright (1922). For the extension to chromosome segments, we have to account for crossing over events in the formation of the parental gametes.

Epistatic inbreeding can be derived from the epistatic kinship of an individual with itself. Consider animal J with sire S and dam D and denote the two homologous strands of individual J at a given chromosome segment as s and d, reflecting the paternal and maternal origin. We sample at random two strands (with replacement) of individual J. The sampled pairs are, with equal probability 0.25, $\{s,s\}$, $\{s,d\}$, $\{d,s\}$, or $\{d,d\}$, respectively. In half of the cases, $\{s,s\}$ and $\{d,d\}$, the two sampled strands are clearly *ibd* because the same strands of animal J were sampled. For the sampled pairs $\{s,d\}$ and $\{d,s\}$, the chromosome segments are only entirely *ibd* if they were already *ibd* in the parents, of which the probability is K_{sd}^x, and if they were both inherited without crossing over. Hence, for a chromosome segment of length x,

$$K_j^x = 0.5 \times 1 + 0.5 \times K_{sd}^x \times (e^{-x})^2 = 0.5 \times (1 + e^{-2x} K_{sd}^x)$$

Using this result,

$$2 \times K_j^x = 1 + e^{-2x} K_{sd}^x = 1 + F_j^x$$

which leads to the definition of the epistatic inbreeding coefficient

$$F_j^x = e^{-2x} K_{sd}^x = 0.5 e^{-2x} R_{sd}^x$$

The epistatic relationship matrix

The epistatic relationship matrix A^x for N individuals is a matrix of dimension $N \times N$ where element

$$A_{ij}^x = R_{ij}^x \qquad \text{for } i \neq j \text{, and}$$

$$A_{ii}^x = 1 + F_i^x$$

Note that for $x = 0$ the epistatic relationship matrix becomes the well-known numerator relationship matrix.

Analogously to the tabular method to set up the numerator relationship matrix (Emik and Terrill, 1949), the following procedure is suggested.

The animals are numbered by age from 1 to N such that the oldest animal is number 1. A pedigree list is defined giving for each animal the sire and dam number. All animals appearing as sires and dams also have to have an animal number between 1 and N. Unknown parents are denoted by a ‘0’.

Using this pedigree list, the following algorithm is performed:

1. Set $i = 1$ and $A^x_{11} = 1$
2. Set $i = i + 1$, read sire s and dam d of animal i from the pedigree list.
3. Set $A^x_{ii} = 1 + 0.5e^{-2x} A^x_{sd}$ if s and d are $\neq 0$, otherwise set $A^x_{ii} = 1$
4. Let j go from 1 to $i-1$, set $A^x_{ji} = 0.5e^{-x}(A^x_{js} + A^x_{jd})$. If $s = 0$ ($d = 0$) use $A^x_{js} = 0$ ($A^x_{jd} = 0$). Finally set $A^x_{ij} = A^x_{ji}$.
5. If $i < N$ continue with step 2.

After going through these steps for all animals, the epistatic relationship matrix is complete. The junction between the epistatic relationship matrix A^x and the epistatic kinship matrix K^x is $\mathrm{K}^x = 0.5\mathrm{A}^x$.

Expected epistatic kinship within and between populations

Assuming an ideal population of size N, the average homozygosity F_t in generation generation t can be computed by the recursive formula (Falconer and Mackay, 1996)

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}. \qquad [1]$$

This equation is made up from two parts: the first expression $\frac{1}{2N}$ is the 'new' homozygosity which is generated in the meiotic sampling of the gametes leading to generation t, and $(1 - \frac{1}{2N})F_{t-1}$ is the 'old' homozygosity which was built up in generations 1 to $t-1$.

If we use the same rationale to derive the expected epistatic kinship for a chromosome segment of length x, we have two processes, which overlay each other: in each generation, new epistatic kinship is generated by the sampling process, while at the same time old epistatic kinship is partly destroyed through crossing over.

In generation t, $2N$ chromosome segments are sampled from the pool of chromosome segments in generation $t-1$. Each chromosome segment will show no crossing over with probability e^{-x}. Therefore, the probability that two randomly chosen chromosome segments in generation t are new epistatic homozygotes is $\frac{e^{-2x}}{2N}$. Old epistatic homozygotes may lose this property in any subsequent generation. The probability that an old epistatic homozygote existing in generation $t-1$ stays homozygote in generation t is e^{-2x}. Combining these findings, the average expected epistatic kinship $\overline{K}_t^x$ in generation t can be calculated by the recursive formula

$$\overline{K}_t^x = \frac{e^{-2x}}{2N} + e^{-2x}(1-\frac{1}{2N})\overline{K}_{t-1}^x = e^{-2x}\left[\frac{1}{2N} + (1-\frac{1}{2N})\overline{K}_{t-1}^x\right]. \qquad [2]$$

Note that the recursion [1] for single loci is a special case with $x = 0$.

The resulting function of $f(t) = \overline{K}_t^x$ is convex and asymptotically goes for $t \to \infty$ to

$$\overline{K}_{\max.}^x = \frac{e^{-2x}}{e^{-2x} + 2N(1-e^{-2x})}. \qquad [3]$$

If a population is split in sub-populations in generation t' and these sub-populations are maintained without genetic exchange, no new epistatic kinship will be generated between these populations. The average epistatic kinship on the level of the time of fission will be maintained as the epistatic kinship between these populations if $x = 0$, but this old epistatic kinship will erode with the rate e^{-2x} in every generation through crossing over with $x > 0$. Thus, the between population expected average epistatic kinship in generation g after fission is

$$\overline{K}_{t'+g}^x = e^{-2xg}\overline{K}_{t'}^x \qquad [4]$$

Note that the rate of erosion of epistatic kinship between separated populations is independent of the population size.

Epistatic kinship based on pedigree information

It is suggested to use the epistatic kinship to differentiate phylogenetically close populations. The hypothesis is, that this metric is more sensitive to small phylogenetic distances caused by short time since separation than conventional distance metrics, which are based on mutation and/or genetic drift as the diversity generating process. It was assumed, that the full pedigree of two sub-populations back to a common base population was known. Samples were taken from the two sub-populations in the latest generation and it was tested, whether the average epistatic kinship between populations differed from the average epistatic kinship within populations.

The test was based on a random sample of M individuals in each of the two populations. For these individuals, L chromosome segments of length x were considered. For each pair of the $2M$ individuals the epistatic kinship was calculated using the tabular method described above.

For the statistical test, it was necessary to take the number of informative comparisons into account. An illustration and the corresponding approximations for the number of informative comparisons within populations N_w and between populations N_b are given in the Appendix.

Because in each comparison four different pairs of chromosome segments can be compared, the number of pairwise comparisons within (V_w) and between (V_b) populations are:

$$V_w = N_w * 4L$$

$$V_b = N_b * 4L$$

Note, that the number of comparisons within and between populations is a linear function of the number of chromosome segments considered, L, and a quadratic function of the number of animals sampled, M.

The average ibd-probability within populations is denoted as p_w and the average ibd-probability between populations is denoted as p_b.

To test the hypothesis

$$H_o : p_w = p_b = p_0$$

versus

$$H_a : p_w > p_b$$

the χ^2-test statistic was calculated using the basic formula

$$X^2 = \frac{(p_w V_w - p_0 V_w)^2}{p_0 V_w} + \frac{\left[(1-p_w)V_w - (1-p_0)V_w\right]^2}{(1-p_0)V_w} + \frac{(p_b V_b - p_0 V_b)^2}{p_0 V_b} + \frac{\left[(1-p_b)V_b - (1-p_0)V_b\right]^2}{(1-p_0)V_b}$$

Using the average ibd-probability p_0 under the null hypothesis

$$p_0 = \frac{p_w * V_w + p_b * V_b}{V_w + V_b}$$

the expected test statistic is

$$E(X^2) = \frac{(p_w - p_0)^2 * V_w + (p_b - p_0)^2 * V_b}{p_0} + \frac{(p_w - p_0)^2 * V_w + (p_b - p_0)^2 * V_b}{(1-p_0)} \quad [5]$$

Since this test statistic is not based on actual, but expected numbers of *ibd* segments under a specific realisation of the alternative hypothesis, we denote $E(X^2)$ as the expected test statistic and assume, that a higher value of this parameter corresponds with a higher power.

Epistatic kinship based on marker information

In applications to real life data, the pedigree of animals from different populations back to common ancestors from one common base population rarely is available. Therefore, it is necessary to assess the *ibd* status of chromosome segments based on genotyping information from marker sets spanning a given chromosome segment length. Typing individuals for certain markers results in genotypes. For the estimation of the epistatic kinship within and between populations haplotypes are relevant. Haplotype reconstruction for individuals without known relationship is of limited efficiency. Therefore it was assumed that genotyping was done for fullsib pairs. Drawing fullsib pairs *(FSP)* for the sample is possible without pedigree information for multiparous species like pigs before weaning.

For the proposed method the genotypes of each pair are compared and it is postulated, that alleles which are common between fullsibs potentially are identical by descent. In the comparison of genotypes three different cases can occur. In the first case there is no common allele found for at least one locus in the two genotypes of the pair. In this case inferring the haplotypes is not possible and the pair is not informative. The second case occurs when for the pair under consideration exactly one common haplotype is possible. In the third case different combinations of common haplotypes are possible, due to common alleles at least at one locus for equally heterozygous animals. If this is the case for m loci, 2^m different common haplotype combinations are possible. For the informative cases 2) and 3) the possible common haplotypes were derived. In case 3, the different possible common haplotype combinations were assigned with probability 2^{-m}, respectively.

The statistical test conducted is based on the assumption, that *ibd* haplotypes are more likely found within than between populations. Consider a situation where two samples of animals are taken. The null hypothesis is, that the two samples originate from the same population, while the alternative hypothesis is, that the two samples originate from different populations.

To verify this, a test statistic based on the accumulation of pairwise individual comparisons is suggested.

We compare two animals, I and J, at one chromosome segment, which, for simplicity of illustration, is assumed to be made up from two loci only. The observed genotypes are $G_i = \{1,2\,;1,2\}$ and $G_j = \{1,2\,;1,3\}$. Haplotype reconstruction results for both animals in $k = 2$ alternative haplotype combinations denoted as $G_i = \begin{Bmatrix} H_{ik1} \\ H_{ik2} \end{Bmatrix}$ and $G_j = \begin{Bmatrix} H_{jk1} \\ H_{jk2} \end{Bmatrix}$.

The possible haplotype combinations and their corresponding probabilities are:

$$G_i = \begin{Bmatrix} H_{i11} \\ H_{i12} \end{Bmatrix} = \begin{Bmatrix} 1-1 \\ 2-2 \end{Bmatrix} \quad p_{i1} = 0{,}5 \qquad G_j = \begin{Bmatrix} H_{j11} \\ H_{j12} \end{Bmatrix} = \begin{Bmatrix} 1-1 \\ 2-3 \end{Bmatrix} \quad p_{j1} = 0{,}5$$

$$G_i = \begin{Bmatrix} H_{i21} \\ H_{i22} \end{Bmatrix} = \begin{Bmatrix} 2-1 \\ 1-2 \end{Bmatrix} \quad p_{i2} = 0{,}5 \qquad G_j = \begin{Bmatrix} H_{j21} \\ H_{j22} \end{Bmatrix} = \begin{Bmatrix} 2-1 \\ 1-3 \end{Bmatrix} \quad p_{j2} = 0{,}5$$

Next, each of the four possible haplotypes of animal I is compared with each of the four possible haplotypes of animal J. At this stage it is not relevant, whether the two individuals are from the same or from different samples. If two haplotypes are identical, the product of the corresponding haplotype probabilities is accumulated in the variable S_{ij}. In the present example, $H_{i11} = H_{j11}$ and $H_{i21} = H_{j21}$, so that

$$S_{ij} = p_{i1}p_{j1} + p_{i2}p_{j2} = 0{,}25 + 0{,}25 = 0{,}5\,.$$

For all within population comparisons, the average value of this variable is denoted as $\overline{S}_w$, while for all between population comparisons, the average value is denoted as $\overline{S}_b$. Since under the alternative hypothesis we assume, that common haplotypes are more likely within than between populations,

$$S = \overline{S}_w - \overline{S}_b \qquad [6]$$

is a suitable test statistic.

To verify the loss of information due to haplotype reconstruction, this test was applied in two forms:

a) It was assumed, that the true haplotypes were observed, i.e. that not only the genotypes, but also the specific haplotype combination of an animal was observable. In this case, only one of the possible haplotype combinations received the probability 1 and all other possible haplotype combinations have the probability 0. Based on these probabilities, the test statistic S was calculated and is henceforth indicated as S_t (*t* standing for 'true').

b) To account for the uncertainty of haplotype reconstruction, the haplotype probabilities derived from full-sib genotypings as indicated above were used, the resulting test statistic is indicated as S_r (*r* standing for 'reconstructed').

In both cases, the expected value under the null hypothesis (the two samples originate from the same population) is $E(S_t) = E(S_r) = 0$, while under the alternative hypothesis, we would expect that S_t and S_r take positive values. The distributions of the test statistics under the null hypothesis need to be determined empirically, either through simulation or through a permutation test approach (Doerge and Churchill, 1996).

Simulation

An existing FORTRAN-Code was extended for the simulations in this study. A base population of 50 males and 50 females was generated. All animals were assumed to be unrelated and genotypes at the required number of loci were assigned at random, assuming the base population to be in Hardy-Weinberg and linkage equilibrium.

Under the null hypothesis, 15 generations of random mating and constant population size were simulated. For testing purposes the number of offsprings was doubled for the creation of the last generation.

Under the alternative hypothesis the population was randomly split after seven populations of random mating in two sub-populations of 50 males and 50 females each. For this purpose, the number of offspring was temporariliy doubled in generation seven. From generation 8 to generation 16, random mating was conducted within these two sub-populations.

In the considered chromosome segments, crossing over events were assumed to follow a Poisson distribution without genetic interference, thus Haldane's mapping function (Haldane, 1919) was applied. For the distribution of the family sizes Poisson distribution was assumed. Under both hypotheses the offsprings of the last generation were simulated as full-sib pairs, this full-sib structure was used for the reconstruction of haplotypes.

Under the null (alternative) hypothesis, a total of 1700 (2500) individuals was generated in one replicate. For these animals, the full pedigree and the simulated genotypes were stored.

For each assumed scenario, 1000 replicates were generated and analysed. To compute the empirical threshold value, the five and one percentile of the test statistic was calculated from the results of the simulation under the null hypothesis. The empirical power then was estimated by determining the proportion of replicates exceeding these empirical thresholds under the alternative hypothesis.

Scenarios studied

For the expected test statistic, $E(X^2)$ was calculated using eq. [5], based on the average epistatic kinship within and between sub-populations. Since this quantity is totally independent of the genotypes, it is only necessary to assume a chromosome segment length x, for which the values x = 0; 0,05; 0,10; 0,15; 0,20 were considered. Note that the results for x = 0 reflect the outcome using the classical single-locus kinship as introduced by Malécot (1948).

For the marker-based estimation of epistatic kinship with the test statistics S_t and S_r a fixed set of 6 equidistant markers per chromosome segment were used, where for simplicity all markers had the same number of alleles, and each allele had the same probability to be drawn in the formation of the base population.

The following quantities were varied:

- The number of alleles per marker was set to N_a = 2, 4, and 6, where N_a = 2 reflects the situation with SNPs and N_a = 6 is a model for microsatellites;

- The length of a chromosome segment was set to x = 0,01; 0,05; 0,10; 0,15; 0,20.
- The number of chromosome segments was set to N_{seg} = 1, 3, and 6;
- The number of full-sib pairs per sample was set to N_{fsp} = 10, 30, and 50.

