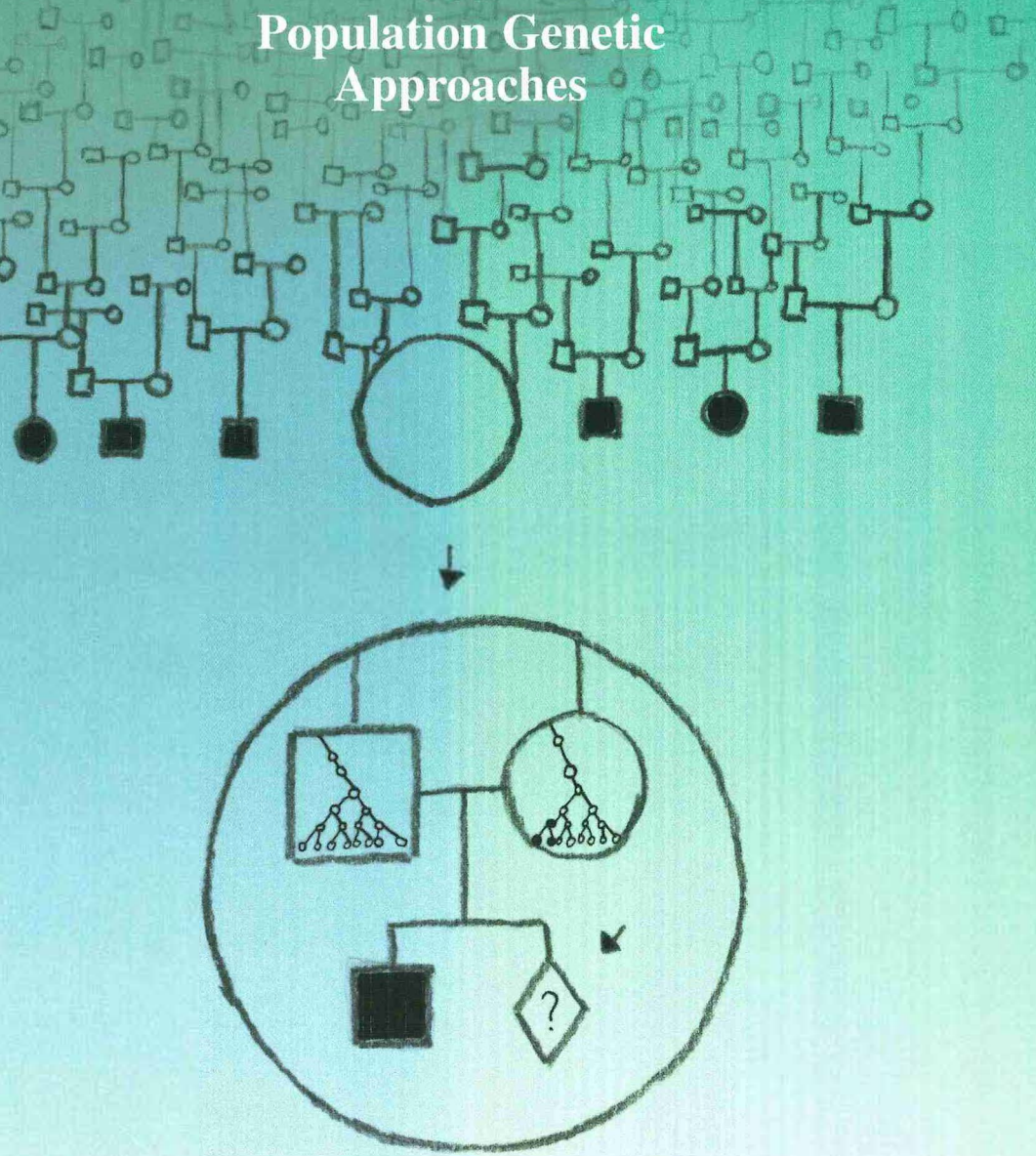
Mutations and Genetic Disease:

Computational and Population Genetic Approaches



Martin A. van der Meulen

Mutations and Genetic Disease:

Computational and Population Genetic Approaches

Stellingen behorende bij het proefschrift van Martin A. van der Meulen
Groningen, 18 december 1996.

1. De mening van Plomin et al. (Science, 17 juni 1994; 1733-39), dat voor het lokaliseren van genen via associatie met een enkele marker een genoomscreen met maximaal 500 kb spacing nodig is, kan zelfs te optimistisch zijn. Er zijn echter betere methoden voor het localiseren van genen (dit proefschrift).

2. Het effect van genetische drift wordt zwaar onderschat (dit proefschrift).

3. De kans op het waarnemen van twee keer dezelfde onafhankelijke mutatie in ziektes met een complexe overerving is niet alleen laag doordat de kans op dezelfde mutatie laag is, maar ook door de lage kans van survival van een mutatie in de bevolking (Fisher, 1930).

4. De door Zhong et al. (Nature Genetics, Nov 1996; 329-334) vermelde hoge sib-pair lodscores zijn hoogst onwaarschijnlijk.

5. De populariteit van computer programma's, zoals Genehunter, voor het analyseren van genetische data lijkt meer gebaseerd op de gerapporteerde lage P waarden dan op de onderliggende 'voor correct aangenomen' statistische methode.

6. Door het krijgen van kinderen op hogere leeftijd, zal in de populatie de vruchtbaarheid op hogere leeftijd toenemen.

7. In de opleiding tot klinisch geneticus krijgt het genetisch redeneren, dat nodig is voor het correct uitvoeren van risico berekeningen, te weinig aandacht.

8. Computerprogramma's voor het uitvoeren van risicoberekeningen zijn niet bedoeld als vervanging van genetisch inzicht, maar als aanvulling.

9. Bewezen dragers van een recessieve ziekte zijn meer geestelijk dan erfelijk belast.

10. Na vele generaties lange selectie op produktiekenmerken bij landbouwhuisdieren is het onwaarschijnlijk dat er nog major QTL's zijn die niet homozygoot aanwezig zijn. Het vinden van deze QTL's is dan ook van weinig nut.

11. Bij de ontsluiting van wetenschappelijke literatuur wordt te weinig aandacht besteed aan het toegankelijk maken en houden van oude kennis.

12. De geringe werkgelegenheid voor gepromoveerden geeft aan dat er eerder een tekort is aan arbeidsplaatsen voor hoger- dan voor lager geschoolden.

13. Het verlagen van de basisbeurs voor studenten levert door verdringing van laag geschoolde arbeid weinig besparing op.

14. Artikelen met nieuwe ideeën hebben een lagere kans op publikatie.

15. In science the credit goes to the man whom convinces the world, not to the man whom the idea first occurs (Sir Francis Darwin).

16. Het Amerikaanse openbaar vervoer lijkt naar Nederlandse begrippen nogal eens op een openluchtmuseum.

17. Het Amerikaanse drugs- en wapenbeleid is met elkaar in strijd.

18. Mathematisch genetici doen alsof.

Rijksuniversiteit Groningen

Mutations and Genetic Disease: Computational and Population Genetic Approaches

Proefschrift

ter verkrijging van het doctoraat
in de Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus Dr. F. van der Woude
in het openbaar te verdedigen op
woensdag 18 december 1996 des namiddags te 4.15 uur

door

Martin Allert van der Meulen

geboren op 6 december 1965
te Eindhoven

Promotor: Prof. Dr. C.H.C.M. Buys

Referent: Dr. Ir. G.J. te Meerman

Aan mijn ouders en Aliki

Promotiecommissie: Prof. Dr. J.H. Edwards
Prof. Dr. Ir. P. Stam
Prof. Dr. W. van Delden

Cover: Misja Sirag
Druk: Grafisch bedrijf Ponsen & Looijen bv, Wageningen

Nugi: 742
ISBN: 90 367 0661 0

This work is performed at the department of Medical Genetics at the University of Groningen in the Netherlands, supported by the Netherlands Organisation for Scientific Research (NWO). Financial support for printing of this thesis by NWO and "Stichting voor Erfelijkheidsvoorlichting te Groningen" is gratefully acknowledged.

Voorwoord

Als eerste wil ik Gerard te Meerman bedanken. Natuurlijk voor het verkrijgen van de subsidie van de Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO), voor het bijeen sprokkelen van de benodigde aanvullende financiën, maar vooral voor de dagelijkse begeleiding. Gerard, ik heb veel van je geleerd en kijk met veel plezier terug op onze samenwerking in de afgelopen vier jaar. Als tweede wil ik Lodewijk Sandkuijl bedanken. Jouw realistische praktische kijk en enthousiasme hebben me erg gestimuleerd. Prof. Charles Buys bedank ik voor de begeleiding vanuit de achtergrond.