Results and Discussion

Epistatic kinship based on pedigree information

In figure 1 the behaviour of the average epistatic kinship is depicted for all generations for the chromosome segment sizes $x = 0$ and $x = 0{,}2$, respectively. From generation 1 to generation 7 the epistatic kinship within the common base population is illustrated. After fission the epistatic kinship between the two subdivided populations is compared with the average epistatic kinship within population 1 and population 2. Figures 1a and 1b show that the empirical results from the simulation (dots) coincided perfectly with the theoretical expectations (lines) from equations [2] and [4].

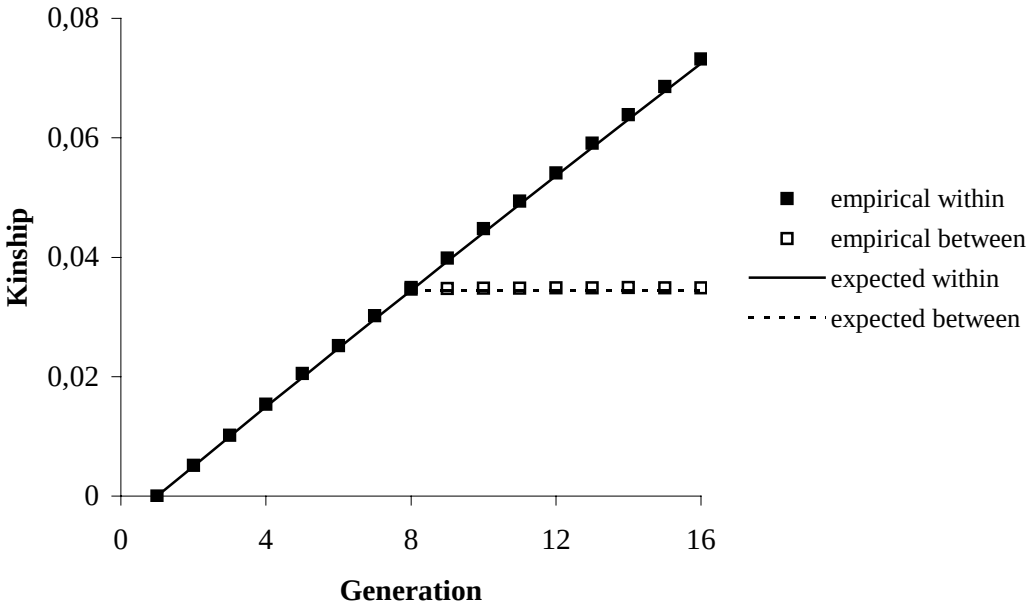


Figure 1a): Empirical and expected average epistatic kinship within and between population; x = 0,00 Morgan.

With $x = 0$ (figure 1a) only one locus is considered and the graph shows the average kinship within and between populations with common origin. The within population average kinship increases linearly with a rate of approximately $1/2N = 1/200 = 0{,}005$ per generation, leading to an average kinship of 0,073 in generation 16. The average kinship between the two sub-population is fixed to the level achieved at the point of fission, i.e. 0,035 in generation 8, and remains constant henceforth.

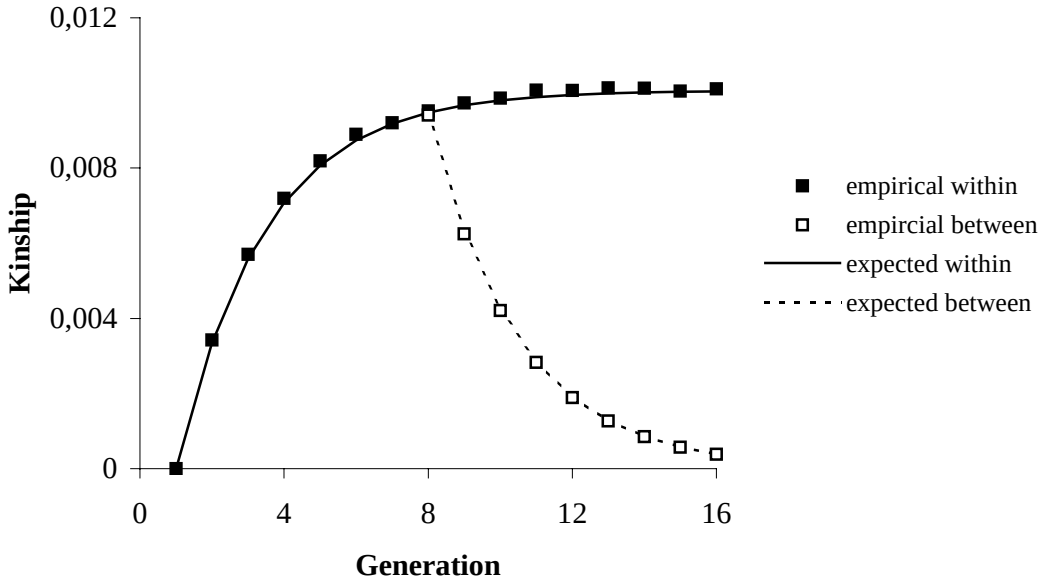


Figure 1b): Empirical and expected average epistatic kinship within and between population; x = 0,20 Morgan.

With $x = 0{,}2$ (figure 1b) the epistatic kinship within population loses the linear behaviour over generations. After generation 9 the increase of kinship within population resulting from coancestry is almost balanced by the loss of *ibd*-status due to crossing over. In generation 16, the expected asymptotic value obtained from eq. [2]

$$\overline{K}^{0,2}_{\max.} = \frac{e^{-0,4}}{e^{-0,4} + 200(1 - e^{-0,4})} = 0{,}010064$$

is achieved to 99,6 per cent.

While after fission in generation 7 the degree of homozygosity between populations remains constant for $x = 0$, it quickly erodes with $x = 0{,}2$ with the rate $e^{-0{,}4} = 0{,}6703$ per generation, so that more than 97 per cent of the expected epistatic kinship present at the time of fission are lost nine generations later.

In the first generations after fission, the difference between expected epistatic kinship within and between populations diverges faster for large chromosome segments compared to short chromosome segments (with the single locus case $x = 0$ as the extreme). However, the suggested test statistic is based on the comparison of expected numbers of *ibd* segments within and between populations. Here, not the ratio, but the absolute difference of observed *ibd* cases is relevant, hence it becomes essential, that the absolute level of *ibd* probabilities is much higher for the single locus case (0,037 at generation 7) compared to the 20 cM case (0,009 at generation 7).

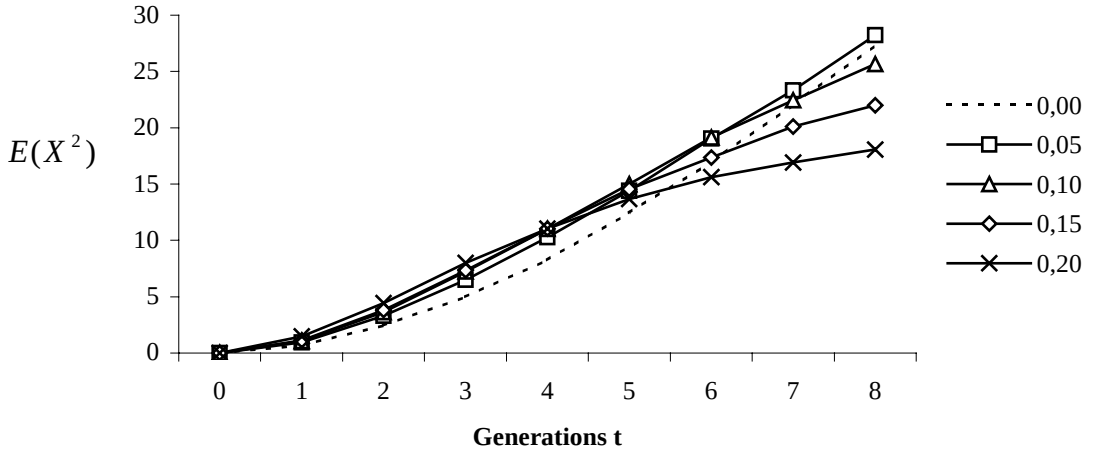


Figure 2: $E(X^2)$ for $x =$ 0,00; 0,05; 0,10; 0,15 and 0,20 Morgan for 1 up to 8 generations after fission.

This difference in the level of the number of cases is reflected in the parameter $E(X^2)$ whose characteristics are depicted in figure 2 for $M = 10$ individuals and $L = 5$ chromosome segments for the five different values for x. The curve for $x = 0$, i.e. considering one locus only, results in each generation with a lower $E(X^2)$ than the curve for some $x > 0$. Further it can be seen that we have different most informative segment lengths for different generations since fission. This is also shown in table 1, where the values of $E(X^2)$ are given for the chosen chromosome segment lengths.

Table 1: $E(X^2)$ for x = 0,00; 0,05; 0,10; 0,15 and 0,20 Morgan for 1 up to 8 generations t after fission.

	Generation t after fission							
x (M)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0,00	0,68	2,46	5,00	8,27	12,46	16,98	22,05	27,33
0,05	0,93	3,31	6,50	10,28	14,42	19,06	**23,31**	**28,22**
0,10	1,06	3,63	7,23	11,03	**15,00**	**19,13**	22,45	25,67
0,15	1,13	3,82	7,31	11,03	14,56	17,35	20,12	22,00
0,20	**1,47**	**4,42**	**8,00**	**11,05**	13,69	15,62	16,91	18,06

For each generation after fission, the highest value is printed in boldface. It is obvious, that in the first generations, the highest value is obtained for larger chromosome segments. With the number of generations increasing, the most informative chromosome segment length decreases. It can be concluded as a general rule, that the closer two populations are expected to be (in terms of generations since fission), the longer the segment length should be chosen. For a large number of generations since divergence, very short segments or, in the extreme, single locus *ibd* status appears to be optimal.

Epistatic kinship based on marker information

The frequencies of the three cases 1, 2, and 3 for the haplotype reconstruction method are depicted in figure 3 for N_a = 2, 4 and 6 alleles per locus and the segment length x from 0,01 up to 0,20 Morgan. Case 1 describes the pairs without a common allele in the genotype of at least one locus, thus the cases where inferring the haplotypes is not possible and the genotyping information can not be used. Case 2 describes the pairs where exactly one common haplotype is possible and case 3 where two or more common haplotypes are possible. The frequency of case 1 is increasing with increasing segment length and to that effect the frequency of case 2 is decreasing. The sum of case 2 and case 3 reflects the frequency of informative comparisons and it is decreasing from 80,5% (for x =0,01) to 71,2% (for x =0,20) with increasing segment length.

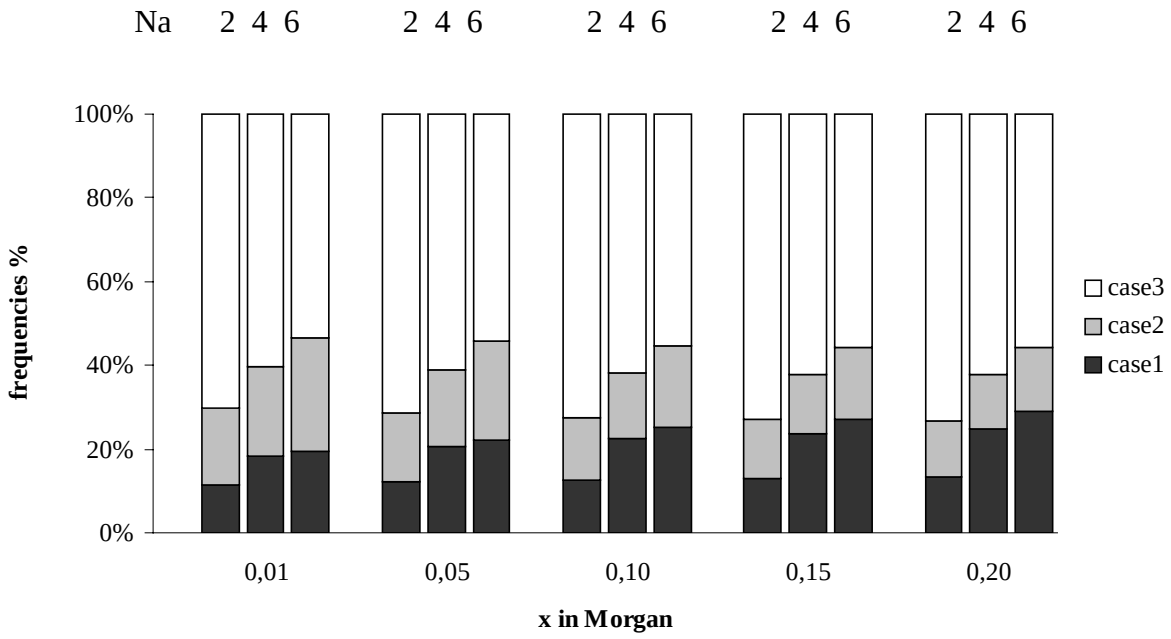


Figure 3: Frequencies for the 3 cases for N_a = 2, 4, 6 and segment length in Morgan.

Case 1 is expected to have a high impact on the efficiency of the haplotype reconstruction method. Again the influence of the segment length becomes obvious. Due to higher probability of recombination events the number of not informative fullsib pairs increases with increasing segment length. Further not informative comparisons increase with increasing number of alleles per locus. For a segment of 0,20 Morgan and N_a=6 the frequency of case 1 is almost 29%.

An overview of the power calculations for all different combinations of x, N_{seg}, N_a, N_{fsp} for true haplotypes are given in table 2a. 118 of totally 135 different combinations simulated result in a power higher 90% (shaded fields in table 2a). This underlines the high potential of the marker based epistatic kinship for short term phylogenetic studies.

Table 2b reports the results for the epistatic kinship based on reconstructed haplotpyes. For reconstructed haplotypes 75 of the totally 135 different combinations simulated yield in a power higher 90% (shaded fields in table 2b). The loss in power between the epistatic kinship with true haplotypes and reconstructed haplotypes is high (up to 57 per cent) for the scenario where only 10 fullsib pairs are genotyped for one segment. Here the power based on reconstructed haplotypes is less than 35% for all segment lengths and for all N_a, thus for this sample size the suggested method has its limitations.

The method for haplotype reconstruction used in this study does not account for linkage disequilibrium (*LD*) in the populations. This leads to a certain loss of information by the estimation of the haplotype frequencies. Excoffier and Slatkin (1995) suggested an EM-algorithm which performed well in the presence of *LD*. A study comparing the efficiency of the epistatic kinship applying the haplotype reconstruction based on the EM-algorithm is in preparation.

The haplotype reconstruction based on fullsib information lacks some generality. For multiparous species such as pig (which we had in mind since this method will be applied in a pig diversity study) it is possible to draw fullsib pairs without pedigree information. For other species (e.g. cattle) this might become a problem. For randomly sampled animals or other simple pedigree structures such as parent-offspring-pairs, the planned implementation of the EM-algorithm is supposed to lead to a general solution.