De stap van de theoretische mathematische genetica naar het toepassen van genetische technieken is groot. Ondersteuning vanuit het lab kwam met name van Rik de Vries, Hans Scheffer, Robert Hofstra, Gerrit van der Steege en Frank Kooij. Geïnteresseerden in mozaïek berekeningen waren met name Ton van Essen en Rolf Sijmons. Alle medewerkers van de vakgroep Medische Genetica en de stichting erfelijkheidsvoorlichting Groningen wil ik bedanken voor de medewerking en de leerzame en gezellige tijd.

I'm honoured and thankful that Dr. John Edwards agreed to be a member of the thesis committee. Prof. Piet Stam en Prof. W. van Delden wil ik ook bedanken voor het zitting nemen in de leescommissie.

I like to thank Prof. Jurg Ott, Prof. Francoise Clerget-Darpoux, Prof. Bertram Muller-Myhsok, Prof. Jan Vijg and Prof. Edwin Mariman for stimulation and discussion.

Tijdens mijn promotieonderzoek is ook samengewerkt met verschillende andere vakgroepen medische genetica in Nederland en met de bedrijven Ingeny (Leiden), Keygene (Wageningen) en Holland Genetics (Arnhem). Betrokkenen wil ik bedanken voor de prettige samenwerking.

Voor de dagelijkse gezelligheid zorgden achtereenvolgens mijn kamergenoten René, Frans, Jacco en Henriette en natuurlijk de buurvrouwen Corien, Hester en Gita. Voor de broodnodige niet genetische afwisseling tijdens de lunch- en koffiepauzes zorgden Pek van Andel, Jaap Lubbers, Han Wassenaar en vele anderen. Gespreksonderwerpen waren onder andere het oplossen van de "Leidse balpen moord", serendipiteit, leerprocessen en de toestand in de wereld.

Voor de ontspanning buiten het werk hebben velen gezorgd, maar met name wil ik Yep en Misja noemen, de enige mensen in Groningen die ik kende toen ik hier kwam wonen, en Tom van Weert, altijd goed voor een opmonterend praatje. Fantastisch dat velen

geïnteresseerd bleven vragen hoe het met mijn onderzoek was, terwijl men vaak geen idee had waar het nu eigenlijk precies over ging. Lezers van dit voorwoord tijdens de promotieplechtigheid verwijs ik naar de samenvatting voor leken, achterin het proefschrift.

De paranimfen Corien Verschuuren en Yep Zeinstra wil ik bedanken voor het organiseren van de bij de promotie behorende formaliteiten.

Mijn broer Meine wil ik bedanken voor het toepassen van Bayesiaanse zaterdagmiddag statistiek, wat de aanzet was voor het tweede gedeelte van dit proefschrift. Mijn ouders wil ik bedanken voor de motivering en de relativering. Aliki, bedankt voor het kunnen relativeren van mijn wisselende buien en voor jouw kritische kijk, alhoewel die niet altijd door mij werd gewaardeerd. Misja, mijn lieve dochter, je kwam precies op tijd, zodat je vader eerst dit proefschrift af kon maken. Na jouw geboorte was mijn concentratie vaak ver te zoeken.

Kortom, bedankt allemaal.

Contents

Scope of thesis

The role of theoretical geneticists in the explanation of disease status can be seen as a two step process. First the genes involved in the disease have to be mapped and their role in the disease inheritance identified. Second this knowledge can be used for genetic counselling in families.

Mapping genes involved in diseases which show a complex mode of inheritance is very difficult, as many large scale projects failed to locate genes. Part 1 of this thesis describes how genes with 'old' mutations involved in complex disease inheritance may be localized by looking at the increased sharing of multiple marker haplotypes between carriers of the mutation in founder populations in comparison to control haplotypes.

When genes involved in disease inheritance have high mutation frequencies and, as is the case in X-linked recessive lethal diseases and in some autosomal dominant diseases, there is a high selection against mutations, 'new' mutations are common and 'old' mutations are rare. Recurrence/carrier risk calculations in families where the disease is possibly caused by a 'new' mutation are described in part 2 of this thesis.

Part I

IBD mapping, or Searching for shared haplotypes to locate genes involved in (complex) diseases

I-1 Introduction

In many hereditary diseases a simple mendelian mode of inheritance can be recognized. These diseases can be mapped by analyzing segregation of markers through pedigrees. When a marker is located close to the disease locus, and in large genes even when a marker is intragenic, recombinations can be observed. Conversely, a marker is located close to a disease locus, when usually the marker allele is transmitted together with the disease locus. This method to map disease loci is also used to map markers. Ott (1991) gives a complete overview of the history and features of linkage analysis and linkage programs available.

In linkage analysis major pitfalls are recognized:

- heterogeneity: A phenotype is genetically heterogeneous, when it has a genetically different etiology in different individuals. When linkage analysis is performed on a heterogeneous disease, it is possible that one family may show complete linkage with a certain locus, while another family shows random segregation. When large families in which segregation is followed are under study, one family may give sufficient power to prove linkage. Although within large families heterogeneity is unlikely, Van Soest et al. (1994) found a large family with heterogeneity for Retinitis Pigmentosa. Special techniques are available for data analysis of multiple small pedigrees in case of heterogeneity (HOMOG computer program)
- reduced penetrance: The phenotype of an individual does not unambiguously express the genotype of an individual. In other words: an individual is a carrier of the disease causing genotype at a specific locus, but is not affected (non-penetrance). This can be caused by age-dependent penetrance, in that case the individual is at risk, or by multifactorial inheritance, where other genes and/or environmental factors also influence the expression of the disease phenotype.
- phenocopies: An affected phenotype identifies an individual as a carrier of the disease allele (or alleles in case of recessive inheritance). However, in many diseases there are individuals whose affection status is not the result of a genetic predisposition, but is due to other factors: phenocopies or false positives. These other factors can be unspecified environmental factors, e.g. infections or drugs.

The three factors mentioned are the major factors reducing the power of linkage studies. The lodscore is usually maximized over input parameters: the penetrance, phenocopy rate and the mode of inheritance. In linkage studies the number of meiosis observed within families is low, resulting in low power for fine mapping of the disease locus.

Next to linkage studies, other, so called non-parametric, methods for gene mapping are used. Affected sibpair studies (Suarez, 1978, Hodge, 1984, Bisshop and Williamson, 1990, Tierney and McKnight, 1993), the Affected Pedigree Member methods (Weeks et al, 1988, 1992, Risch, 1990, Ward, 1993, Brown et al, 1994) and the Transmission Disequilibrium Test (Spielman et al., 1993, Ewens and Spielman, 1995, Thomson, 1995). These three methods compare genotypes of two or more affected individuals from a pedigree, to find genomic areas which are in excess shared by affecteds. This avoids the problem of non-penetrance, because only affecteds are used, but the problem of heterogeneity and phenocopies are still present. A problem for these methods is that in case of a complex mode of inheritance the availability of pedigrees with multiple affecteds is low. When affected sibpairs are not available some studies focus on sibpairs, where one of the sibs is affected and (multiple) non-affected sibs are used for comparison (Penrose, 1935). The power of this kind of studies is low due to all three factors mentioned above.

Next to diseases which show a simple mendelian mode of inheritance, diseases with a complex mode of inheritance exist. A complex mode of inheritance is defined as a mode of inheritance which is not simple mendelian. This can be caused by one or more of the factors mentioned above or by more complex factors as multiple interacting loci involved in disease expression, etc. Increased relative risk is an indication of the involvement of genetical factors in the etiology of the disorder.

As is the case in autosomal recessive diseases, disorders with a complex mode of inheritance may show a relatively high incidence in certain populations. This is caused by the introduction of a mutation in a founder population followed by drift of this mutation to sizeable numbers. When a severe autosomal recessive disorder, which results in reduced fitness of homozygotes, has a high incidence in a large population, the

explanation must lie in either a very high mutation rate or in a heterozygote advantage (Mueller and Young, 1995). A heterozygote advantage in several recessive disorders is however speculative and hard to distinguish from random drift.

Until recently the great majority of rare conditions listed in McKusicks's *Mendelian inheritance in man* were classified as X-linked recessive. X-linked diseases can be recognized easily (for instance Harper, 1993). Well over 4000 disorders and polymorphic systems showing definite or probable dominant inheritance have been documented (McKusick, 1994, OMIM update Febr. 1996 in Vogel and Motulsky). The greater ease of detection of dominant mutations in comparison to autosomal recessive mutations is especially emphasized in humans, where the recognition of a rare defect as due to a single mendelian factor depends upon genealogical evidence; for the simplest pedigree, such as almost always available, will reveal the character of a dominant effect, while the collation of the statistical evidence of extensive pedigree collections is usually necessary to demonstrate the Mendelian character of a simple recessive. It is a consequence of this difficulty that in humans more dominant defects are known than recessives, although there can be no doubt that the great bulk of human defects, physical and mental, are, as in other animals, autosomal recessive (Fisher, 1930). In contrast to an autosomal dominant disorder, selection against an autosomal recessive disorder will have only a very slow affect. The reason for this is that in autosomal recessive conditions most of the genes in the populations are present in healthy heterozygotes.

Lander and Botstein (1986, 1987) launched a new method to map disease genes in recessive diseases, in which affected individuals are supposed to be homozygous for the disease mutation and therefore for the closely linked markers: homozygosity mapping. This method is based on the assumption that affected individuals will be homozygous for a specific mutation, due to inbreeding, because there is likely only one mutation segregating in the founder population in that geographic area. This method has been extended by Houwen et al. (1995), who extended homozygosity mapping to compare marker genotypes between individuals and over multiple linked markers in haplotypes. They illustrated their method with the mapping of benign recurrent intrahepatic cholestasis (BRIC) in only three patients. Later the same haplotype surrounding the mutation found

in The Netherlands was also found in the United States (personal communication Lodewijk A Sandkuijl).

The method used by Houwen et al. (1995) to map recessive diseases can also be used for mapping of gene loci involved in complex diseases by looking in populations for shared segments of chromosomes between patients, with such a degree of polymorphism that Identity by Descent (IBD) is likely rather than Identity by State (IBS = the same marker haplotype, but not the same segment of a chromosome). Shared segments can be recognized by sharing of the same haplotype of linked markers. Identity by Descent is caused by the introduction of a mutation and the surrounding haplotype by a founder a long time ago, followed by drift and population growth which caused the surviving mutation to be present many times. Affecteds, belonging to this founder population, are then likely to be affected due to the same mutation. Common ancestry of affecteds is likely but not proven. This method of gene mapping has been recognized since long, but the general opinion is that this method can work only with genome screens using marker spacing of 1 cM or less. (e.g. Lander and Schork, 1994, Plomin et al, 1994, Kruglyak and Lander, 1995).

Outline of thesis

In chapter 2 the IBD mapping method is worked out in more detail. The probability of a genomic overlap between haplotypes surrounding a mutation depending on the number of meioses between those haplotypes is calculated. This graph can be used both for the overlap between two haplotypes and for the overlap between all haplotypes surrounding a mutation. IBD mapping uses both the overlap between closely related haplotypes to determine the genomic location roughly and the overlap between all haplotypes for fine mapping of the disease locus. From the same graph in chapter two it can be concluded that the number of meioses needed to get a high probability that the size of the region surrounding the mutation is clipped of to approx. 1 cM equals about 1000. In IBD mapping the number of meioses implicitly observed is high, leading to high power for fine mapping.

Empirical evidence that haplotype sharing between carriers of the same mutation extends over several cM is described in chapter three, where haplotypes of proven carriers of the A455E CF mutation from French Canada and from the Netherlands are compared, within and between countries. The overlap of haplotypes between carriers of the A455E CF mutation within countries was up to 25 cM. Note that with these patients, expressing all a distinct mild type of CF, the CF gene could have been located and fine mapped. In this study the mapping of the CF locus due to sharing around the A455E CF mutation can be seen as a model for mapping a dominant disease, with a reduced penetrance, because the penetrance of the A455E mutation is regulated by the other chromosome. The haplotypes of the other chromosomes show almost no overlap, because different mutations or very old mutations with multiple introductions (e.g. ΔF508) are observed.

In chapter four, the population genetic background of IBD mapping is elaborated. The number of copies of a mutation present in a specific generation of a founder population can be high enough to cause multiple affecteds, even when only a small fraction of the present alleles can be observed in patients. Simulations over 60 generation are used to evaluate the expected number of copies of a disease allele, the total number of meioses connecting all affecteds and the coalescence time for alleles. The total meiotic count is a measure for the length of haplotype sharing between all identical alleles. The coalescence time is a measure for the length of sharing between two haplotypes. In these simulations recombination is also incorporated, so evaluation of the haplotype sharing between haplotypes is possible and the theoretical graph from chapter 2 can be evaluated, with the additional effect of drift.

In populations descending from a relatively small number of founders another complicating factor is observed: Due to random drift other than at the disease locus the patients will also share marker haplotypes more than expected on the basis of linkage equilibrium between markers. This is called haplotype drift-recombination equilibrium, because haplotypes will disappear due to random drift and will be formed because of recombination.

A test statistic is proposed for the evaluation of IBD mapping studies. The power of this method is evaluated in simulations by ranking the real disease locus, between all other possible loci. Ranking is chosen, because the haplotype drift-recombination equilibrium causes inflated lodscores on many loci, when sharing in patients is compared to sharing under the nul-hypothesis of linkage equilibrium between markers.

References

Bisshop DT and Williamson JA. The power of Identity by State methods for linkage analysis. Am J Hum Genet 1990; 46:254-65

Brown DL, Gorin MB, Weeks DE. Efficient Strategies for Genomic Searching Using the Affected-Pedigree-Member Method of Linkage Analysis. Am J Hum Genet 1994; 54:544-552

Ewens WJ and Spielman RS. The Transmission/Disequilibrium Test: History, Subdivision, and Admixture. Am J Hum Genet 1995; 57:455-465

Fisher RA. The genetical theory of natural selection. Dover publications, 1st ed. 1930 (2nd rev. ed. 1958) New York.

Harper PS. Practical Genetic counselling. Butterworth Heineman, Oxford, 4th ed. 1993

Hodge DE. The information contained in multiple sibling pairs. Genet Epidemiol 1984; 1:109-22

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA and Freimer NB: Genome Screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. Nature Genet 1994; 8:380-386

Kruglyak L and Lander ES. High-Resolution genetic mapping of complex traits. Am J Hum Genet 1995; 56:1212-1223

Lander ES and Botstein D: Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 1986;51:49-62

Lander ES and Botstein: Homozygosity Mapping: A way to Map Human Recessive Traits with the DNA of Inbred Children. Science 1987; 236:1567-1570

Lander ES and Schork NJ: Genetic dissection of Complex Traits. Science 1994; 265:2037-2048

McKusick VA. Mendelian inheritance in man. The Johns Hopskins University Press, 10th ed. 1992, Baltimore and London.

Mueller RF and Young ID. Emery's elements of medical genetics. Churchill Livingstone, 1995, 9th ed, Edinburgh

Ott J. Analysis of human genetic linkage. The Johns Hopkins University Press, rev edition, 1991, Baltimore and London

Penrose LS. The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. Ann Eugen 1935; 6:133-38

Plomin R, Owen MJ, McGuffin P. The genetic basis of complex human behaviors. Science 1994; 264:1733-1739

Risch N. Linkage strategies for genetically complex traits: II. The power of affected relative pairs. Am J Hum Genet 1990; 46:229-241

Spielman RS, McGinnis RE and Warren WJ. Transmission Test for Linkage Disequilibrium: The Insulin gene region and insulin-dependent Diabetes Mellitus (IDDM). Am J Hum Genet 1993; 52:506-516

Suarez BK. The affected sibpair IBD distribution for HLA-linked disease susceptibility genes. Tissue Antigens 1978; 12:78-93

Thomson G. Analysis of complex human traits: an ordered-notation method and new tests for mode of inheritance. Am J Hum Genet 1995; 57(2):474-86

Tierney C and McKnight B. Power of affected sibling method tests for linkage. Hum Hered 1993; 43(5):276-87

Van Soest S, Van den Born LI, Gal A, Farrar GJ, Bleeker-Wagemakers LM, Westerveld A, Humphries P, Sandkuijl LA and Bergen AAB. Assignment of a gene for autosomal recessive retinitis pigmentosa (RP12) to chromosome 1q31-q32.1 in an inbred and genetically heterogeneous disease population. Genomics 1994; 22:499-504

Vogel and Motulsky. Human genetics, Third edition, 1996, Springer Verlag.

Ward PJ. Some Developments on the Affected-Pedigree-Member Method of Linkage Analysis. Am J Hum Genet 1993; 52:1200-1215

Weeks DE and Lange K. The Affected-Pedigree-Member Method of Linkage Analysis. Am J Hum Genet 1988; 42:315-326

Weeks DE and Lange K. A Multilocus Extension of the Affected-Pedigree-Member Method of Linkage Analysis. Am J Hum Genet 1992; 50:859-868

I-2. Perspectives of identity by descent (IBD) mapping in founder populations.

G.J. te Meerman, M.A. van der Meulen and L.A. Sandkuijl

Clin Exp Allergy 1995; 25(suppl 2):97-102.