Table 2a): Power for the different scenarios based on true haplotypes, ▢: power >90%

		N_{fsp}=10			N_{fsp}=30			N_{fsp}=50		
N_{seg}	x	N_a=2	N_a=4	N_a=6	N_a=2	N_a=4	N_a=6	N_a=2	N_a=4	N_a=6
1	0.01	0.657	0.735	0.696	0.990	0.998	0.999	0.999	0.999	1.000
"	0.05	0.605	0.742	0.751	0.987	1.000	1.000	0.999	1.000	1.000
"	0.10	0.603	0.732	0.680	0.989	1.000	1.000	1.000	1.000	1.000
"	0.15	0.550	0.669	0.690	0.985	1.000	1.000	1.000	1.000	1.000
"	0.20	0.512	0.702	0.655	0.990	1.000	0.999	0.994	1.000	1.000
3	0.01	0.916	0.994	0.996	1.000	1.000	1.000	1.000	1.000	1.000
"	0.05	0.939	0.991	0.997	1.000	1.000	1.000	1.000	1.000	1.000
"	0.10	0.935	0.984	0.985	1.000	1.000	1.000	1.000	1.000	1.000
"	0.15	0.883	0.985	0.981	1.000	1.000	1.000	1.000	1.000	1.000
"	0.20	0.835	0.966	0.958	1.000	1.000	1.000	1.000	1.000	1.000
6	0.01	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
"	0.05	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
"	0.10	0.996	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
"	0.15	0.987	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000
"	0.20	0.975	0.995	0.992	1.000	1.000	1.000	1.000	1.000	1.000

Table 2b): Power for the different scenarios based on reconstructed haplotypes, ▢: power >90%

		N_{fsp}=10			N_{fsp}=30			N_{fsp}=50		
N_{seg}	x	N_a=2	N_a=4	N_a=6	N_a=2	N_a=4	N_a=6	N_a=2	N_a=4	N_a=6
1	0.01	0.264	0.228	0.240	0.779	0.826	0.831	0.943	0.957	0.951
"	0.05	0.340	0.220	0.243	0.781	0.779	0.812	0.920	0.976	0.981
"	0.10	0.313	0.165	0.212	0.808	0.795	0.750	0.933	0.980	0.931
"	0.15	0.293	0.170	0.132	0.790	0.794	0.749	0.933	0.976	0.951
"	0.20	0.288	0.166	0.138	0.785	0.782	0.667	0.923	0.975	0.955
3	0.01	0.550	0.550	0.465	0.988	0.991	0.997	1.000	1.000	1.000
"	0.05	0.582	0.447	0.481	0.992	1.000	0.996	1.000	1.000	1.000
"	0.10	0.574	0.451	0.391	0.995	0.999	0.994	1.000	1.000	1.000
"	0.15	0.599	0.358	0.322	0.985	0.998	0.999	1.000	1.000	1.000
"	0.20	0.534	0.433	0.282	0.990	0.993	0.994	0.998	1.000	1.000
6	0.01	0.837	0.731	0.752	1.000	1.000	1.000	1.000	1.000	1.000
"	0.05	0.844	0.702	0.661	1.000	1.000	1.000	1.000	1.000	1.000
"	0.10	0.811	0.713	0.607	1.000	1.000	1.000	1.000	1.000	1.000
"	0.15	0.784	0.710	0.571	1.000	1.000	1.000	1.000	1.000	1.000
"	0.20	0.728	0.707	0.580	1.000	1.000	1.000	1.000	1.000	1.000

Other than in using marker-based estimated kinships (Lynch, 1988) we do not correct for the probability that an identical haplotype may be only identical by state, but not identical by descent. This possibility is neglected, because the probability of such a case is minor. With equal allele frequencies in the base population, the probability that two haplotypes made up from N_{loc} loci with N_a alleles each are identical in the founder population is $N_a^{-N_{loc}}$. Since in our study, the number of loci per haplotype was fixed to $N_{loc} = 6$, this probability varies between $1{,}56 \times 10^{-2}$ for $N_a = 2$ and $2{,}14 \times 10^{-5}$ for $N_a = 6$. Therefore, identity of haplotypes is expected to be almost exclusively due to identity by descent and correction is unnecessary.

Table 2a and 2b highlight, that the power of the marker based epistatic kinship depends on the segment length x in Morgan. The power is decreasing with increasing x. The decrease in power with increasing x is smaller than expected, though, table 2a shows that a power greater 65% is feasible for a single segment of 0.20 Morgan when genotyping highly polymorphic markers. While for true haplotypes, the power reduction is mainly due to a reduced rate of identity by descent due to recombination in the generations between fission and the final generation, the loss of power between true and reconstructed haplotypes is due to failure or disturbance of haplotype reconstruction through crossing over events in the formation of the fullsib pairs.

The lower power for $N_a = 2$ with true haplotypes (figure 2a) underlines the information loss with single nucleotid polymorphisms due to their biallelic nature (Vignal et al., 2002). The loss of power in this case is caused by the high proportion of ambiguous haplotypes. This becomes evident by the fact, that at each locus 50 per cent of the animals are expected to be homozygous for a biallelic SNP, while this rate is only 16.7 per cent with a microsatellite with 6 loci. Since homozygous loci add no information to discriminating between haplotypes, the informativeness of reconstructed haplotypes is minor for biallelic markers due to the low heterozygosity. This confirms the suggested analogy of 1 microsatellite being equivalent to two to three SNPs in linkage studies suggested by Evans and Cardon (2004).

For reconstructed haplotypes the increase from 2 alleles to 4 alleles leads to a loss in power. This loss can be explained with the increase in non-informative comparisons between fullsib pairs when increasing the number of alleles per locus (figure 3). Again, the need of a more powerful method for haplotype reconstruction is highlighted.

Classical distance measures reflect differences between populations which are mainly due to genetic drift and mutation (Oldenbroek, 1999). In our approach, mutation is totally disregarded. Yue et al. (2002) estimated the mutation rate of microsatellites in swine to be $7{,}5\times10^{-5}$ per generation. Using this rate, the probability that a mutation occurs in a haplotype of 6 microsatellite loci over 10 generations is less than 0,5 per cent.

Mutations may occur, though, in the chromosome segments considered. A segment of 0,2 M contains on average 2×10^{7} basepairs. Nachman and Crowell (2000) estimated the human mutation rate to be $2{,}5\times10^{-8}$ per nucleotide and generation. Assuming this value to be valid for mammals in general, the probability is 8 per cent that such a mutation occurs in a 20 cM interval in one generation, and the probability that at least one base change due to a mutation appears in 10 generations is 56.6 per cent and thus non-negligible. However, this mutation will never be detected unless it affects a marker site, which was shown to be highly unlikely above, or if it causes a major reorganisation of the chromosome, e.g. through a translocation, deletion or inversion of a major chromosome segment, which is equally unlikely to appear *de novo* in viable offspring.

The second ‘classical’ driving force of population divergence is genetic drift which of course also operates on chromosome segments. However, in the assumed scenario of a limited number of generations since fission, drift is a much weaker process than crossing over, especially when longer chromosome segments are considered. As shown in eq. [4], crossing over reduces the rate of epistatic kinship between populations in every generation with the rate e^{-2x}, independent of the population size. Disregarding

crossing over, the drift variance of chromosome segment frequency is $Var(p_1) = \frac{p_o(1-p_o)}{2N_e}$ (Falconer and Mackay, 1996), where p_o is the initial frequency of the chromosome segment and p_1 is the frequency in the subsequent generation. To give an example: with $p_o = 0{,}2$ and N_e=100, the frequency of the chromosome segment in the next generation will lie with a 95 per cent probability between $p_1 = 0{,}1446$ and $p_1 = 0{,}25540$, respectively. Drift is an undirected mechanism, which may both increase and decrease the chromosome segment frequency in a population. For a comparison of chromosome segment frequencies between lines, the probability that both frequencies change through drift by, say, more than 10 per cent in the same direction (from $p_o = 0{,}2$ either to $p_1 = 0{,}22$ or $p_1 = 0{,}18$) is only 5,8 per cent. Crossing over strictly reduces the probability of chromosome homozygosity. In the example discussed, we expect a change of epistatic kinship between lines from $p_o = 0{,}2$ to $p_1 = 0{,}18$ already with a chromosome segment length of $x = 0{,}053$. Since this process, other than drift, is independent of effective population size, we expect the epistatic kinship based approaches to have higher sensitivity in cases where the effective size of the populations to compare is high.

As was argued in reference to table 1, the suggested method even allows to ‘adapt’ the sensitivity of the method by choosing the optimal chromosome segment length depending on the (expected) number of generations since divergence, with long (20 cM and more) segments for less than four generations and short (5 cM and less) segments for more than seven generations.

The suggested approach is primarily targeted to the analysis of short-term phylogenies through subdivision of populations. Although this does not necessarily imply that the populations included are small, this will often be the case, leading to a relative small degree of polymorphism due to drift and eventually selection. Based on the results in table 2a and 2b we suggest to overcome this information loss nature by genotyping

multiple segments. The number of segment genotyped N_{seg} has an immediate impact on the efficiency of the approach. Especially genotyping three segments instead of a single segment raises the power distinctively.

Another important factor is the sample size N_{fsp}, i.e. number of fullsib pairs drawn in each population. An increase in the tested animals from 10 to 30 fullsib pairs per population genotyped for one segment with microsatellites leads to doubled power for reconstructed haplotypes. Those findings with marker based epistatic kinship support the linear influence of the number of segments typed and the squared influence of the sample size found when estimating the epistatic kinship with pedigree information.

At this point it is important to make some practical and economic considerations. Consider a case where two populations are compared based on $N_{seg} = 1$ segment of length $x = 0{,}05$ Morgan with six microsatellite markers with $N_a = 6$ alleles based on $N_{fsp} = 10$ fullsib pairs. In this case, the power to statistically prove the difference between the two populations on the 5 per cent error level is 0,751 based on true, but only 0,243 based on reconstructed haplotypes, respectively (table 2a and b). This result can either be improved by typing three instead of one segment, or by considering 30 instead of 10 fullsib pairs. In both cases, the number of necessary genotypings is tripled. While in both cases the power based on true haplotypes increases to >0,99, the power based on reconstructed haplotypes is increased to only 0,481 with $N_{seg} = 3$ chromosome segments, while with $N_{fsp} = 30$ fullsibs it is 0,812. Thus, the alternative to increase the number of fullsib pairs is much more efficient, which again reflects the quadratic effect of sample size.

However, increasing the sample size often has considerable extra cost, especially if samples have to be collected under field conditions. Adding chromosome segments, on the other hand, yields almost no extra cost, given the required markers are established in the lab (remember that the total number of genotypings is identical). The results in table

2a show, that with only 10 fullsib pairs per population and 3 to 6 chromosome segments carrying polymorphic markers, sufficient power to differentiate populations can be achieved. With a further improvement of the haplotype reconstruction algorithm based on Excoffier and Slatkin's (1995) approach, it will be possible to get closer to this results when the analysis is based on reconstructed haplotypes. This demonstrates the potential of the suggested method to develop analytical tools of high sensitivity based on limited samples to be used in phylogenetic studies of domesticated, feral, or wild populations.

References

Andersson, L. 2001. Genetic dissection of phenotypic diversity in farm animals. Nature Genetics 2: 130-138.

Brockmann, G. A., Kratzsch, J., Haley, C. S., Renne, U., Schwerin, M., and Karle, S. 2000. Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F2 variance of growth and obesity in DU6i x DBA/2 mice. Genome Research 10: 1941-1957.

Caballero, A., and Toro, M. A. 2000. Interrelations between effective population size and other pedigree tools for the management of conserved populations. Genetical Research 75: 331-343.

Cockerham, C. C. 1967. Group inbreeding and coancestry. Genetics 56: 89-104.

Coppieters, W., Blott, S., Farnir, F., Grisart, B., Riquet, J., and Georges, M. 1999. From Phenotype to Genotype: Towards Positional Cloning of Quantitative Trait Loci in Livestock? In: From Jay L. Lush to Genomics: Visions for Animal Breeding and Genetics, Iowa State University

Doerge, R. W., and Churchill, G. A. 1996. Permutation Tests for Multiple Loci Affecting a Quantitative Character. Genetics 142: 285-294.

Eding, H., and Meuwissen, T. H. E. 2001. Marker based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118: 141-159.

Eding, J. H. 2002. Conservation of genetic resources. Assessing genetic variation using marker estimated kinships, Wageningen Agricultural University, Wageningen.

Emik, L. O., and Terrill, C. R. 1949. Systematic procedures for calculating inbreeding coefficients. Journal of Heredity 40: 51-55.

Evans, D. M., and Cardon, L. 2004. Guidelines for genotyping in genomwide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. American Journal of Human Genetics 75: 687-692.

Excoffier, L., and Slatkin, M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.

Falconer, D. S., and Mackay, T. F. C. 1996. Introduction to Quantitative Genetics. 4. ed. Longman Group Ltd., Essex.

Farnir, F., Coppieters, W., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., and Georges, M. 2000. Extensive Genome-wide Linkage Disequilibrium in Cattle. Genome Research 10: 220-227.

Flury, C., Täubert, H., and Simianer, H. 2005. Extension of the concept of kinship, relationship and inbreeding to account for linked epistatic complexes. Livestock Science (in press).

Haldane, J. B. S. 1919. The combination of linkage values and the combination of distance between the loci of linkage factors. J. Genet. 8: 299-309.

Hayes, B. J., Visscher, P. M., McPartlan, H., and Goddard, M. E. 2003. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research 13: 635-643.

Lynch, M. 1988. Estimation of relatedness by DNA fingerprinting. Molecular Biological and Evolution 5: 584-599.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie., Paris.

Meuwissen, T. H. E., and Goddard, M. E. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155: 421-430.

Nachmann, M. W., and Crowell, S. L. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.

Nezer, C., Collette, C., Moreau, L., Brouwers, B., Kim, J. J., Giuffra, E., Buys, N., Andersson, L., and Georges, M. 2003. Haplotype Sharing Refines Location of an Imprinted Quantitative Trait Locus With Major Effect on Muscle Mass to a 250-kb Chromosome Segment Containing the Porcine IGF2 Gene. Genetics 165: 277-285.

Niu, T. 2004. Algorithms for Inferring Haplotypes. Genetic Epidemiology 27: 334-247.

Oldenbroek, J. K. 1999. Genebanks and the conservation of farm animals genetic resources. DLO Institute for Animal Science and Health, Lelystad.

Ruane, J. 1999. A critical review of the value of genetic distance studies in conservation of animal genetic resources. Journal of Animal Breeding and Genetics 116: 317-323.

Sambraus, H. H. 2001. Farbatlas der Nutztierrassen. 6 ed. Verlag Eugen Ulmer, Stuttgart.

Scherf, B. D. (Editor), 2000. World watch list for domestic animal diversity. FAO, Rome.

Simianer, H. 1994. Derivation of single locus relationship coefficients conditional on marker information. Theoretical and Applied Genetics 88: 548-556.

Simianer, H. 2002. Vorstudie zum Projekt 'Molekulargenetische Differenzierung verschiedener Rotviehpopulationen'. In: E. u. L. Bundesministerium für Verbraucherschutz (ed.) Molekulargenetische Differenzierung verschiedener Rotviehpopulationen. No. 493. p 7-32. Landwirtschaftsverlag GmbH Münster-Hiltrup, Münster.

Takezaki, N., and Nei, M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144: 389-399.

Toro, M., and Caballero, A. 2004. Characterisation and conservation of genetic diversity between breeds. In: 55th EAAP Annual Meeting, Bled, Slovenia

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetic Selection Evolution 34: 275-305.

Visscher, P. M. 2003. Principles of QTL mapping, manual PhD - course Salzburg, Edinburgh.

Windig, J. J., and Meuwissen, T. H. E. 2004. Rapid haplotype reconstruction in pedigrees with dense marker maps. Journal of Animal Breeding and Genetics 121: 26-39.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56: 330-339.

Yue, G. H., Beeckmann, P., and Geldermann, H. 2002. Mutation rate at swine microsatellite loci. Genetica 114: 113-119.

Appendix

The necessity to account for the number of informative comparisons is illustrated with the following example: Consider individuals A and B in population 1 and C and D in population 2. We find that for one chromosome segment A is *ibd* with both C and D. B is also found to be *ibd* with C, then B has to be *ibd* with D as well. In this case, only three of the four comparisons between populations are in fact informative.

For the number of informative comparisons for a given chromosome segment we derived the following approximations

$$N_w = 2\left[(M-1) + (M^2/2 - 3M/2 + 1)(1-p_w^2)^{(M-2)}\right]$$

$$N_b = M + M(M-1)(1-p_b^3)$$

where

N_w is the number of effective segment-specific pairwise comparisons within populations

N_b is the number of effective segment-specific pairwise comparisons between populations

p_w is the average ibd-probability within populations

p_b is the average ibd-probability between populations.