Clinical and Experimental Allergy, 1995, Volume 25, Supplement 2, pages 97–102

Perspectives of identity by descent (IBD) mapping in founder populations

G. J. TE MEERMAN,* M. A. VAN DER MEULEN* and L. A. SANDKUIJL*†‡

**Department of Medical Genetics, University of Groningen, Groningen, †Institute of Clinical Genetics, Erasmus University, Rotterdam, ‡Department of Human Genetics, Leiden University, Leiden, The Netherlands*

Summary

In a founder population patients with a genetic disease are likely to share predisposing genes from a common ancestor. We show that, depending on the distance of the relationship, patients are expected to share extended segments of DNA around the disease gene. Because of the size of the shared segment, a genomic search with DNA markers for such shared segments, identity by descent (IBD) mapping, can efficiently find the map position of genes, particularly due to genetic drift leading to reduction of heterogeneity and the large number of meioses that is implicitly observed. The statistical power of this method and the approximate cost are given as a function of the density of the map of tested markers and the number of generations since a common ancestor. Initial marker spacings between 5 and 15 centiMorgans are shown to be optimal. IBD mapping is applicable to many genetic diseases, because it does not presuppose a specific genetic model.

Introduction

In recent years much attention has been paid to the possibility of mapping disease genes through association studies. Patients' chromosomes may show a highly increased frequency of one or a few alleles at DNA markers in the immediate vicinity of a disease gene, as has been amply documented for Huntington's disease [1], and cystic fibrosis [2]. Lander and Botstein [3,4] have advocated the search for association on a genome-wide basis, or 'disequilibrium mapping', when no adequate family material can be obtained for linkage mapping. Populations formed only relatively recently from a small group of founders should be most suitable for such disequilibrium or identity by descent (IBD) mapping. In practice, however, association methods are almost exclusively applied to evaluate candidate genes, or for fine mapping of genes within a small candidate region. A complete genomic search for association is generally perceived as very laborious, as it is thought to involve testing thousands of DNA markers at intervals of approximately 1 centiMorgan (cM) or less [5]. We will show here that in appropriately selected populations IBD mapping may be very efficient both in terms of required marker maps and required sample sizes.

In order to obtain a quantitative understanding of the required marker spacing, we firstly compute the expected length of chromosomal segments shared by patients around a shared disease gene. Next, we define a sequential method to screen for shared multimarker haplotypes, where allele sharing at a few adjacent markers in a widely spaced map triggers subsequent saturation of that region with more markers from a denser map. The power and cost of genome screening with this sequential method are derived. Thirdly, we argue that the variety of disease predisposing genes and alleles will be strongly reduced in founder populations due to genetic drift. In selected populations the aetiological complexity of multifactorial diseases should therefore be sufficiently reduced to make IBD mapping feasible.

Correspondence: Dr G. J. te Meerman, Department of Medical Genetics, University of Groningen, A. Deusinglaan 4, 9713 AW Groningen, The Netherlands.

Expected extent of chromosome sharing around a shared disease allele

When association is observed in a founder population

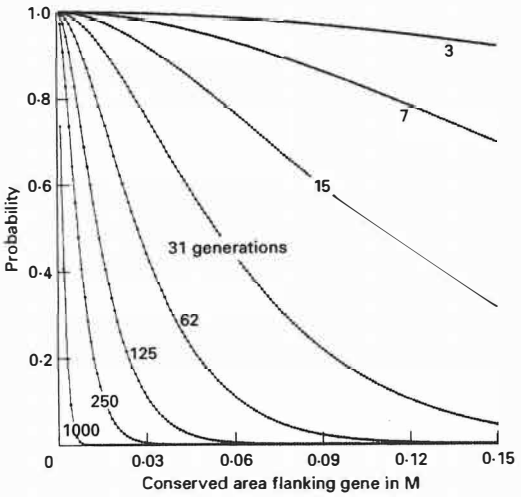


Fig. 1. Probability that the size of the region shared IBD around a common disease allele exceeds x M, for various meiotic counts (generations) between individuals.

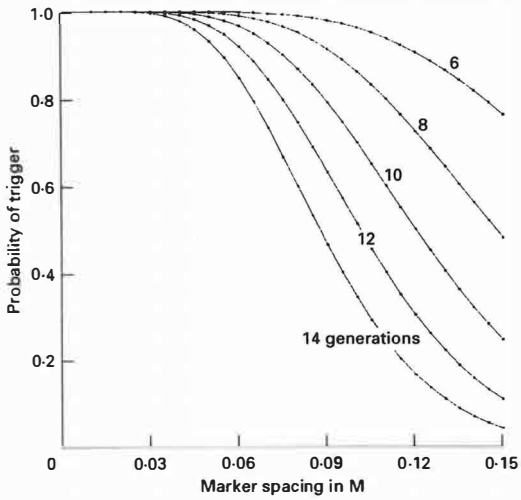


Fig. 2. Probability that a shared haplotype is detected for at least three adjacent markers in the region of the disease gene in at least three out of 10 patients with IBD genes between an arbitary number of patients. All 10 IBD patients are assumed to have inherited their disease gene from a common ancestor a specified number of generations ago.

between a disease and certain marker alleles, it is highly likely that the affected individuals share a common ancestor, who introduced the disease mutation surrounded by the respective marker alleles into the population. While the disease allele is transmitted through generations, the region on either side of the gene that remains identical by descent (IBD) will decrease in size because of cross-overs. The size of the region on one side of the disease gene shared by all affected persons in the last generation has a cumulative probability function (cdf) $1-(1-x)^N$, with N indicating the number of independent meioses connecting the patients (called 'meiotic count') and x the probability of a cross-over in a single meiosis (corresponding for small values of x to the genetic distance in Morgan). For calculation of the size of the region shared on both sides of the disease gene a more complex equation is given in Appendix 1, with its complete derivation. In Fig. 1, we have plotted the probability that the shared region is larger than x for various values of the meiotic count.

The expected size of the chromosomal area shared by two affected individuals follows directly from Fig. 1, while for more than two patients the meiotic counts that separate those patients from their common ancestor should be summed. Obviously, the expected area shared IBD by all decreases in size as more relatives are compared, since the total meiotic count increases with the number of descendants studied.

The power of genome screening for shared haplotypes

As can be seen from Fig. 1, approximately 500 meioses are required to reduce the conserved area around a gene to less than 0·5 cM with 95% probability. Boehnke [6] has recently published similar results, using a different method. Testing a series of markers on a sample of patients with a total meiotic count of 500 or more would be unlikely to yield complete allele sharing by all patients at any of the tested markers. While screening the genome for association one should, however, not search for areas of complete identity initially. Instead, the first step of a screening procedure should identify all areas that show sufficient sharing to merit further attention. The precise choice of the amount of sharing that should be present before an area is regarded as interesting is crucial: a threshold set too low will lead to many false positive signals, while setting the threshold too high may lead one to miss the area containing the disease gene entirely. The probability to localize a disease gene through a random search for association with a complete map of equally spaced markers can be computed accurately when assumptions are made about the proportion of patients that share disease allele(s) IBD, the number of generations that those patients are removed

from their common ancestor, and the definition of the trigger that leads to testing additional markers in a region. As an example, we present results for a sample of arbitrary size with 10 patients sharing a disease mutation of common origin on one of their chromosomes, for various meiotic counts and the worst case assumption that the gene is located precisely between two marker loci (Fig. 2; exact probability calculations are given in Appendix 2).

In this example, we regard areas as promising when a shared haplotype is detected for at least three adjacent markers in at least three of the patients. Additional information can be used from sharing of two locus haplotypes. Therefore, the results give conservative estimates of actual power. Phase is known from genotyping a parent or another close relative of the patient.

In this approach, there are two possible sources of false-positive triggers. A region may be truly IBD, not because there is a shared disease gene in that region, but merely by chance. The probability of such events is minimal, as has been discussed in detail elsewhere [7]. Another possibility is that a region appears identical initially for a few markers, but that it is in reality not completely identical, as will become evident as additional markers in the region are tested. We have assessed the probability of occurrence of this second type of false positive trigger by simulation, assuming markers with 10 alleles each, with randomly chosen allele frequencies. For a sample of 10 patients with 20 chromosomes and known phase of marker alleles, false positive triggers were found for fewer than 2% of all tested regions. When the sample size was increased to 40 chromosomes up to 12·5% of regions yielded false positive triggers. All such regions should be included in the next series of marker tests, but typically one would select first the most promising areas by gradually decreasing the threshold. Exclusion of false positive findings requires the additional testing of only one or two highly polymorphic markers. Confirmation will occur when sharing of a haplotype over a smaller area is observed in more individuals.

Causes of heterogeneity can be divided in three broad categories: allelic heterogeneity exists when multiple disease predisposing mutations at a single disease locus segregate in a population, locus heterogeneity when mutated alleles at more than one locus can independently lead to disease, and aetiological heterogeneity when the disease is due to non-genetic causes in some of the patients (phenocopies). Our subsequent examples will concentrate on the more realistic assumption that 50% of all patients share their disease mutation from a common ancestor, while the remaining patients will be treated as phenocopies (other types of heterogeneity lead to more favourable results, see Discussion).

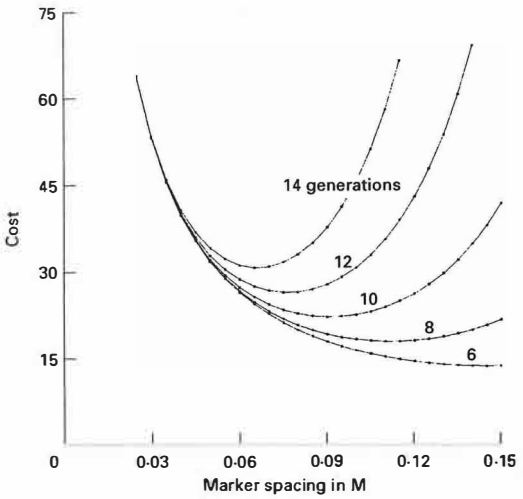


Fig. 3. Relationship between cost of a genome screen (arbitrary units) and marker spacing for various numbers of generations separating the tested individuals from their common ancestor. Twenty affected persons in the analysis and a homogeneity of 50%.

Optimal initial spacing of markers for monogenic diseases

The cost of a genomic search for a monogenic disease will depend on the number of markers and the number of persons that will have to be genotyped. We present a cost calculation for a genomic search in 20 affected individuals under 50% allelic homogeneity (10 disease alleles IBD). The cost will be directly proportional to the number of markers tested, i.e. inversely proportional to marker spacing: testing a too dense map will lead to unnecessary expenses. When the marker spacing is chosen too wide, however, the probability of missing the true location of the disease gene increases rapidly. The risk of not finding the gene in the first screening, which would necessitate testing a second series of markers at considerable expense, leads to an approximately inverse relationship between costs and statistical power in the initial screening map; in Fig. 3 cost is calculated as 2/(spacing*power).

The lowest cost was found for a rather coarse grid of markers spaced at 11 cM, assuming that the sample included 10 copies of a disease mutation inherited from

a common ancestor six generations ago. In a similarly sized sample optimal spacing was reduced to approximately 5 cM when the common ancestor lived 14 generations ago. Selection of suboptimal spacing will lead to increased costs, particularly for larger meiotic counts. Testing markers at intervals of less than 3 cM will lead to a sharp rise of costs without potential benefit, even for relatively larger distances between the patients in the sample and their common ancestor.

Discussion

Linkage mapping in extended families has been successfully applied for mapping clearly Mendelian inherited disorders. Rare recessive diseases, diseases with considerable genetic or aetiological heterogeneity and oligogenic diseases are less amenable to linkage mapping, as suitable families are scarce and probably not representative. Alternative methods of statistical analysis, such as the affected pedigree member method or sib-pair method address this problem only partially, as they still focus on closely related patients [8–11]. Association-based methods have been advocated, as they can accommodate samples of mostly distantly related patients, but were generally perceived as very inefficient. We have demonstrated here that in carefully selected populations IBD mapping may be as efficient as linkage mapping for finding the approximate location of a disease gene.

The IBD mapping procedure that we propose uses haplotype sharing at several markers rather than differences in allele frequencies at individual markers to identify regions of interest. Our statistical approach, which is open for further refinements, resembles existing non-parametric approaches in that it concentrates on affected persons only, thereby obviating the need for exact statistical parameters to describe disease transmission. Essential for the method is the high degree of polymorphism of the currently available marker maps. The method is limited, however, to disease genes that are subject to genetic drift and have an appreciable lifetime in the population.

IBD mapping further compares favourably to linkage mapping for a later stage of mapping projects: genetic fine mapping. While fine mapping in independent pedigrees requires scoring of hundreds of meioses as recombinant or non-recombinant, fine mapping in founder populations may be achieved on the basis of overlap of haplotypes in patients. Each patient adds several meioses to the meiotic count. Consistent overlap occurs if and only if the disease allele originates from a common founder.

Empirical support for our calculations comes from several recent mapping studies of complex diseases, including melanoma and colon cancer, in more or less isolated communities [12–15]. These studies reported sharing of extended haplotypes by patients removed as far as 450 years from their common ancestor. Such findings had until now not resulted in a reappraisal of the efficiency of association methods for gene mapping. Puffenberger *et al.* report the successful application of some variant of the method we propose for mapping genes predisposing for Hirschsprung disease [16]. We show that their results are not accidental, but can be expected.

As genes are mapped for an increasing number of simple disorders, other diseases with a possible genetic component have received the epithet 'complex'. Genetically complex diseases constitute an ill-defined group, ranging from clearly Mendelian diseases that show extensive locus heterogeneity (e.g. for retinitis pigmentosa more than 12 genes have been identified showing autosomal dominant, autosomal recessive and X-linked recessive inheritance [17]) to diseases with only a presumed polygenic background. It is conceivable that complex diseases are caused by such complex interplay of environment and allelic interactions that most mapping efforts will fail, but it is probable that IBD mapping will be instrumental in mapping genetic contributions for some complex diseases. Founder populations, the target populations for IBD mapping, characteristically show the effects of genetic drift. While the number of mutant genes introduced in the small founder population of centuries ago will already be restricted, it will be further reduced by genetic drift which leads to chance elimination of alleles over time. For diseases that show extensive locus heterogeneity worldwide, the expected heterogeneity in founder populations is reduced. Locus heterogeneity (and allelic heterogeneity) are not as detrimental for IBD mapping as is the existence of non-genetic cases. Each disease allele segregating in a founder population represents a potential source for haplotype sharing in patients. The triggers that we described earlier for mapping monogeneic diseases might be easily adjusted to detect unusual sharing of more than one haplotype in a particular region, or alternate sharing of haplotypes at two different locations in the genome.

Multifactorial diseases represent a more serious challenge for genetic mapping. It is important to realize, however, that the exact mechanism of disease aetiology is not directly relevant for the power of IBD mapping. The limiting factor is instead what proportion of patients will share a predisposing mutation. In our previous examples, this proportion was fixed at 50%. Similar calculations may be carried out for smaller genetic contributions, probably requiring adjustment of thresholds. In our statistical evaluation, no comparison is made with a

control population. As a consequence, genetic factors may be detected that contribute only mildly to the total disease risk. It is possible, and under conditions of genetic drift even likely, that different contributing factors will be detected in different founder populations.

The theory discussed leads to a simple advice to investigators currently using affected sib pair methods. Absolute scoring of alleles will allow analysis of data as if they had been gathered in an IBD context. Although such scoring is more difficult, it is the only way to use the information contained in the present high degree of polymorphism of current marker loci.

Acknowledgments

We thank Dr Jan Vijg (BIH, Harvard) and Professor Charles Buys for comments on the manuscript. This work was supported in part by The Netherlands Organisation for Scientific Research.

Table 1. Combination of recombination events, 14 generations, spacing between markers 10cM, disease allele 2 cM from marker 2, 8 cM from marker 3

		P(4)	P(5)	P(6)
		0·689	0·249	0·062
P(1)	0·246	0·170	0·061	0·015
P(2)	0·587	0·404	0·146	0·037
P(3)	0·167	0·115	0·042	0·010

Example computation:

$P(4) = 1-(1-0{\cdot}08)^{14} = 0{\cdot}689$ (at least one recombinant on area of 8 cM)

$P(6) = (1-0{\cdot}18)^{14} = 0{\cdot}062$ (no recombinant on area of 18 cM)

$P(5) = 1-P(4)-P(6)$

The markers with the same allele as in the founder:

none	3	34
2	23	234
12	123	1234

Appendix 1. Derivatrion of the size of the chromosome fragment shared after N meioses

In N meioses the probability of no recombination at a genetic distance X from a gene is $(1 - X)^N$. The meiotic count for two children with the same great-grandparents is, e.g. 6. An additional brother of one of these children would increase the meiotic count by one.