For p_w and p_b, the corresponding values calculated with recursion [2] and eq. [4] can be used. Note that the proportion of informative segment-specific comparisons within and between populations is inversely proportional to the *ibd* probabilities p_w and p_b, respectively. Note further that for $p_w = p_b = 0$, $N_w = M(M-1)$ and $N_b = M^2$, i.e. the effective number equals the true number of comparisons.

4th CHAPTER

EPISTATIC KINSHIP FOR THREE SUBDIVIDED POPULATIONS OF THE GOETTINGEN MINIPIG

Christine Flury[1], Steffen Weigend[2], Xiangdong Ding[1], Helge Täubert[1,3] and Henner Simianer[1]

[1]Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany
[2]Institute for Animal Breeding, Federal Agricultural Research Centre (FAL), Hoeltystrasse 10, 31535 Neustadt-Mariensee, Germany
[3] Smurfit Institute, Department of Genetics, Trinity College, Dublin 2, Ireland

Submitted for publication

EPISTATIC KINSHIP FOR THREE SUBDIVIDED POPULATIONS OF THE GOETTINGEN MINIPIG

Christine Flury[1], Steffen Weigend[2], Xiang Dong Ding[1], Helge Täubert[1,3] and Henner Simianer[1]

[1]Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany
[2]Institute for Animal Breeding, Federal Agricultural Research Centre (FAL), Hoeltystrasse 10, 31535 Neustadt-Mariensee, Germany
[3] Smurfit Institute, Department of Genetics, Trinity College, Dublin 2, Ireland

Summary

To overcome limitations of some diversity measures applied to livestock breeds marker based estimations of kinship coefficients within and between populations were proposed. This concept was extended from the single locus consideration to chromosomal segments of a given length in Morgan. Algorithms for the derivation of the so called epistatic kinship were published. Further the behaviour of marker based epistatic kinship was investigated theoretically. In the present study the results of the first practical application of this concept are presented. Full sib pairs of three separated sub-populations of the Goettingen minipig were genotyped for six chromosome segments. After haplotype reconstruction the haplotypes were compared and epistatic kinships were estimated within and between populations. A distance measure is proposed which is approximatively linear with the number of generations since fission. The epistatic kinship distances, the respective standard errors and the pedigree-based expected values are presented. As theoretically expected, the level of epistatic kinship is shown to decrease with increasing length of the chromosome segments. Even tough the marker estimated epistatic kinship reveals variable among segments which leads to high standard errors of the respective distances. Possible reasons for this phenomen are discussed. A pedigree-based approach to correct for identical haplotypes which are not identical by descent is proposed. Further it is presumed, that some of the segments studied are influenced by selection, as several QTLs and candidate genes reported in literature were found in proximity.

Keywords

Genetic diversity, short term phylogeny, kinship, epistatic kinship, ibd-haplotypes

Introduction

Genetic diversity is the variety of alleles and genotypes present in a population. It is required for populations to cope with environmental change and therefore the maintenance of genetic diversity is a primary objective in the management of threatened populations (Frankham et al., 2002). Numerous projects have been conducted in different livestock breeds with the goal to help decision makers to identify genetically unique breeds to be included in conservation activities (Ruane, 1999). In subdivided populations like livestock species total genetic diversity consists of within and between subpopulation diversity. Within population diversity can be described with observed and expected heterozygosities, allelic diversity (i.e. the average number of alleles per locus) and the percentage of polymorphic loci (Kantanen et al., 2000; Toro and Caballero, 2004). Between breed diversity is mostly assessed on the basis of genetic distances, for which allele frequencies are used as basic information. For visualization of the between population diversity dendrogramms or phylogenetic trees are constructed from distance data.

In the last years genetic distances estimated from polymorphic microsatellite markers have been the most popular method for the assessment of the phylogenetic structure in animal genetic resources (Baumung et al., 2004; Toro and Caballero, 2004). Based on a survey Baumung et al. (2004) showed that on average three different genetic distance measures were calculated per diversity project. The most favoured is Nei's standard genetic distance D_s used in 74% of the studies (Nei, 1972) followed by Nei's distance DA (Nei, 1972) used in 51% of the studies and Reynold's genetic distance (Reynolds, 1983) for short term evolution used in 30% of the studies. The high popularity of genetic distance projects are explained by Ruane (1999) with the instinctive appeal and the putative objectiveness of genetic distance values in contrary to the subjective evaluations of cultural values of breeds and their current and future value to mankind. Genetic distances have statistical and biological properties which often are based on assumptions which do not hold for livestock populations. Without the consideration of

those limitations genetic distance values might become missleading and lose the explanatory power for genetic diversity in livestock breeds. The properties and limitations related to the subject of the study are presented in the next section, for more detailed discussion a reference to literature is made (Eding and Laval, 1999; Laval et al., 2002; Nagamine and Higuchi, 2001).

Genetic distances have a base in population genetics, initially they have been developed with species in mind, thus for an evolutionary time span. For the creation of livestock breeds this assumption does not hold, as those have been domesticated and improved by man (Toro and Caballero, 2004). Most of today's breeds go back to the 19th or the beginning of the 20th century and crossbreeding was commonly practised 50 to 100 generations (Sambraus, 2001; Visscher, 2003) ago. Therefore the role of mutation in creating differences is assumed to be small and the often made assumption of no or negligible migration between populations is not applicable.

After the assessment of the uniqueness of different breeds with genetic distances a decision is required. Financial resources are limited for conservation activities and therefore not all breeds can be given the same priority. The question is which breeds lead to the highest future genetic diversity. Weitzman (1992) suggested a method that uses genetic and non genetic information to calculate the current diversity and the expected change in total diversity over a certain time horizon for a group of species (Reist-Marti et al., 2003). The properties of this approach have been evaluated in detail (Reist-Marti et al., 2003; Thaon d'Arnoldi et al., 1998). The Weitzman approach was criticised by several authors (Caballero and Toro, 2002; Eding, 2002; Laval et al., 2002) as it does not consider within population variability. Ignoring within population diversity is not only a drawback of the Weitzman method but of all diversity studies relying on genetic distances only. Hall (2004) mentions two reasons for the conservation of within breed variation, first to retain the capacity to respond to selection and second to prevent animals to become homozygous for harmful alleles. When neglecting the within breed diversity, the increase of genetic distances with increasing levels of inbreeding of populations might lead to the conservation of highly inbred populations (Eding and Meuwissen, 2001). To overcome this problem Eding and Meuwissen (2001) and Toro and Caballero (2002) proposed to evaluate genetic

variability within and between population based on the kinship coefficient. Eding (2002) evaluated marker estimated kinships between and within populations and developed corresponding distances. The driving force for the kinship as measure of genetic diversity is solely random drift. Thus, the short term evolution of livestock breeds is accounted for to some extent. However, drift is inversely proportional to the effective population size (Falconer and Mackay, 1996) so that the diversification of large populations will be slower than the one of small populations. For decision making the authors proposed a core set method based on average kinship coefficients (Bennewitz and Meuwissen, 2005; Eding and Meuwissen, 2003).

In this study the single locus concept of kinship was extended to chromosomal segments of a given length in Morgan units. A similar idea was applied for the estimation of past effective population size by Hayes et al. (2003). For the proposed measure based on segments identical by descent (ibd) called epistatic kinship a force additional to random drift becomes crucial - recombination. Thus it goes one step further, regarding „short“ developing time of small populations. Algorithms were derived for the calculation of the epistatic kinship based on pedigree (Flury et al., 2005a). As pedigree information is often missing for small endangered livestock populations (Ruane, 1999) the epistatic kinship was estimated based on marker information. Those investigations showed the promising potential of the concept for the differentiation of short term phylogenies (Flury et al., 2005b).

The goal of the present study is the evaluation of the epistatic kinship based on data from an existing population. The concept is illustrated with a small diversity study for three subdivided populations of the Goettingen Minipig. The estimates for marker based epistatic kinships within and between the three subpopulations are derived. The expected values for the respective segment lengths are calculated based on pedigree information. Further epistatic kinship distances and the corresponding standard errors are presented.

Material and methods

The Goettingen Minipig was established 1960 at the University of Goettingen for laboratory use. The goal was the development of a small pig as a human model (Glodek and Oldigs, 1981). The founder population (*GE*) was separated in 1992 and an additional population was built up in Denmark (*DK1*). In 1998 the Danish population was split, resulting in the third population *DK2*. Today the three populations *GE*, *DK1* and *DK2* are kept closed under specific pathogen free conditions and without any exchanges between the populations. From the actual stock of the three populations *GE*, *DK1* and *DK2* tissue samples of randomly chosen full sib pairs were taken. An insight in the actual relationships within and between the three populations for the pedigree of the sampled animals is provided in table 1. The diagonal reflects the kinship coefficient within population and the corresponding standard error and the off-diagonals the between population kinship and the standard error.

Table 1: Average kinship coefficients within and between populations and the corresponding standard errors for the animals genotyped from populations *GE*, *DK1* and *DK2*.

	GE	**DK1**	**DK2**
GE	0,172 ± 0,029	0,148 ± 0,005	0,148 ± 0,003
DK1		0,176 ± 0,031	0,159 ± 0,005
DK2			0,178 ± 0,026

From the two porcine genetic maps USDA_MARC_v1 and USDA_MARC_v2 six segments on five different chromosomes were chosen (Rohrer et al., 1996; Rohrer et al., 1994). The segments were defined based on five or six microsatellites. The first criterion for the choice of the markers was the segment length in Morgan. Additionally constant order of the markers on both maps, the heterozygosity and the annealing temperature were considered.

The PCR products were obtained in a total volume of 9µL using *Qiagen HotStarTaq Master Mix* Kit (Qiagen GmbH, Hilden, Germany). Each PCR tube contained 20 ng of genomic DNA, 0.3 µM of each primer, 3mM tetramethylammoniumchloride, and 4µL of master mix containing 1 x reaction buffer, 200µM of each dNTP and 0.4 units Taq

polymerase. The amplification protocol of the *Hot Start PCR* were: 15' 95°C; [1' 94°C; 1' Z°C; 1' 72°C] x 35; 10' 72°C; 4°C. The annealing temperature Z varied from 55°-63°C. DNA fragments amplified were visualized by 8% polyacrylamind gel electrophoresis using a LI-COR automated DNA analyzer (LI-COR GmbH, Bad Homburg, Germany). The allele scoring between gels were standardised using internal DNA standard alleles. Standard alleles were calibrated in size using an commercially available external size ladder (MWG Biotech AG, Ebersberg, Germany). For comparability with other studies, a set of standard alleles is available.

The DNA content was not sufficient for some samples. Furthermore some markers did not amplify during PCR. For marker *SW775* only one allele was present in the populations, thus *SW775* was discarded. Finally 334 genotypes (106 from *GE*, 108 from *DK1* and 120 from *DK2*) for 6 segments and totally 33 microsatellites were available for the statistical analysis. In table 2 the 33 microsatellites defining the 6 segments, the chromosome number, the position and the total segment length in Morgan based on USDA_MARC _v2, the number of alleles found in the three populations and the average number of alleles for the segments are given.

Haplotype Determination

For the estimation of the marker based epistatic kinship haplotypes are relevant. Therefore an efficient method for haplotype reconstruction is needed. Excoffier and Slatkin (1995) used the Expectation Maximization (EM) algorithm (Dempster et al., 1977) for the derivation of haplotypes with several loci and several alleles per locus. The EM-algorithm uses information on linkage disequilibrium and pedigree information is not requested. To full account for the available full sib information an extended version of Excoffiers and Slatkin's EM-algorithm was developed (Ding, X., Zhang, Q., Flury, C. and Simianer H., in preparation). The EM-algorithm may lead to biased haplotype frequencies if markers are not in Hardy-Weinberg-Equilibrium (Excoffier and Slatkin, 1995; Tenesa et al., 2003). Therefore the test for Hardy-Weinberg-Equilibrium (HWE) implemented in ARLEQUIN (version3.0, (Excoffier et al., 2005) was conducted for each marker in the three populations. Finally, haplotype reconstruction was conducted for all 33 markers.

Table 2: Definition of the 6 segments, the microsatellites used, the chromosome number, the position based on USDA_MARC_v2 and the number of alleles found in the three populations *GE*, *DK1* and *DK2*. The segment number, the chromsome number, the respective segment length and the average number of alleles alleles are printed in italics.

Segment	Marker	Chromosome	Position	Length	Alleles
1	*SW970*	1	83,700		5
1	*SW216*	1	82,400		3
1	*SW780*	1	81,000		4
1	*SW962*	1	80,500		3
1	*S0082*	1	79,400		4
1	*SW157*	1	78,700		3
1		*1*		*0,050*	*3,67*
2	*SW1536*	14	47,100		5
2	*SW210*	14	46,300		3
2	*SWR1113*	14	45,200		2
2	*SW288*	14	44,600		4
2	*SW69*	14	41,500		2
2		*14*		*0,056*	*3,20*
3	*SW328*	14	59,300		7
3	*SWR2063*	14	57,900		4
3	*SWR925*	14	56,900		4
3	*SW63*	14	54,200		5
3	*SW342*	14	53,200		3
3	*SWR84*	14	52,600		4
3		*14*		*0,067*	*4,50*
4	*SW304*	7	88,600		5
4	*SW732*	7	85,800		2
4	*SWR2152*	7	85,200		5
4	*SWR1210*	7	82,800		4
4	*SW1122*	7	82,300		3
4	*SW175*	7	81,500		5
4		*7*		*0,071*	*4,00*
5	*SW1823*	6	90,700		5
5	*SW316*	6	89,300		3
5	*SW446*	6	88,100		3
5	*SWR987*	6	86,500		3
5	*SW122*	6	83,300		3
5		*6*		*0,074*	*3,40*
6	*SW139*	3	52,400		4
6	*SWR978*	3	52,900		2
6	*SW1315*	3	55,700		4
6	*S0094*	3	57,800		8
6	*SW1066*	3	60,500		8
6		*3*		*0,081*	*5,20*

Epistatic kinship

For the **m**arker **e**stimated **e**pistatic **k**inship (*MEEK)* between and within populations y and z the haplotypes of each full sib pair were compared with the haplotypes of all other full sib pairs. In the case of common haplotypes the product of the haplotype probabilities was summed up.

In a fullsib pair i, we have j = 2 individuals with k = 2 gametes each in the chromosome segment considered. Suppose in the population are l = 1, ... L different haplotypes for this segment. We denote the probability that gamete k of animal j in fullsibgroup i is identical to haplotype l as P_{ijkl}. Note that $\sum_{l=1}^{L} P_{ijkl} = 1$. To compare fullsib group i with fullsib group i', we sum up all products of haplotype probabilities, i.e.

$$S_{ii'} = \sum_{l=1}^{L}\sum_{j=1}^{2}\sum_{j'=1}^{2}\sum_{k=1}^{2}\sum_{k'=1}^{2} P_{ijkl}P_{i'j'k'l}$$

This statistic can vary between 0 (if all haplotypes with probability > 0 differ between the two fullsib groups) and 16 (if all four individuals are homozygous for the same haplotype).

The marker estimated epistatic kinships are derived for each of the six segments seperately and summed up. Finally the sum is averaged over the number of segments.

Pedigree information for the genotyped animals was available back to 1975. This led to a total pedigree consisting of 2081 animals. With the algorithm proposed for the derivation of the epistatic kinship based on pedigree (Flury et al., 2005a) the expected values for segment length x = 0.01 up to 0.15 Morgan were derived in 1 cM steps. For the **p**edigree **e**stimated **e**pistatic **k**inship the abbreviation *PEEK* is used. The average segment length for the 6 segments based on the 33 markers is $x = 0{,}0665$, thus the corresponding *PEEKs* were derived for this average.