The probability distribution of the unchanged fragment formed by the combined area around the gene is the convolution of the pdf $(Nx(1 - X)^{N-1})$ with itself:

$$\mathrm{con}(Y) = \int_0^Y N^2((1 - Y + x)(1 - x))^{N-1}\,dx$$

The probability that the area around the gene is larger than Z, is given by

$$P(\text{area} > Z) = 1 - \int_0^Z \mathrm{con}(Y)dY$$

Although partial integration, and therefore an exact solution, is possible, we have used double numerical integration for the computations. The PASCAL program used for the calculations is available on demand by E-mail (email:g.j.te.meerman@med.rug.nl).

Appendix 2. Derivation of the probability of a triggering event, which is equal to the power of the IBD mapping strategy

Assume that there are four equally spaced marker loci, numbered 1–4. The disease locus is between the second and the third marker. Then a combination of the following events may occur for each observed person sharing a gene IBD:

- a recombinant occurs between the gene and marker 2,
- a recombinant occurs between markers 1 and 2,
- a recombinant occurs outside marker 1.

the same applies for recombinants on the other side of the gene:

- a recombinant occurs between the gene and marker 3,
- a recombinant occurs between markers 3 and 4,
- a recombinant occurs outside marker 4.

These two events are considered to be independent, because most of them will occur in different meioses, so interference will not influence the calculations to any significant degree. All nine possible events leading to sharing of alleles for marker 1 and 2, 2 and 3, 3 and 4, are indicated by the loci for which sharing occurs. The probability is then calculated that no, one, two and three non-informative locus haplotype sharing events occur (e.g. two haplotypes 123 and one 234) for each gene assumed to be present from the founder, given the meiotic distance from the founder. Table 1 gives, as an example, the matrix of all probabilities of combinations of events 1–3 and 4–6, for an individual 14 generations from the founder, with marker spacing of 10 cM, and the position of the gene 2 cM from marker 2.

The probability that an informative set is found, if N founder genes are present, is then computed from a binomal distribution, combining all non-informative

events and subtracting the probability from 1. This probability is then the power to detect a gene through shared haplotypes. It is assumed that after a triggering event additional markers within and two either side of the haplotype are investigated. If sharing occurs between adjacent intervals, more information results that is not taken into account for the calculation. All calculations that are shown are minimum power calculations, where the gene is assumed to be located exactly between markers 2 and 3. A slight increase in power results if the disease allele is nearer to one marker. The derivation of optimal test statistics is beyond the scope of this article.

Three shared haplotypes with a carrier frequency of less than 1% is about the minimum number required to find statistically sound evidence for involvement of a gene. The probability of such sharing due to random effects among 10 or less affected individuals is less than 10^{-4}, equivalent to a LOD score of 4.

References

1 MacDonald ME, Noveletto A, Lin K et al. The Huntington's disease candidate region exhibits many different haplotypes. Nature Genet 1992; 1:99–103.

2 Kerem BS, Rommens JM, Buchanan JA et al. Identification of the cystic fibrosis gene. Genetic analysis science 1989; 245:1073–80.

3 Lander ES, Botstein D. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 1986; 51:49–62.

4 Lander ES, Botstein D. Homozygosity Mapping: A way to map human recessive traits with the DNA of inbred children. Science 1987; 236:1567–70.

5 Plomin R, Owen MJ, McGuffin P. The genetic basis of complex human behaviors. Science 1994; 264:1733–9.

6 Boehnke M. Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 1994; 55:379–90.

7 Donnelly KP. The probability that related individuals share some section of genome identical by descent. Theor Pop Biol 1983; 23:34–63.

8 Ward PJ. Some developments on the affected-pedigree-member method of linkage analysis. Am J Hum Genet 1993; 52:1200–15.

9 Brown DL, Gorin MB, Weeks DE. Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. Am J Hum Genet 1994; 54:544–52.

10 Weeks DE, Lange K. The affected-pedigree-member method of linkage analysis. Am J Hum Genet 1988; 42:315–26.

11 Weeks DE, Lange K. A Multilocus Extension of the Affected-Pedigree-Member Method of Linkage Analysis. Am J Hum Genet 1992; 50:859–68.

12 Cannon-Albright LA, Goldgar DE, Gruis NA et al. Localization of the 9P Melanoma susceptibility locus to a 2cM region between D9S736 and D9S171 Genomics 1994; 23:265.

13 Gruis N, Sandkuijl LA, Bergman W, Frants RR. Common 9P haplotype in Dutch FAMMM families. PhD thesis N. Gruis, Genetics of the Familial Atypical Multiple Mole-Melanoma Syndrome, Leiden, 1994.

14 Nystrom-Lahti M, Sistonen P, Mecklin J-P et al. Close linkage to chromosome 3p and conservation of ancestral founding haplotypes in hereditary nonpoloyposis colorectal cancer families. Proc Natl Acad Sci USA 1994; 91:6054–58.

15 Sulisalo T, Francomano CA, Sistonen P et al. High-resolution genetic mapping of the cartilage-hair hypoplasia (CHH) gene in Amish and Finnish families. Genomics 1994; 20:347–53.

16 Puffenberger EG, Kauffman ER, Bolk S et al. Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 1994; 3:1217–25.

17 Soest S van, Born LI van den, Galu A et al. Assignment of a gene for autosomal recessive retinitis pigmentosa (RP12) to chromosome 1q31-q32.1 in an inbred and genetically heterogeneous disease population. Genomics 1994; 22:499–504.

I-3. Haplotype identity between individuals who share a CFTR mutation allele Identical by descent: demonstration of the usefulness of the haplotype sharing concept for gene mapping in real populations.

Hendrik G. de Vries, Martin A. van der Meulen, Rima Rozen, Dickie J.J. Halley, Hans Scheffer, Leo P. ten Kate, Charles H.C.M. Buys and Gerard J. te Meerman.

Human Genetics 1996; 98:304-309

Hum Genet (1996) 98 : 304–309

ORIGINAL INVESTIGATION

**Hendrik G. de Vries · Martin A. van der Meulen
Rima Rozen · Dickie J. J. Halley · Hans Scheffer
Leo P. ten Kate · Charles H. C. M. Buys
Gerard J. te Meerman**

Haplotype identity between individuals who share a CFTR mutation allele "identical by descent": demonstration of the usefulness of the haplotype-sharing concept for gene mapping in real populations

Received: 30 November 1995 / Revised: 11 April 1996

Abstract Cystic fibrosis (CF) patients with the A455E mutation, in both the French Canadian and the Dutch population, share a common haplotype over distances of up to 25 cM. French Canadian patients with the 621+1G→T mutation share a common haplotype of more than 14 cM. In contrast, haplotypes containing the ΔF508 mutation show haplotype identity over a much shorter genomic distance within and between populations, probably because of the multiple introduction of this most common mutation. Haplotype analysis for specific mutations in CF or in other recessive diseases can be used as a model for studying the occurrence of genetic drift conditional on gene frequencies. Moreover, from our results, it can be inferred that analysis of shared haplotypes is a suitable method for genetic mapping in general.

Introduction

In recent times, it has been repeatedly observed that haplotypes surrounding rare alleles of a gene are large (Rozen et al. 1990; Cannon-Albright et al. 1994; Gruis et al. 1994; Nystrom-Lahti et al. 1994; Houwen et al. 1994; Hastbacka et al. 1994; Sulisalo et al. 1994; Meyers et al. 1994; Heyer and Tremblay 1995). Sharing of large genomic areas can be used as a method for mapping disease genes: this is termed identity by descent (IBD) mapping (Houwen et al. 1994; te Meerman et al. 1995). An empirical question is whether haplotype sharing can be observed in real populations to an extent where IBD mapping using haplotype sharing is feasible.

As an empirical model for high and low frequency alleles weakly associated with a disease or with dominance and low penetrance, we have chosen to study mutations leading to the recessive disease cystic fibrosis (CF). Only a small fraction of the disease alleles present in the population can be observed in diseased individuals. To date, more than 500 presumed mutations have been identified in the CF transmembrane conductance regulator gene. The most common mutation (ΔF508) has a high frequency in Caucasian populations (up to 1.5%). A less frequent mutation is the A455E missense mutation. According to data from the CF consortium (Cystic Fibrosis Genetic Analysis Consortium 1994), this mutation is mainly detected among French Canadian and Dutch CF patients. The A455E mutation comprises 8% of all CF mutations in the French Canadian population of the Saguenay-Lac St. Jean region of Quebec (Rozen et al. 1992) and 3% in The Netherlands (unpublished data). The overall CF carrier frequency is 1 in 15 in this particular French Canadian population (Rozen et al. 1992) and 1 in 30 in The Netherlands (ten Kate 1977). This results in allele frequencies of 1/400 (8% of 1/30) and 1/2000 (3% of 1/60), respectively. The concentrated geographical distribution indicates that this mutation has been recently introduced.