Marker estimated kinships were derived for all 33 microsatellites (*MEK*). For better understanding the differences between the single locus approach, i.e. the kinship coefficient and the epistatic kinship, regressions of the *MEK* values and the *MEEK*

values on the corresponding expected values were calculated. Pairwise comparisons between the genotypes at the 33 marker loci of the 334 genotyped animals were conducted in analogy to Eding and Meuwissen (2001) and average similarity indices were estimated. No correction for alleles being identical by state but not identical by descent was implemented, as the fraction is assumed to be the same in all three populations. The similarity indices found for each pair were compared with the pedigree based expected kinship coefficients for the same individuals resulting in 55611 pairwise comparisons. Secondly pairwise comparisons were condcuted for all 334 animals and the 6 segments and again the expected epistatic kinships for $x = 0{,}0665$ Morgan, i.e. equal the average segment length was derived for the 55611 pairs.

In both approaches, the baseline similarity i.e. the probability of identity by state without identity by descent, can be estimated with the intercept of the linear regression. The intercept of the regression of the *MEEKs* on the *PEEKs* of each segment separately therefore is proposed as correction factor for the probability of identical haplotypes which are not identical by descent. Hence the substraction of the intercept from each element of the *MEEK*-matrix for the segment under consideration is proposed as corrected marker estimated epistatic kinship, indicated by *MEEK_corr*.

Genetic Distances

Eding and Meuwissen (2001) suggested the following distance between two populations i and j based on kinship coefficients

$$D_{ij} = f_{ii} + f_{jj} - 2 f_{ij} \qquad [2]$$

where: D_{ij} = the kinship distance between populations i and j.

f_{ii} = the average kinship coefficient within population i.

f_{jj} = the average kinship coefficient within population j.

f_{ij} = the average kinship coefficient between population i and j.

The average kinship coefficient between the two populations stays constant after population fission, thus the distance between the two populations is determined by the increase of within population kinship.

In the case of epistatic kinship we suggest a different distance metric, which will be shown to be approximately linear with the number of generations since fission under certain conditions.

Consider a population which at the time of fission has the average epistatic kinship K_o^x. This population is split in two subpopulations i and j with effective population size N_i and N_j, respectively. If we assume that fission has taken place in generation t, then the average epistatic kinship both the within subpopulations, denoted as $K_{i(t)}^x$ and $K_{j(t)}^x$, and between subpopulations, denoted as $K_{ij(t)}^x$, are equal to K_o^x.

Flury et al. (2005b) have shown, that for generation $t+1$ the expected average epistatic kinship in a closed population i can be calculated as

$$K_{i(t+1)}^x = e^{-2x}\left[\frac{1}{2N_i} + (1-\frac{1}{2N_i})K_{i(t)}^x\right] \qquad [3]$$

and the expected average epistatic kinship between populations i and j is

$$K_{ij(t+1)}^x = e^{-2x}K_{ij(t)}^x \qquad [4]$$

for generation T after fission the expected epistatic kinship between breeds then is

$$K_{ij(T)}^x = e^{-2xT}K_o^x \qquad [5]$$

A distance measure should be based on the relation of between and within breed epistatic kinship.

Consider the following one

$$d_{ij}^{x} = \frac{K_i^{x} K_j^{x}}{\left(K_{ij}^{x}\right)^2}$$

As was also shown by Flury et al. (2005b) the epistatic diversity in a closed population for $t \to \infty$ asymptotically approaches an equilibrium value

$$K_{i(\infty)}^{x} = \frac{e^{-2x}}{e^{-2x} + 2N_i(1-e^{-2x})}$$

in which 'new' homozygosity is generated in the same rate as 'old' diversity is destroyed through recombination. It can be shown that this equilibrium value is approached rapidly if the chromosome segment is not too small. Therefore, close to the equilibrium $C = K_i^{x} K_j^{x}$ will remain approximately constant over generations and the change of the diversity is only depending on the kinship between populations. Hence, d_{ij}^{x} approximately is

$$d_{ij}^{x} \approx \frac{C}{(K_{ij}^{x})^2}$$

and, making use of eq. [5], the diversity in generation T after fission approximately is

$$d_{ij(T)}^{x} \approx \frac{C}{(e^{-2xT} K_o^{x})^2} = \frac{C}{e^{-4xT}(K_o^{x})^2}$$

Taking the natural logarithm of this diversity, we get

$$\ln(d_{ij(T)}^{x}) \approx \ln(C) - \ln(e^{-4xT}) - \ln(K_o^{x})^2 = \ln(C) - 2\ln(K_o^{x}) + 4x \times T$$

This shows that the natural logarithm of d_{ij}^{x} is an approximately linear function of the number of generations since fission, with slope $4x$. Therefore, we suggest to use the diversity

$$D_{ij}^{x} = 2\ln(d_{ij}^{x}) = \ln(K_i^{x}) + \ln(K_j^{x}) - 2\ln(K_{ij}^{x}) \qquad [6]$$

which has the value 0 at the time of fission and increases approximately linear with slope $4x$ per generation.

To assess the expected distances, $E(D_{ij}^{x})$, based on the pedigree information PEEK values were used in eq. [6]. For marker based distance estimates, $\hat{D}_{ij}^{x}$, MEEK values were put in eq. [6].

The variance for the *MEEK* distances was estimated with the following formula.

$$Var(\hat{D}_{ij}^{x}) = Var(\ln(\hat{K}_{i}^{x})) + Var(\ln(\hat{K}_{j}^{x})) + 4 \times Var(\ln(\hat{K}_{ij}^{x})) + 2 \times Cov(\ln(\hat{K}_{i}^{x}), \ln(\hat{K}_{j}^{x}))$$
$$- 4 \times Cov(\ln(\hat{K}_{i}^{x}), \ln(\hat{K}_{ij}^{x})) - 4 \times Cov(\ln(\hat{K}_{j}^{x}), \ln(\hat{K}_{ij}^{x}))$$

The required variances and covariances were calculated based on the obtained epistatic kinships within and between populations. The square root of the variance was taken as the standard error of the *MEEK* -distances. Again, the distances and the respective standard errors were calculated for the two scenarios 1) and 2) separately.

Results and Discussion

Table 3 reports the results from HWE-testing for the 33 genotyped markers and the three populations. Markers with significant derivation from HWE (p-values < 0.01) are marked grey. HWE departures in all of the three populations was found for the microsatellites *SWR2063* and *SW1066*. *SW328* and *SWR2152* show a significant excess of homozygotes in populations *DK1* and *DK2*. Additionally, *SW175* is not in HWE in population *DK1* and *SW780*, *SW1536* and *S0094* are not in HWE in population *DK2*.

Excoffier and Slatkin (1995) mentioned that the use of markers which are not in Hardy-Weinberg-Equilibrium might lead to biased haplotype frequencies when applying the EM-algorithm. In contrary to this Tenesa et al. (2003) observed that departures from HWE do not lead to a notable degree of bias in the estimates of haplotype frequencies using the EM-algorithm. Neglecting the 8 markers which are not in HWE (table 3), 24% of the initial available marker information would be lost. The decreasing number of markers defining the 6 segments and the decrease in the average number of alleles per locus force the occurance of identical haplotypes, which leads to a lower resolution of the suggested method. Therefore the use of all 33 markers is advised.

Table 3: Observed heterozygosity, expected heterozygosity and the p-value from HWE-test for the 33 microsatellites and the three populations.

	Population GE			Population DK1			Population DK2		
Marker	obs.het	exp.het	p-value	obs.het	exp.het	p-value	obs.het	exp.het	p-value
1 SW970	0,71	0,69	0,6867	0,71	0,67	0,0153	0,78	0,71	0,0217
SW216	0,65	0,62	0,5038	0,55	0,58	0,3311	0,58	0,55	0,1526
SW780	0,69	0,65	0,3229	0,73	0,68	0,0752	0,78	0,71	0,0001
SW962	0,68	0,65	0,8807	0,62	0,60	0,7081	0,59	0,59	0,3431
S0082	0,68	0,66	0,5369	0,70	0,65	0,2169	0,61	0,61	0,1302
SW157	0,69	0,64	0,0245	0,57	0,60	0,2363	0,57	0,60	0,0468
2 SW1536	0,62	0,64	0,6770	0,76	0,74	0,5643	0,73	0,71	0,0094
SW210	0,44	0,42	0,6079	0,66	0,60	0,3373	0,48	0,53	0,1759
SWR1113	0,04	0,05	1,0000	0,12	0,12	1,0000	0,02	0,02	1,0000
SW288	0,59	0,56	0,7975	0,55	0,60	0,1115	0,59	0,49	0,1220
SW69	0,22	0,22	0,6713	0,26	0,25	0,6886	0,16	0,15	1,0000
3 SW328	0,55	0,70	0,0205	0,43	0,69	0,0000	0,58	0,74	0,0003
SWR2063	0,44	0,65	0,0004	0,31	0,62	0,0000	0,42	0,60	0,0000
SWR925	0,42	0,52	0,0415	0,51	0,52	0,2343	0,63	0,63	0,3285
SW63	0,74	0,73	0,1041	0,74	0,77	0,2263	0,76	0,74	0,4642
SW342	0,62	0,66	0,0433	0,57	0,62	0,5303	0,63	0,62	0,6285
SWR84	0,54	0,54	0,1771	0,54	0,62	0,0603	0,71	0,68	0,2237
4 SW304	0,58	0,52	0,2499	0,50	0,46	0,7136	0,63	0,62	0,7820
SW732	0,38	0,32	0,0204	0,23	0,22	0,3503	0,18	0,17	0,5957
SWR2152	0,59	0,61	0,1759	0,56	0,65	0,0000	0,62	0,57	0,0027
SWR1210	0,46	0,49	0,3957	0,56	0,51	0,4663	0,53	0,45	0,1809
SW1122	0,28	0,27	0,4578	0,32	0,31	0,6024	0,05	0,06	1,0000
SW175	0,61	0,59	0,1241	0,69	0,65	0,0022	0,47	0,44	0,8195
5 SW1823	0,62	0,68	0,3043	0,81	0,77	0,9216	0,74	0,73	0,1490
SW316	0,58	0,56	0,6253	0,61	0,58	0,7599	0,45	0,39	0,3394
SW446	0,36	0,36	0,7591	0,50	0,53	0,4530	0,33	0,30	0,6469
SWR987	0,53	0,50	0,2636	0,54	0,57	0,5197	0,52	0,53	0,4209
SW122	0,58	0,51	0,4174	0,47	0,46	0,8333	0,48	0,50	0,0752
6 SW139	0,61	0,60	0,9008	0,56	0,64	0,4054	0,65	0,62	0,8457
SWR978	0,31	0,28	0,2984	0,18	0,18	1,0000	0,23	0,24	1,0000
SW1315	0,75	0,73	0,7559	0,67	0,68	0,2843	0,66	0,75	0,1537
S0094	0,76	0,69	0,3516	0,64	0,69	0,0197	0,59	0,69	0,0020
SW1066	0,64	0,63	0,0057	0,63	0,69	0,0000	0,64	0,67	0,0000

p< 0.01

The relation between the single locus focused similarity indices (*MEK*) of the 55611 pairwise comparisons between the 334 genotyped animals and the respective pairwise kinship coefficients based on pedigree information are depicted in figure 1. The estimated linear fit was

$Y = 0{,}35461 + 0{,}56197X$ with $R^2 = 0{,}0291$.

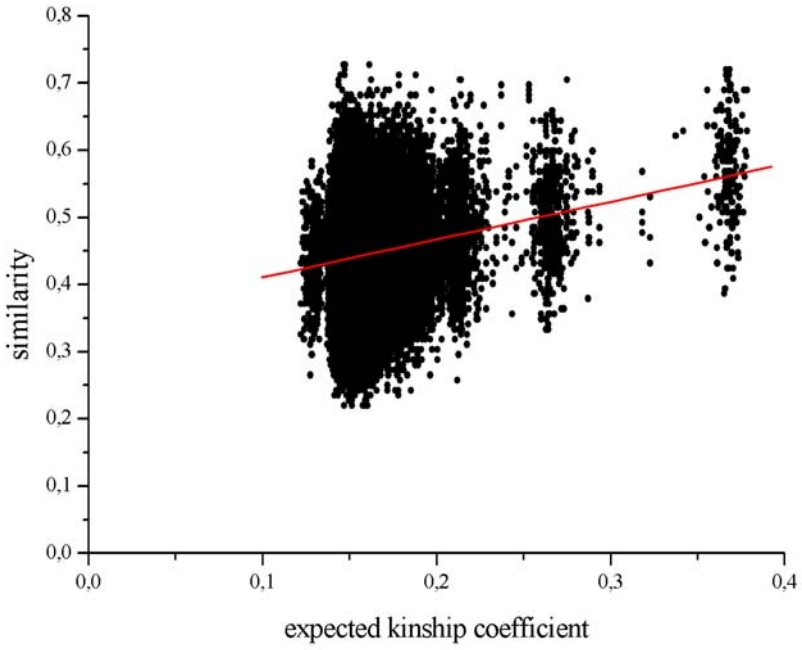


Figure 1: Relation between the average similarity index for the 33 markers and the pedigree based kinship coefficient: 55611 pairwise comparisons between the 334 individuals and the linear regression.

Analogously, figure 2 shows the relationship between the 55611 pairwise comparisons of the *MEEKs* and the *PEEKs*. The estimated linear fit for this regression is

$Y = 0{,}03319 + 0{,}81818X$ with $R^2 = 0{,}0796$.

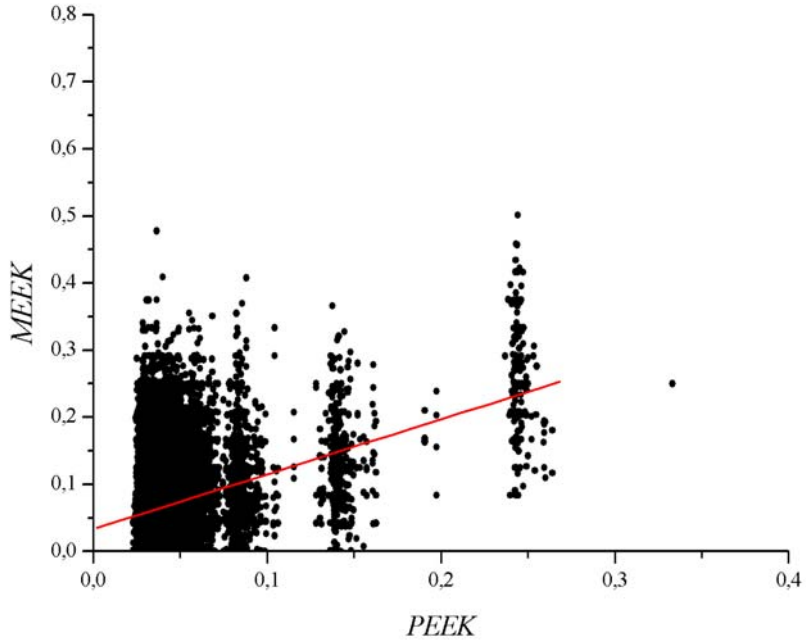


Figure 2: Relation between the marker estimated kinship (*MEEK*) for the 6 segments (i.e. 33 markers) and the expected epistatic kinship (*PEEK*) for $x = 0{,}0665$ Morgan: 55611 pairwise comparisons between the 334 individuals and linear regression.

At this stage it was of interest to compare the regression coefficients of Eding and Meuwissen (2001) single locus consideration with the coefficients of the regression of marker based epistatic kinship in figure 2. The stability index for the regression of similarity indices on pairwise kinship coefficients was low (figure 1). The stability index for the pairwise comparisons of *MEEKs* against *PEEKs* presented in figure 2 is low as well, but higher than for the similarity index. The intercept for the single locus approach results at 0,35 (figure 1) which is much higher than 0,03 for the six segments (figure 2). Since the intercept reflects the probability of loci of segments being identical by state, this quantity can be used to correct for this effect. This underlines, that the bias due to identity by state is much higher for the single locus consideration than for segments.

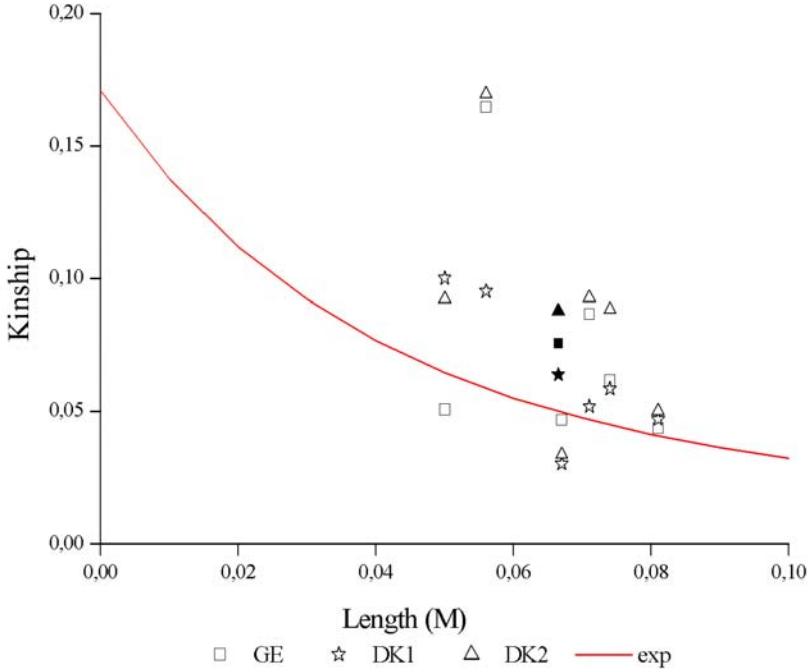
The history of development of the three populations is rather short. After bottleneck the extension of the population size was important as there was an increasing demand for miniature pigs on the market for laboratory animals. Thus it is assumed that populations

were more influenced by increased inbreeding than by drift. Under such circumstances the role of drift in creating differences is small and solely drift based methods like marker estimated single locus kinship are assumed to fail to investigate the differences between populations properly.

In figure 3 the average marker estimated epistatic kinship for the six different segments and their average (at segment length = 0,0665 Morgan) is presented. The marker estimated epistatic kinship within the three populations *GE*, *DK1* and *DK2* are depicted in 3a) and the marker estimated kinship between the three populations in 3b), respectively. The line reflects the averaged *PEEK*, i.e. the expected values based on pedigree information averaged over the three populations. Based on the close relatedness between the three populations, the expected values for the within and the between population *PEEKs* were very similar thus the averaged *PEEK* - value is given as single curve in figure 3 a) and 3 b), respectively.

The results for the marker estimated epistatic kinships are variable. With decreasing segment length the epistatic kinship is supposed to increase due to higher probability of identical haplotypes. This expectation is confirmed with the trend of increasing marker estimated epistatic kinship with decreasing segment length x in figures 3a) and 3b). Even tough, an upward bias of the average marker based estimation in comparison with the pedigree based expectation was observed. The second and fourth segment heavily deviate from the expected values at the corresponding segment lengths (i.e. *PEEK* at 0,056 and 0,071) within as well as between populations.

a) within popoulations



b) between popoulations

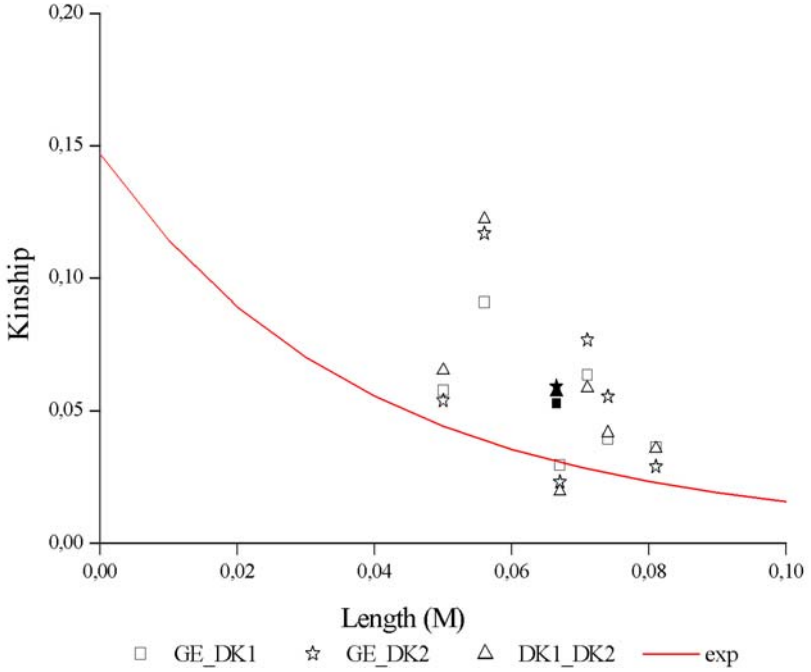


Figure 3: Marker estimated epistatic kinship for the six segments and their average (■★▲) and the course of the expected values (—) within (a) and between (b) the three populations.

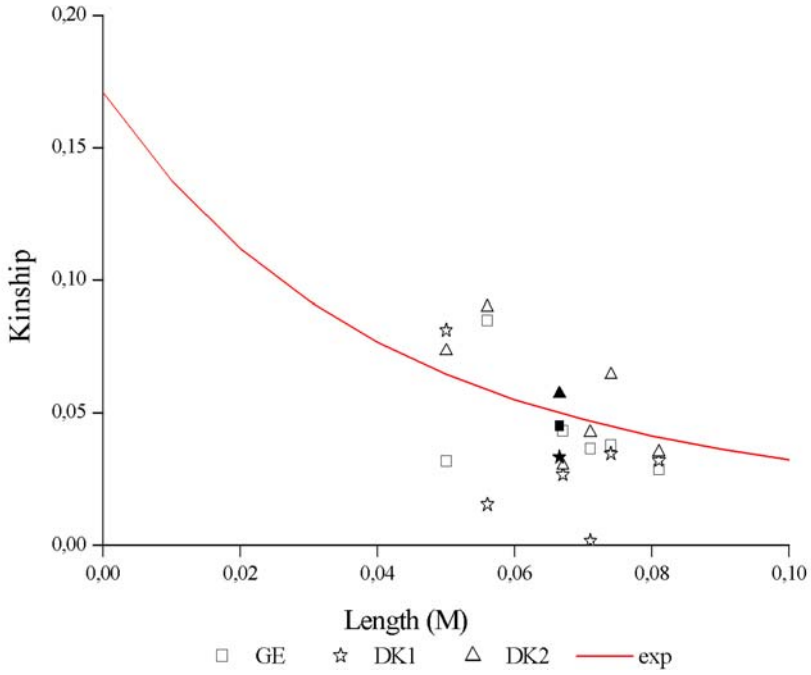
The intercept of 0,03 in figure 2 already suggests a certain overestimation applying marker based epistatic kinship due to segments being identical by state. For further quantification, regressions of the *MEEKs* on the corresponding *PEEKs* were derived for the six segments separately. The corresponding intercepts, the slopes and the stability indices of the regressions and their average are given in table 4. Again the results from segment 2 and 4 are eye-catching due to high intercepts and low stability indices.

Table 4: Intercept, slope and stability index for the linear fit of the regressions from *MEEK* on *PEEK* for the six segments separately and their average.

	Intercept	**Slope**	**R^2**
Segment 1	0,019	0,940	0,017
Segment 2	0,080	0,940	0,011
Segment 3	0,004	0,675	0,025
Segment 4	0,050	0,600	0,007
Segment 5	0,024	0,942	0,026
Segment 6	0,015	0,819	0,023
all	0,033	0,818	0,080

For the comparison of identical haplotypes a certain fraction of conformity arises due to identical haplotypes which are not identical by descent. This can occur due to not unique founder haplotypes or due to recombination randomly resulting in an already existing haplotype. Assuming that all identical haplotypes found are identical by descent the intercept of the regressions should be zero. Based on this assumption the intercepts for the 6 segments seperately (listed in table 4) are applied as correction factors for identical haplotypes which are not identical by descent, named as *MEEK_corr*.

a) within popoulations



b) between popoulations

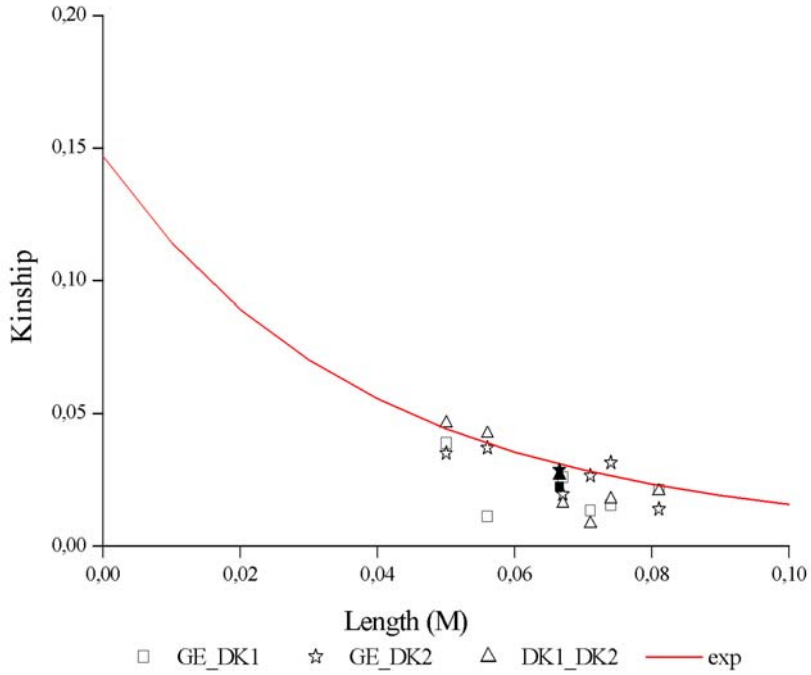


Figure 4: Corrected marker estimated epistatic kinship (*MEEK_corr*) for the six segments and their average (■★▲) and the course of the expected values (—) within (a) and between (b) the three populations.

The results of the corrected marker epistatic kinship *(MEEK_corr)* within and between populations for each segment and their average are depicted in figures 4 a) and b), respectively. The figures show that variability between segments is reduced without loosing the expected trend of increasing marker estimated epistatic kinship with decreasing segment length. Therefore the use of the intercept seems an efficient correction factor.

In table 5a) the elements of the *PEEK*-matrix are listed for the average segment length at 0,0665 Morgan, where the diagonal reflects the within population kinship for each of the three populations and the off-diagonals the corresponding between population kinships. The respective elements of the uncorrected *MEEK*-matrix and their standard errors are given in table 5b). Analogously, in table 5c) the elements of the corrected marker estimated kinship matrix (*MEEK_corr*-matrix) and their standard errors are given. Comparing the standard errors of the elements of the uncorrected *MEEK* matrix in 5b) with the standard errors of the *MEEK_corr* matrix in 5c) further indicates the higher accuracy of the corrected marker estimated epistatic kinships.

Table 5: Epistatic kinship matrices based on pedigree information *(PEEK)* in a) and based on marker information for all 33 markers *(MEEK)* in b) with the corresponding standard errors and for the corrected (*MEEK_corr*) in c) respectively, for the average segment length $x = 0.0665$.

a) *PEEK*

	GE	**DK1**	**DK2**
GE	0,050	0,030	0,029
DK1		0,052	0,035
DK2			0,049

b) *MEEK*

	GE	**DK1**	**DK2**
GE	0,076 ± 0,019	0,053 ± 0,009	0,059 ± 0,014
DK1		0,064 ± 0,011	0,057 ± 0,015
DK2			0,088 ± 0,019

c) *MEEK_corr*

	GE	**DK1**	**DK2**
GE	0,044± 0,008	0,021± 0,004	0,027 ± 0,004
DK1		0,032 ± 0,011	0,025 ± 0,006
DK2			0,056 ± 0,009

The corresponding distance matrices to 5a), b) and c) in b) and c) are given in tables 6a), b) and c), again with the standard errors in 6 b) and 6c). Based on pedigree information (*PEEK)* the two danish populations *DK1* and *DK2* are less distinct than *DK1* with *GE*, respectively, *DK2* with *GE*. The same order was found for the distances based on marker information (*MEEK and MEEK_corr).* This underlines the promising potential of the epistatic kinship as measure for genetic diversity. Overestimation of the marker based epistatic kinship (table 5b) leads to distances at a lower level (table 6b). Correction for identity by state without ibd removes this bias to a larger extent and leads to distance estimates with a slight upward bias compared to the pedigree based expectations, but well within the expected range (table 6c).

Table 6: Distances for *PEEK* in a) *MEEK* and its standard errors in b) and the corrected *MEEK_corr* and its standard errors in c) for the average segment length $x = 0.0665$.

a) *PEEK*

	DK1	**DK2**
GE	0,997	1,051
DK1	0	0,717

b) *MEEK*

	DK1	**DK2**
GE	0,547 ± 0,201	0,640 ± 0,260
DK1	0	0,540 ± 0,278

c) *MEEK_corr*

	DK1	**DK2**
GE	1,157 ± 0,516	1,190 ± 0,231
DK1	0	1,029 ± 0,437

In theoretical investigations it was shown that the number of alleles per segment influences the power for the distinction between populations with the marker estimated kinship (Flury et al., 2005b). With decreasing number of alleles per locus the probability for identical haplotypes is increasing for the same average coancestry between individuals. For the microsatellites defining the segments two and four, on average only 3,20 and 4,00 alleles were found in the three populations. Thus the low heterozygosity of the markers seems a possible explanation for the high deviation from

the pedigree based epistatic kinship and the marker estimated epistatic kinship within and between population especially in the extreme case of segment 2. The influence of the low heterozyosity is supported by the results of the corrected *MEEKs* in figure 4a) and 4b), as the overestimation decreases implementing the correction factor for identical haplotypes which are not identical by descent.

Theoretical investigations yielded a high power for the distinction between populations with the marker estimated kinship under varying number of segments, number of full sib pairs genotyped and number of alleles per (Flury et al., 2005b). However, neutrality of the segments was assumed and therefore selection was not accounted for. In a QTL study of a Meishan x Goettingen minipig cross Wada et al. (2000) found QTLs for vertebra number and birth weight on chromosome 1, for teat number on chromsomes 1 and 7 and for backfat thickness on chromosome 7. For further investigation of the QTL on vertebrae number F2 families of different Asian, Europe and miniature pig breeds were produced (Mikawa et al., 2005). In this study the QTL on chromosome 1 was confirmed and an additional QTL for the same trait was found on chromosome 7 in 6 families but not in the Meishan x Goettingen minipig family. In Rothschild and Plastow's (1999) review on the recent discoveries of gene mapping in commercial pig, QTLs for growth rate, immune response and the candidate gene of the ESR (Estrogen Receptor) are reported on chromosome 1. The authors mention the associations between several traits and the pig major histocompatibility complex on chromosome 7.

Those findings suggest that the markers used for the definition of segment one and four (on chromosome 1 and 7, respectively) might be influenced by selection. The main focus of selection in the three Goettingen minipig populations was set on decreasing body weight by keeping litter size at an acceptable level. The actual mean of piglets born alive is 5,68 ± 2,32 (N=140) and 35,49 ± 9,05 (N=85) for the 345- to 385- day weight in population *GE*. The deviations of piglets born alive and body weight in comparison with commercial pigs indicate the high selection pressure in the Goettingen minipig populations. This might also be a force for the fraction of markers deviating from HWE.