In models for multifactorial disease, the mutated gene concept is not directly applicable, since alleles act as risk factors in combination with other factors. However, if gene/gene or gene/environment interactions are modelled in founder populations, where the multifactorial background is more in common than in mixed populations, specific alleles of genes may be involved, leading to association at the population level. The implication is that such alleles show increased frequencies in affected individuals. It cannot be ruled out that multifactorial diseases are determined by only a few genes; this would make an increased

H. G. de Vries · M. A. van der Meulen · H. Scheffer · C. H. C. M. Buys · G. J. te Meerman (✉)
Department of Medical Genetics, University of Groningen, Antonius Deusinglaan 4, 9713 AW Groningen, The Netherlands
Tel.: +31-50-3632925; Fax: +31-50-3632947;
email: G.J.te.meerman@med.rug.nl

R. Rozen
Departments of Human Genetics, Pediatrics and Biology, Montreal, Canada

D. J. J. Halley
Department of Clinical Genetics, Erasmus University, Rotterdam, The Netherlands

L. P. ten Kate
Department of Human Genetics, Free University of Amsterdam, The Netherlands

frequency of specific alleles in affected individuals even more likely (van Ommen 1995). By studying haplotype sharing in two populations, one with a very late origin (French Canadians) and one with a longer history (southern part of The Netherlands), we demonstrate the usefulness of the haplotype-sharing concept for gene mapping in real populations.

Materials and methods

Recruitment of CF patients

The A455E mutation, which is associated with a less severe CF phenotype, has mainly been detected in Canada and in The Netherlands (Cystic Fibrosis Genetic Analysis Consortium 1994). In Canada, this mutation was introduced by French immigrants during the period 1650–1900 (Rozen et al. 1990, 1992). The Dutch patients with an A455E mutation all come from southern parts of The Netherlands (unpublished results). Blood samples from 15 independent Dutch CF patients with the A455E mutation were collected in Groningen and Rotterdam. Samples from 10 French Canadian CF patients with the A455E mutation, from a subpopulation in north-eastern Quebec, viz., the Saguenay-Lac St.Jean region (Rozen et al. 1990, 1992), were collected in Montreal. The father of proband 3 and the mother of proband 6 were carriers of the 621+1G→T mutation and were sibs. Because all patients were compound heterozygotes, haplotype sharing could also be determined around two other CF mutations (ΔF508, 621+1G→T). Clinical diagnosis in all patients was confirmed by demonstrating the ΔF508 (Scheffer et al. 1989), A455E (Kerem et al. 1990), and 621+1G→T (Zielenski et al. 1991a) mutations. In all cases, DNA from one of the parents was used for phase determination.

Microsatellite analysis

Three intragenic microsatellites, IVS8CA (intron 8), IVS17BTA, and IVS17BCA (intron 17b; Zielenski et al. 1991b) and 7 extragenic microsatellites, D7S518, D7S501, D7S523, D7S486, D7S480, D7S490, and D7S635 (Gyupay et al. 1994), were analyzed by single amplification. The primer sequences are shown in Table 1. The polymerase reaction (PCR) was performed with 400 ng DNA, 50 mM KCl, 10 mM TRIS (pH 9.0), 1.5 mM $MgCl_2$, 0.01% gelatin, 0.1% Triton X-100, 0.5 μM of each microsatellite primer (biotinylated), and 0.25 U Super *Taq* (SphearoQ), in a total volume of 50 μl. Part of the PCR product was diluted to 1:20 and mixed with a formamide solution. Absolute PCR product sizes were determined by adding standard size markers (generated from Pharmacia m13mp18) to each sample. The reaction products were detected in a 6% denaturing polyacrylamide gel with an Automated Laser Fluorescent system (Pharmacia LKB). In total, a region of 25 cM flanking the CF gene was genotyped. For certain microsatellite alleles, it was not possible to establish phase. In these cases, haplotypes were assigned by comparing alleles of flanking microsatellites between the CF patients.

Results

The results of the haplotype analysis of the Canadian and Dutch CF patients with a A455E mutation are shown in Table 2. The haplotype of the intragenic markers on the A455E chromosome of the 10 French Canadian patients is identical in all cases, namely 22–35–13, in which the numbers represent the numbers of repeats of IVS8CA, IVS17BTA, and IVS17BCA, respectively. The extragenic alleles are numbered according to the relative length of the PCR product. Of the 10 Canadian patients, 9 share a region of at least 5 cM surrounding the A455E mutation (Fig. 1). Patient nos. 1, 2, and 3 share a region of more than 25 cM. This is also the case, with a different haplotype, for patient nos. 4, 5, and 6. The haplotype of the intragenic markers of 14 of the 15 Dutch patients are identical to the Canadian patients. They share a region of more than 4 cM surrounding A455E (Fig. 2). The remaining patient shows a different intragenic haplotype (22–37–13).

The haplotypes of the non-A455E chromosomes of the Canadian and Dutch patients are shown in Table 3. Four Canadian patients have a 621+1G→T mutation on the other CF chromosome, and six have a ΔF508 mutation.

Table 1 PCR primers of microsatellites at and flanking the CF locus. Data have been derived from Zielenski et al. (1991b) and Gyapay et al. (1994)

Marker	Primer sequences 5′→3′	Hetero-zygosity	Distance to CFTR gene (cM)
D7S518	CAGTAGGCAGGGGTGG GGGTGTGTCTGTGTGACAAC	0.87	15
D7S501	CACCGTTGTGATGGCAGAG ATTTCTTACCAGGCAGACTGCT	0.81	7
D7S523	CTGATTCATAGCAGCACTTG AAAACATTTCCATTACCACTG	0.80	1
D7S486	AAAGGCCAATGGTATATCCC GCCAGGTGATTGATAGTGC	0.81	0.2
IVS8BTA	TCTATCTCATGTTAATGCTG GTTTCTAGAGGACATGATC	0.41	0
IVS17BTA	GACAATCTGTGTGCATCG GCTCGATTCTATAGGTTATC	0.89	0
IVS17BCA	AAACTTACCGACAAGAGGA TGTCACCTCTTCATACTCAT	0.38	0
D7S480	CTTGGGGACTGAACCATCTT AGCTACCATAGGGCTGGAGG	0.86	2
D7S490	CCTTGGGCCAATAAGGTAAG AGCTACTTGCAGTGTAACAGCATTT	0.78	5
D7S635	CCAGGCCATGTGGAAC AGTTCTTGGCTTGCGTCAGT	0.81	10

Table 2 Microsatellite haplotypes for independent A455E chromosomes from a French Canadian and a Dutch population (! phase unknown)

	5	3	2	Intragenic			1	6	8	(cM)
	D7S-635	490	480	8CA	17BTA	17BCA	486	523	501	518
French Canadian patients										
1	7!	1	3	22!	35!	13	8	4	5	10
2	7	1	3	22	35!	13	8	4!	5	10
3	7	1	3	22	35	13	8	4	5!	10
4	3!	4	3	22!	35	13	8	4	5	10
5	3	4	3	22	35	13	8	4	5	10
6	3	4	3!	22	35	13	8	4	5!	10
7	1!	1	2	22!	35	13	8	4	5	2
8	1	1	2	22	35	13	8	4!	4	10
9	4!	2	2	22	35	13	8	4	5	10
10	7	1!	3	22	35	13	8	3	3	2
Dutch patients										
1	7	1	3	22	35!	13	8	4	3	2
2	1	1	3	22!	35	13	8	4	1	12
3	1	4	6	22	35	13	8	4	1!	4!
4	6	1	3	22	35	13	8	5	7	9
5	6	1	3	22	35	13	8	5	7	9
6	6	1	3	22!	35	13	8	5	5	2
7	6	1!	3	22	35	13	10	5	5	2
8	1	1	3!	22	35	13	8	5!	3	11
9	4	1	3	22	37	13	8	5	3	10
10	3	1	3	22!	35	13	8	3	3	12!
11	5	6	3	22	35	13	8	5	6	2
12	8	2	3	22!	35	13	8	5	7	2
13	5!	6	3	22	35	13	8	3	7!	10
14	10	2	3	22	35	13	8!	2	3	3
15	1	4	5	22	35!	13	8	5	7	3