Therefore the knowledge of QTLs and candidate genes should be considered by the choice of the segments, even though at the actual state it might be a problem to define 6 segments with 5 to 6 microsatellites spanning a region of less than 0,10 Morgan which is selectively neutral. The aspect of selective neutrality for the choice of the segments is further ambivalent as selection can be an important force for the conservation of genomic regions, on which the epistatic kinship relies. The effect of selection on LD between linked loci was investigated by Nsengiama et al. (2004) in five populations of commercial pigs for regions of the two porcine chromosomes 4 and 7 where QTLs affecting growth rate and fat deposition had been reported to be located. The effect of selection was not discarded by the authors, even though with a p-value of 0,06 no significance could be found.

The lengths of the six different segments were not calculated based on own data, but they were taken from the existing porcine map USDA_MARC_v2. The position and the order of the markers seems robust for the six segments. Thus, a change of the segment length with significant influence on the *MEEK*s is not expected.

Conclusions

The results of this study empirically confirm some properties of the suggested kinship and diversity measures (Flury et al., 2005b), but at the same time illustrates some aspects which need to be further studied and discussed. The hypothesis that the epistatic kinship is decreasing with increasing chromosome segment size is clearly confirmed (figures 3a, 3b, 4a and 4b). This allows to adapt the molecular tool, i.e. the length of chromosome segments genotype, flexibly to the phylogenetic structure studied.

The results also show that the problem of chromosome segments being identical by state but not identical by descent is much less relevant compared to single locus approaches (Eding and Meuwissen, 2001), but is not negligible. The suggested correction based on the linear regression of the pedigree based epistatic kinship on the marker based epistatic kinship works well in the example studied, but depends on the availability of

pedigree information. Corrections that can be used in situations where less information is available need to be developed.

As was shown in another study (Ding et al., 2005) sampling incomplete nuclear families is more informative than sampling the same number of unrelated individuals, both with respect to the estimation of haplotype frequencies and to individual haplotype reconstruction. The type of families sampled, however, is depending on the species. While in multiparous species as in pigs, sampling fullsibs is appropriate, it will be more practical to sample e.g. mother – offspring pairs in species like cattle or small ruminants, especially in field studies.

The suggested diversity to our knowledge is the first such measure which was especially designed to study short term phylogenies, and which is not using genetic drift and mutation, but recombination as the major force creating population differences. This will be especially useful, when SNP genotyping platforms will provide massive data on many chromosome segments spread across the entire genome. We expect that the method proposed here has a considerable potential to develop a better understanding of short-term phylogenetic structures in farm animal populations.

Acknowledgements

We gratefully acknowledge Prof. Gary Rohrer for the helpful comments on mapping positions and the Deutsche Forschungsgemeinschaft (DFG) for the financial support.

References

Baumung, R., Simianer, H., and Hoffmann, I. 2004. Genetic diversity studies in farm animals - a survey. Journal of Animal Breeding and Genetics 121: 361-373.

Bennewitz, J., and Meuwissen, T. 2005. A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance. Genetics Selection Evolution 37: 315-337.

Caballero, A., and Toro, M. 2002. Analysis of genetic diversity for the management of conserved subdivided populations. Conservation Genetics 3: 289-299.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society 39: 1-38.

Ding, X., Zhang, Q., Flury, C., and Simianer, H. 2005. A new method for haplotype inference including full sib information. Genetics, submitted.

Eding, H., and Meuwissen, T. H. E. 2001. Marker based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118: 141-159.

Eding, J. H. 2002. Conservation of genetic resources. Assessing genetic variation using marker estimated kinships, Wageningen Agricultural University, Wageningen.

Eding, J. H., and Laval, G. 1999. Measuring the genetic uniqueness in livestock. In: J. K. Oldenbroek (ed.) Genebanks and the conservation of farm animals genetic resources. p 33-58. DLO Institute for Animal Science and Health, Lelystad.

Eding, J. H., and Meuwissen, T. H. E. 2003. Linear methods to estimate kinships from genetic marker data for the construction of core sets in genetic conservation schemes. Journal of Animal Breeding and Genetics 120: 289-302.

Excoffier, L., Laval, G., and Schneider, S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online (submitted).

Excoffier, L., and Slatkin, M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.

Falconer, D. S., and Mackay, T. F. C. 1996. Introduction to Quantitative Genetics. 4. ed. Longman Group Ltd., Essex.

Flury, C., Täubert, H., and Simianer, H. 2005a. Extension of the concept of kinship, relationship and inbreeding to account for linked epistatic complexes. Livestock Science (in press).

Flury, C., Tietze, M., and Simianer, H. 2005b. Epistatic Kinship a new measure of genetic diversity for short term phylogenetic structures -theoretical investigations. Journal of Animal Breeding and Genetics (in press).

Frankham, R., Ballou, J., and Briscoe, D. A. 2002. Introduction to Conservation Genetics. Cambridge University Press.

Glodek, P., and Oldigs, B. 1981. Das Göttinger Miniaturschwein. Paul Parey, Berlin.

Hall, J. G. 2004. Livestock biodiversity: genetic resources for the farming of the future. Blackwell Science Ltd.

Hayes, B. J., Visscher, P. M., McPartlan, H., and Goddard, M. E. 2003. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research 13: 635-643.

Kantanen, J., Olsaker, I., Holm, L.-E., Lien, S., Vilkki, J., Brusgaard, K., Eythorsdottir, E., Danell, B., and Adalsteinsson, S. 2000. Genetic diversity and population structure of 20 north European cattle breeds. Journal of Heredity: 446-457.

Laval, G., SanCristobal, M., and Chevalet, C. 2002. Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genetics Selection Evolution 34: 481-507.

Mikawa, S., Hayashi, T., Nii, M., Shimanuki, S., Morozumi, T., and Awata, T. 2005. Two quantitative trait loci on Sus scrofa chromosomes 1 and 7 affecting the number of vertebrae. Journal of Animal Science 83: 2247-2254.

Nagamine, Y., and Higuchi, M. 2001. Genetic distance and classification of domestic animals using genetic markers. Journal of Animal Breeding and Genetics 118: 101-109.

Nei, M. 1972. Genetic distance between populations. American Naturalist 106: 283-292.

Nsengimana, J., Baret, P., Haley, C. S., and Visscher, P. M. 2004. Linkage Disequilibrium in the Domesticated Pig. Genetics 166: 1395-1404.

Reist-Marti, S. B., Simianer, H., Gibson, J., Hanotte, O., and Rege, J. E. O. 2003. Weitzman's Approach and Conservation of Breed Diversity: an Application to African Cattle Breeds. Conservation Biology 17: 1299-1311.

Reynolds, J. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767-779.

Rohrer, G. A., Alexander, L. J., Hu, Z., Smith, T. P., Keele, J. W., and Beattie, C. W. 1996. A comprehensive map of the porcine genome. Genome Research 6: 371-391.

Rohrer, G. A., Alexander, L. J., Keele, J. W., Smith, T. P., and Beattie, C. W. 1994. A Microsatellite Linkage Map of the Porcine Genome. Genetics 136: 231-245.

Rothschild, M., and Plastow, G. 1999. Advances in pig genomics and industry applications. AgBiotechNet 1.

Ruane, J. 1999. A critical review of the value of genetic distance studies in conservation of animal genetic resources. Journal of Animal Breeding and Genetics 116: 317-323.

Sambraus, H. H. 2001. Farbatlas der Nutztierrassen. 6 ed. Verlag Eugen Ulmer, Stuttgart.

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L., and Visscher, P. M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. Journal of Animal Science 81: 617-623.

Thaon d'Arnoldi, C., Foulley, J.-L., and Ollivier, L. 1998. An overview of the Weitzman approach to diversity. Genetics Selection Evolution 30: 146-161.

Toro, M., and Caballero, A. 2004. Characterisation and conservation of genetic diversity between breeds. In: 55th EAAP Annual Meeting, Bled, Slovenia

Visscher, P. M. 2003. Principles of QTL mapping, manual PhD - course Salzburg, Edinburgh.

Wada, Y., Akita, T., Awata, T., Furukawa, T., Sugai, N., Ishii, K., Ito, Y., Kobayashi, E., Mikawa, S., Yasue, H., Inage, Y., Kusumoto, H., Matsumoto, T., Miyake, M., Murase, A., Shimanuki, S., Sugiyama, T., Uchida, Y., and Yanai, S. 2000. Quantitative trait loci (QTL) analysis in a Meishan x Goettingen cross population. Animal Genetics 31: 376-384.

Weitzman, M. L. 1992. On diversity. Quarterly Journal of Economics 107: 363-405.

5th CHAPTER

GENERAL DISCUSSION

GENERAL DISCUSSION

Algorithms

In chapter two and three the derivation of the alogrithms for the epistatic kinship, epistatic relationship and epistatic inbreeding was presented. Additionally the extended rules to set up the numerator relationship matrix and its inverse for segments of a predefined length were provided (chapter 2 and 3) and illustrated for a small pedigree (chapter 2).

For the proposed algorithms the Haldane (1919) mapping function was applied, thus poisson distribution was assumed for crossing over events. Therefore the algorithms for the epistatic kinship and the corresponding measures contained of the term e^{-x} reflecting the probability, that a chromosome segment of length x reaches the next generation unrecombined. The main difference between mapping functions is to what extent genetic interference is taken into account (Windemuth et al., 1998) and not so much the probability, that a single crossing over events happens in a short chromosome segment which is not affected by interference. Therefore the given probability is assumed to hold over a variety of mapping functions and the influence of other mapping functions was not further investigated.

There is a straight relation between the new concept and well established measures for relationships between animals, i.e. Malécot's (1948) kinship coefficient and Wright's (1922) relationship coefficient and the relationship within individuals, i.e. the usual inbreeding coefficient. For all algorithms the single locus approach, was found to be a special case of the extended approach as the terms e^{-x} and e^{-2x} equal 1 for $x = 0$. It was shown, that this also holds for the direct set up of the epistatic numerator relationship matrix and it inverse. The special properties of the inverse NRM, like its sparseness and that simple rules can be applied for the derivation of non zero elements, observed by Henderson (1976) and Quaas (1976) are also valid for the inverse of the epistatic NRM.

For the suggested algorithms of the inverse epistatic relationship matrix epistatic inbreeding coefficients were initially derived. Thus the algorithm for the set up of the epistatic inverse is linear with N. The direct set up in one recursive algorithm in analogy to Quaas (1976) would have been possible, even tough the computations then are proportional to N^2. For large pedigrees this might become prohibitive and therefore the derivation in two steps was implemented.

The usual NRM and its inverse are widely used for different tasks in animal breeding and genetics. The analouge characteristics of the epistatic numerator relationship matrix and its inverse enable the use for a variety of tasks as well. In this thesis two possible applications were presented. In chapter 2 the proposed algorithms were applied for the estimation of additive x additive epistatic variance components and in chapters 3 and 4 as a new tool for the assessment of genetic diversity. Those applications in simulation studies (chapter 2 and 3) and on real data (chapter 4) confirmed the ease in the use of the extended algorithms.

Estimation of additive x additive interactions

Non additive effects like dominance effects or epistatic effects are often ignored or treated as nuisance parameters in practical applications of quantitative genetics. Even though, no firm empirical evidence exists, that epistasis is of negligible importance (Lynch and Walsh, 1998). Wright (1968) argued that epistasis is rather the rule than the exception. In chapter 2 the epistatic kinship was used to theoretically investigate additive effects and additive x additive interaction effects of linked loci lying in the same region of predefined length in Morgan units. The extended model was compared and discussed in relation to the pure additive model (i.e. segment length $x = 0$).

The use of the epistatic relationship model in a mixed model has the potential to pick up some of the non-additive genetic components, which are ignored in the pure additive models. So far the possible linkage of interacting genes was ignored by other models accounting for epistatic effects. The epistatic model suggested here accounts for the

entity of unspecified an non localised gene complexes of a given segment length and sums such effects over the whole genome.

However, some limitations for disentangling the additive from the epistatic variance components were found, due to small sample sizes. With small values of x the matrices A and A^x become rather similar. In the mixed additive and epistatic model, the power to estimate additive and epistatic variance components simultaenously depends on the size and structure of the data and the magnitude of the true effect. The proposed algorithms can be applied to much larger data sets which may help to overcome some of those limitations.

The use of the term ‘epistatic’ in the name of the new measure was criticised several times. Considering the implementation of the epistatic kinship as a new measure of genetic diversity the term ‘chromosome segment homozygosity’ proposed by Hayes et al. (2003) for the estimation of past effective population size would have been an alternative. However, the discussion of this section underlines why the term was chosen and is appropriate.

Assessment of genetic diversity

The main focus of this thesis was the development of a new measure for the assessment of genetic diversity in short term phylogenies. Therefore the single locus approach of Eding and Meuwissen’s (2001) average kinship was extended to chromosomal segments and investigated theoretically in chapter 3 and on practical data in chapter 4.

Generally the resolution of the epistatic kinship for the differentiation between populations depends on the number of segments typed, the number of animals sampled and the segment length x in Morgan. With pedigree based epistatic kinship the influence of the number of segments typed was linear and the influence of the number of animals sampled was quadratic. In theoretical investigations of marker based epistatic

kinship a distinct raise in power was found when the number of segments typed was increased from 1 to 3 as well as with an increase of the number of animals sampled from 10 to 30 fullsibpairs. But again, the gain in power was higher for increasing the number of animals sampled than for the increase in the number of segments typed. Therefore the higher influence of the number of animals sampled on the resolution of the average epistatic kinship as tool for the assessment of genetic diversity was determined as a general rule.

The influence of the segment length x in Morgan is the third factor influencing the informativeness of the method. Generally, the ibd probability is decreasing with increasing segment length. For marker based epistatic kinship decreasing power was found for increasing x even tough at a lower level than expected (chapter 3). With increasing segment length the probability of recombination events destroying ancestral segments is increasing. This explains the power reduction for larger segments.

For the single locus approach the average kinship in a closed population is increasing almost linearly. The kinship between populations stays constant after fission at the level of the average within population kinship in the last generation before fission. The behaviour is different considering segments. It was shown, that the average within population kinship loses the linear behaviour as old coancestry is destroyed due to crossing over. The between populations kinship quickly erodes with increasing segment length and number of generations since fission.

Those characteristics led to the result that the informativeness of the segment length depends on on the number of generations since fission. The conclusion was, that the closer two populations are expected to be, the longer the segment length should be chosen. For a large number of generations since fission very short or in the extreme case single loci might be optimal. This allows the adaption of the sensitivity of the method by choosing the optimal chromosome segment length depending on the number of generations since fission.

The first fission in the Goettingen Minipig populations took place in the early nineties, thus about 5 to 6 generations ago (the generation intervall is slightly higher for the german population). Based on the theoretical findings segments in the range of 0,10 to 0,05 Morgan seem optimal. Thus, segments less than 0,10 Morgan were searched based on USDA_MARC_2 (1996). Seven segments fulifilled this request, one of which was omitted due to amplifying problems during PCR. With an average segment length of 0,0665 (± 0,0116) Morgan for the 6 segments finally analysed the target length was reached. However, under some circumstances the acutal density of the available genetic maps might become limiting (i.e. for species like goat, llama and turkey) and the optimal segment length eventually cannot be reached.

To overcome the problem of missing pedigree information between breeds and poor administration within breeds the estimation of the epistatic kinship based on marker information was proposed. The theoretical investigations (chapter 3) and the practical application in chapter 4 revealed, that for the marker estimated epistatic kinship additional aspects like the method of haplotype reconstruction, number of alleles per locus and the influence of selection are relevant. Those aspects are discussed in the following sections.

For the estimation of marker estimated kinship haplotypes are relevant. Therefore a method for haplotype determination was needed for the investigations of chapter 3 and 4. For the theoretical investigations an own method for haplotype reconstruction was derived based on the available fullsib information. As true haplotypes were known, the efficiency of marker based epistatic kinship for true haplotypes could be compared with the efficiency of the marker based epistatic kinship for reconstructed haplotypes.

The developped method for haplotype reconstruction relied on identical alleles at all loci for the fullsib pair under investigation. If at one locus no common allele was found, haplotype reconstruction was not possible for this pair. This information loss highly influenced the power of the marker estimated epistatic kinship. The number of non informative fullsib pairs was highly correlated with the segment length x in Morgan

and the number of alleles per locus. With increasing segment length the number of informative comparisons decreased, due to higher chance of crossing over events during the two meioses for the formation of the fullsib pair. Additionally, the increase from 2 to 4 alleles per locus led to a loss in power. This loss in power also led to an increasing fraction of not informative fullsib pairs due to increased number of alleles per locus. The last aspect contradicts the problem of low heterozygosity of biallelic markers (Vignal et al., 2002) and underlines again that a more powerful method for haplotype reconstruction was needed. The application of an iterative method allows the derivation of haplotypes for all genotypes, thus a more efficient use of the promising potential presented with true haplotypes in chapter 2 would be possible.

For this reason in the practical application of chapter 4 the EM-algorithm of Excoffier and Slatkin (1995) was implemented. With this algorithm haplotype derivation is possible for all animals, thus no genotyping information is lost during haplotype determination. Genotypes for fullsib pairs were available, for the full account of this partial pedigree information, the extended version of the EM-algorithm of Ding et al. (2005) was applied. Excoffier and Slatkin (1995) annotated that Hardy-Weinberg-Equilibrium (HWE) is a prerequisite for the application of the EM-algorithm. However, ignoring the markers which depart from HWE results in a high loss of initially available information, which leads to higher standard errors for marker estimated epistatic kinships. Tenesa et al. (2003) did not find any bias in haplotype frequencies due to markers which are not in HWE. Therefore reconstructed haplotypes for all markers, even those not in HWE, were used for marker estimated epistatic kinship.

The information loss applying the EM-algorithm in comparison with true haplotypes was not quantified in this study, altough this would be highly interesting. Further the influence of different sampling schemes i.e. the comparison of sampling fullsib pairs, parent offspring pairs and random animals was not determined. The information loss during haplotype reconstruction is assumed to be highest for random animals which further influences the standard errors of the marker estimated epistatic kinship. Sampling fullsib pairs is possible for multiparous species like pigs but for other species,

e.g. cattle, or under other circumstances this might not be possible. In most species, sampling dam – offspring pairs will be a realistic option. The investigation of sampling random indiviudals remains of high interest for further improvement of the proposed method.

In this thesis the practical application was tested on three populations. Based on a survey Baumung et al. (2004) found, that on average 18 breeds were investigated within a livestock diversity project. The extension of the marker based epistatic kinship from 3 to a larger set of breeds is straightforward. Even though, computing time for haplotype reconstruction will increase with increasing number of breeds. Ding et al. (2005) compared the efficiency of the EM-algorithm and the haplotype reconstruction method implemented in PHASE (Stephens and Donnelly, 2003; Stephens et al., 2001). The EM-algorithm was found to be consistently more efficient. Thus at the actual state the EM-algorithm seems the optimum method for haplotype reconstruction if no or partial pedigree information is available.

The influence of the number of alleles per locus became evident in chapter 3 for true haplotypes as well as in the practical application in chapter 4. Under low number of alleles per locus, as for example with biallelic markers, the fraction of ambiguous haplotypes is increased. Homozygous loci add no new information for the discrimination between haplotypes, thus the informativeness is less for biallelic markers due to their low heterozygosity. This drawback of biallelic markers is also reported in literature (Evans and Cardon, 2004; Vignal et al., 2002). Further the probability for identical haplotypes which are not identical by descent is increasing with decreasing degree of polymorphism. Those aspects explain some of the variability found between the six segments in chapter 4. Especially for segment 2 the marker estimated epistatic kinship within and between populations is much higher than the expected value. The average number of alleles for the 5 microsatellites used for this segment was 3.20 only. Three of the five markers used for the definition of segment 2 did break the rule of thumb proposed by the FAO – guidelines for genetic distances that loci should have at least 4 different alleles (Hoffmann, 2004). Further it is concluded, that at the actual state

(as long as development and genotyping costs for single nucleotid polymorphisms stay at the actual level) microsatellites are the marker of choice for the proposed method. If the method is used in combination with SNP haplotypes, about 3 to 4 SNP's should replace one microsatellite to achieve the same level of heterozygosity.

A correction factor for identical haplotypes which are not identical by descent is needed. Eding et al. (2002) also found that the unbiased estimation of the kinship coefficient from marker data depends on an accurate correction factor for not unique alleles in the founder population. As mentioned above the importance for such a correction factor is increasing with low numbers of alleles per locus and low numbers of loci used for the definition of a segment (of which the single locus consideration is the extreme case). This was further supported by the comparison of the regression from marker estimated epistatic kinship on pedigree estimated epistatic kinship with the regression from similarity indices on the kinship coefficient (chapter 4). The intercept for the single locus consideration was found at 0,35 where the intercept for the average of the six segments was found at 0,03. This means that 35 percent of the single locus identities were not due to identity by descent, while this was only the case for 3 percent of the chromosome segment identities. Disentangling the fraction of haplotypes/alleles identical by state from the fraction of haplotypes/alleles identical by descent becomes different when no pedigree information is available. Eding and Meuwissen (2001) proposed the use of the similarity index of the pair of populations with the lowest per locus similarity as correction factor. This quantity is supposed to indicate the population similarity just prior to fission. In chapter 3 of this thesis it was shown that this does not hold considering segments, as the epistatic kinship between populations does not stay constant after fission. Therefore this correction is not applicable for the epistatic kinship.

The suggested correction (chapter 4) based on the linear regression of the pedigree based epistatic kinship on the marker based epistatic kinship worked well in this example, but depends on the availability of pedigree information. A general correction which can be estimated without pedigree information is supposed to enable the use of

the marker estimated epistatic kinship under different circumstances. Therefore further investigations on this subject are of high interest.

The effect of selection was ignored in chapter 2, thus selective neutrality was assumed. The results of the practical application suggest, that selection might be an additional force leading to an overestimation of the marker estimated epistatic kinship. Therefore literature on QTL- and candidate gene studies was consulted. Based on those findings it was concluded that for the two segments 1 and 4 used for the practical application selective neutrality was eventually not given, which may be a reason for the excess of epistatic kinship.

The distance measure proposed by Eding and Meuwissen (2001) for the kinship coefficient, relies on the assumption that the average kinship coefficient between populations remains constant after population fission. Thus differences between two populations are mainly determined due to the increase of within population kinship. We propose a distance measure which is based on the relation of between and within breed epistatic kinship. In chapter 3 it was presented, that due to the destroying effect of recombination and due to increase in relationship, old homozygosity and new homozygosity rapidly reach an equilibrium value. Therefore, within population epistatic kinship remains constant over generation and the diversity depends mainly on the between population kinship. Those results were used for the derivation of a diversity measure which has the value 0 at the time of fission and increases approximatively linear with the slope $4x$ per generation.

The distance measures for the three minipig populations were investigated based on pedigree estimated epistatic kinship and for marker based epistatic kinship in chapter 4. The same order of the distance measures was found for the expected distances as well as for the marker based estimations. Further, the order found fully agreed with the documented population history. However, high standard errors were found for the marker based estimations. Here, the variability between the six segments and the corresponding standard error of the average seems an appropriate explanation.

Applying the epistatic kinship to populations with different selection purposes, selection might create high homozygosity within populations and low homozygosity between populations, thus large differences. However in our example the main focus of selection was set on decreasing body weight by keeping litter size constant in all of the three minipig populations. This explains the overestimation of the epistatic kinship within as well as between populations. Assuming the same selection goals the overestimation is assumed to influence the level of the distances, but not the ranking of the breeds.

Based on the investigations of this study and the above discussed aspects some general advises for the application of the marker based epistatic kinship as tool for the assessment of genetic diversity are given. The consideration of the following points is recommended:

- Definition of the optimum segment length

Is there some information on the population history available? Then the optimum segment length can be roughly derived based on the number of generations since fission.

- Choice of the segments

For the choice of the segments reliable maps should be consulted. If different maps are available the comparison of the order between markers contains additional information concerning the reliability of the position. We further recommend to consider the degree of polymorphism in the mapping population and the annealing temperature. Further the consultation of literature might indicate candidate genes or QTLs in the region of a putative segment.

- Haplotype determination

The knowledge of the relationship structure between the sampled animals can be implemented for a more efficient haplotype reconstruction. HWE-testing is advised in any case, as this might contain some information on extreme markers or extreme entire segments.

Analogies with other approaches

Modern breeding programmes with intense selection are expected to cause fixation of larger haplotypes blocks compared with the less intensive animal breeding that was carried out before the twentieth century (Andersson and Georges, 2004). The effect of selection on linkage disequilibrium (LD) between linked loci is termed genetic hitchiking. Thus, genetic hitchhiking describes the situation where a favourable mutation arises and increases to fixation, and with this gives a selective advantage to all genes it was originally associated with (Barton, 2000). This leads to homozygosity at the selected locus and also at the flanking loci.

Identity by descent (ibd) mapping is used to identify the minimum haplotype identical by descent that is shared among the carriers of the mutant allele (Andersson and Georges, 2004; Meuwissen and Goddard, 2000). The method combines the information arising from genetic hitchhiking and recombination to narrow down the location of a QTL. Here a certain analogy to the epistatic kinship becomes evident. Both approaches use the information from conserved genomic regions which go back to a common ancestor and recombination as driving force, however with different goals. For ibd-mapping the minimum ibd-haplotype is of interest where in our method, the fraction of undestroyed ibd-haplotypes of predefined length x in Morgan is used to determine if animals belong to the same population or not.

Ibd-mapping is one method to fine map QTLs. A different approach, using current recombinations was investigated with two methods by Thaller and Hoeschele (2000). The proposed methods are based on recombinant offspring of a known QTL-heterozygous sire (grandsire). The mapping of the QTL to a region of 2-4 cM was feasible. Again recombination is used as driving force, but within very few meioses.

Another similarity exists between the epistatic kinship and the chromosome segment homozygosity proposed by Hayes et al. (2003). The chromosome segment homozygosity and the epistatic kinship both describe the probability that two

chromosome segments of the same size and location drawn at random from a population go back to a common ancestor. Based on this measure the estimation of past effective population size was proposed. In analogy to our results Hayes et al. (2003) found that LD over large genetic distances estimates the effective population size in the more recent past than LD over short genetic distances.

The estimation of the effective population size contains relevant information for conservation activities. It comprehends information about the degree of endangerment of a population, thus one of the seven criteria mentioned by Ruane (1999). As approximation for the actual effective population size long segments should be chosen. Since many of the threatened breeds in developing countries have even not been properly characterised (Ruane, 1999) and no documentation exists, this method is proposed as an efficient and uncomplicated instrument to generate meaningful information.

The analogies with other actually discussed apporoaches underline that the basic idea of the epistatic kinship is on the cutting edge. The method is promising as a tool to assess genetic diversity as well as for other scopes in animal breeding and genetics. The presented results provide interesting aspects of the genetics of closely linked loci. However, genetic variation remains complex. Meanwhile one should not forget practice, every week one to two breeds are lost on the global scale. Thus, there is a high need for immediate activities concerning livestock diversity on the practical level, to ensure that as much (genetic and cultural) diversity as possible survives into the future.

References

Andersson, L., and Georges, M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics 5: 202-212.

Barton, N. H. 2000. Genetic hitchhiking. Phil. Trans. R. Soc. Lond. B 355: 1553-1562.

Baumung, R., Simianer, H., and Hoffmann, I. 2004. Genetic diversity studies in farm animals - a survey. Journal of Animal Breeding and Genetics 121: 361-373.

Ding, X., Zhang, Q., Flury, C., and Simianer, H. 2005. A new method for haplotype inference including full sib information. Genetics, submitted.

Eding, H., and Meuwissen, T. H. E. 2001. Marker based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118: 141-159.

Eding, J. H., Crooijmans, R. P. M. A., Groenen, M. A. M., and Meuwissen, T. H. E. 2002. Assessing the contribution of breeds to genetic diversity in conservation schemes. Genetics Selection Evolution 34: 613-633.

Evans, D. M., and Cardon, L. 2004. Guidelines for genotyping in genomwide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. American Journal of Human Genetics 75: 687-692.

Excoffier, L., and Slatkin, M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.

Haldane, J. B. S. 1919. The combination of linkage values and the combination of distance between the loci of linkage factors. J. Genet. 8: 299-309.

Hayes, B. J., Visscher, P. M., McPartlan, H., and Goddard, M. E. 2003. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research 13: 635-643.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. Biometrics 32: 69-83.

Hoffmann, I. 2004. FAO Guidelines. Proceedings ISAG.

Lynch, M., and Walsh, B. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie., Paris.

Meuwissen, T. H. E., and Goddard, M. E. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155: 421-430.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32: 949.

Rohrer, G. A., Alexander, L. J., Hu, Z., Smith, T. P., Keele, J. W., and Beattie, C. W. 1996. A comprehensive map of the porcine genome. Genome Research 6: 371-391.

Ruane, J. 1999. A critical review of the value of genetic distance studies in conservation of animal genetic resources. Journal of Animal Breeding and Genetics 116: 317-323.

Stephens, M., and Donnelly, P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics 73: 1162-1169.

Stephens, M., Smith, N. J., and Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68: 978-989.

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L., and Visscher, P. M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. Journal of Animal Science 81: 617-623.

Thaller, G., and Hoeschele, I. 2000. Fine-mapping of quantitative trait loci in half-sib families using current recombinations. Genetical Research 76: 87-104.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetic Selection Evolution 34: 275-305.

Windemuth, C., Simianer, H., and Lien, S. 1998. Fitting genetic mapping functions based on sperm typing: Results for three chromosomal segments in cattle. Animal Genetics 29: 425-434.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56: 330-339.

Wright, S. 1968. Evolution and the genetics of populations. Univ. Chicago Press, Chicago.

LEBENSLAUF

Name:	Flury
Vorname:	Christine
Geburtsdatum:	14. Dezember 1975
Heimatorte:	Ammannsegg und Deitingen, Schweiz

Berufliche Tätigkeiten

Seit 10/2002	Doktorandin Institut für Tierzucht und Haustiergenetik, Universität Göttingen, D-37073 Göttingen
11/2001 – 09/2002	Assistentin Landwirtschaftliches Institut des Kantons Freiburg, Grangeneuve, Station Tierproduktion, CH-1725 Posieux

Praktika

08/2000 – 09/2000	Applied Genetics Network, CH-8852 Altendorf
07/1999 – 09/1999	Schweizerische Vereinigung der Ammen- und Mutterkuhhalter SVAMH, CH-5201 Brugg
04/1998 – 10/1998	Landwirtschaftlicher Betrieb der Familie Hockenjos, CH-1607 Palézieux

Studium

1996 – 2001	Studium der Agrarwissenschaften, Fachstudium Tierproduktion, ETH Zürich, CH-8092 Zürich
01/2000 – 05/2000	Austauschsemester, Institute of Animal Science, Wageningen University, N-6700 Wageningen

Schulen

1991 – 1996	Wirtschaftsgymnasium des Kanton Solothurn, CH-4500 Solothurn
1988 – 1991	Bezirksschule, CH-4562 Biberist
1982 – 1988	Primarschule, CH-4573 Lohn

Göttingen, 15. Dezember 2005