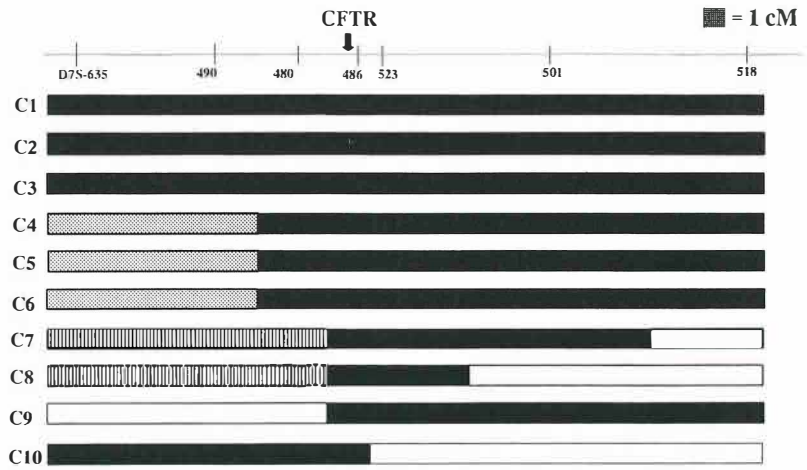


Fig. 1 DNA sharing around the CF gene of 10 unrelated French Canadian CF patients (*C1–10*) with the A455E mutation. DNA regions with the same grade of *shading* represent identical haplotypes. *White regions* represent unique haplotypes within this population

Out of the 15 Dutch CF patients, 13 have a ΔF508 mutation on the other chromosome. The two remaining Dutch patients, nos. 2 and 10, have the 1717–1G→A and the R553X mutation, respectively, on their other chromosome. The French Canadian patients with the 621+1G→T mutation have identical intragenic haplotypes (21–31–13) and share a DNA region of more than 14 cM. Little or no sharing can be observed when the ΔF508 chromosomes in and between the two populations are compared.

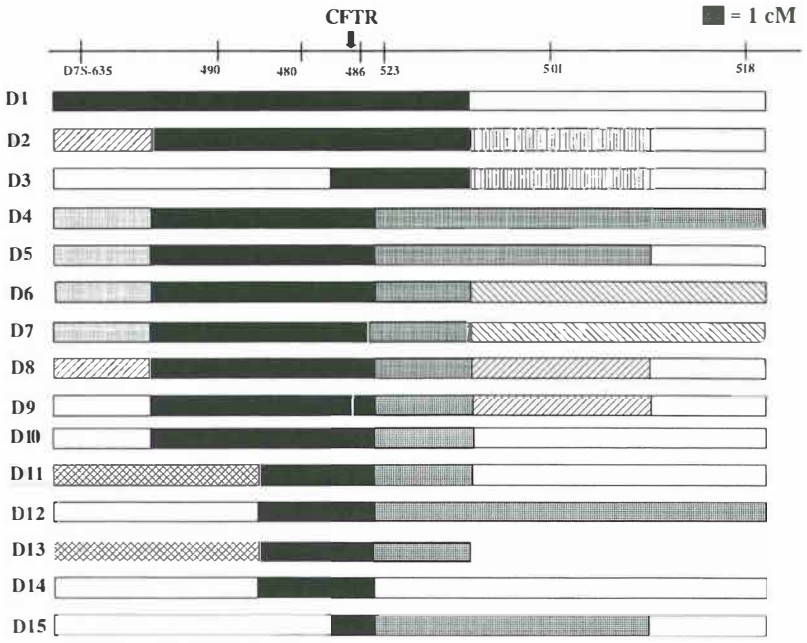


Fig.2 DNA sharing around the CF gene of 15 unrelated Dutch CF patients (*D1–15*) with the A455E mutation. DNA regions with the same grade of *shading* represent identical haplotypes. *White regions* represent unique haplotypes within this population

Table 3 Microsatellite haplotypes for independent ΔF508 and 621+1G→T chromosomes from a French Canadian and a Dutch population (! phase unknown)

		5	3	2	Intragenic			1	6	8 (cM)	
	D7S-635	490	480		8CA	17BTA	17BCA	486	523	501	518
French Canadian patients											
2	621+1G→T	1	4	5	21	31!	13	8	5!	7	3
3	621+1G→T	1	4	5	21	31	13	8	5	1!	6
6	621+1G→T	1	4	5!	21	31	13	8	5	7!	5
9	621+1G→T	1!	4	5	21	31	13	8	5	5	2
1	ΔF508	9!	3	1	17!	32!	13	8	7	1	3
7	ΔF508	9!	3	1	17!	32	13	8	1	5	2
4	ΔF508	7!	4	1	17!	32	13	8	7	1	4
5	ΔF508	9	6	7	17!	32	13	5	5	7	5
8	ΔF508	2	4	3	17	35	17	3	5!	8	7
10	ΔF508	8	4!	2	23	31	13	11	5	3	4
Dutch patients											
5	ΔF508	6	1	6	17	31	13	10	4	3	2
7	ΔF508	1	6!	8	17	31	13	10	6	5	9
8	ΔF508	4	2	7!	17	31	13	10	3!	6	11
15	ΔF508	1	6	5	17	31!	13	10	2	7	4
1	ΔF508	4	1	3	17	31!	13	3	3	4	2
9	ΔF508	8	6	3	17	31	13	8	4	1	7
6	ΔF508	7	3	3	17!	32	13	8	2	1	2
11	ΔF508	1	4	7	17	32	13	8	6	5	9
14	ΔF508	1	2	1	17	32	13	5!	3	7	7
4	ΔF508	3	3	2	17	35	17	10	5	4	2
12	ΔF508	8	1	2	17!	35	17	10	2	7	4
13	ΔF508	4!	1	5	17	47	13	10	4	8!	3
3	ΔF508	1	2	1	23	31	13	8	5	5!	10!
10	1717 1G–A	1	1	3	17!	35	16!	8	3	7	6!
2	R553X	1	1	6	11	55	17!	10	5	8	7

Discussion

The Saguenay-Lac St.Jean region was settled by approximately 5000 immigrant families during 1838–1911 (Rozen et al. 1990). The identical haplotype of the intragenic markers and the sharing of large DNA regions on the A455E chromosomes indicate that the A455E mutations that we now observe result from a single introduction into this population. The three intragenic markers in the Dutch population show the same haplotype. Furthermore, some of the Canadian and Dutch patients share a region of more than 4 cM around the CF gene. This indicates that the A455E mutation has been inherited from a not-too-distant predecessor. In the same French Canadian population, extensive sharing of haplotypes surrounding the mutated myotonic dystrophia and rickettsia genes has been demonstrated (Bétard et al. 1995). Dutch patient no. 9 shows the only aberrant haplotype (22–37–13). This can best be explained by a mutation in the original haplotype, because the neighboring markers in this patient are comparable with almost all the other Dutch patients.

The ΔF508 mutation is a very old mutation that was introduced at least 52 000 years ago into Europe (Morral et al. 1994). Much variation is observed in intragenic markers; this is probably because of repeat length mutations subsequent to the ΔF508 mutation. This most frequent CF mutation has been widely distributed, especially in Caucasian populations, over a long period of time. In the Saguenay-Lac St.Jean region, we have detected 3 different intragenic haplotypes in only 6 patients with a ΔF508 mutation. This means that there were probably multiple independent introductions of the ΔF508 mutations into this population. Sharing of DNA regions surrounding ΔF508 mutations is limited to individuals with the same haplotype of intragenic markers. The observed intragenic haplotypes are also the most common haplotypes in Europe (Morral et al. 1994). To detect haplotype sharing surrounding ΔF508 mutations, more observations with identical intragenic haplotypes are necessary.

If the size of a shared haplotype is large, the common predecessor must be recent. The difference in the extent of sharing found for the ΔF508 mutation and for the other mutations shows that haplotype sharing can be expected to occur only if a limited number of independent introductions of a gene has been made into the population. Haplotype sharing and other association studies can therefore be expected to be successful in founder populations of a specific size relative to the gene frequency. Because of genetic drift, the independent introduction of 10–20 copies will lead to perhaps 1–4 copies remaining in widely varying numbers of individuals after 8 or more generations. This is in agreement with the theoretical expectation according to Fisher (Vogel and Motulski 1986), who computed that the probability of survival of a single gene in a stable population after 10 generations is about 10%. It appears from the French Canadian genealogical data that gene flow cannot be traced back to unique predecessors, even if all genealogical data are complete. The informativity of genealogical data therefore seems limited relative to the information being learned from direct haplotype comparison.

The similarity in intragenic ΔF508 haplotypes in the two populations suggests a multiple introduction of the same intragenic haplotypes, of which some have survived. In Finland, genetic drift has led, by this process, to a low CF frequency and to a distribution of mutations different from that of the rest of Europe (de la Chapelle 1993).

The suitability of a population for haplotype-sharing studies will increase if there is a larger population with which the founder population is connected. Once a genomic region has been identified that possibly contains an allele predisposing to a disease, it is easy to perform a genome screening in affected individuals at the 1 cM level. This situation seems to exist both for the French Canadians and in The Netherlands, where more isolated areas are connected with larger populations. At genomic distances of 1 cM, association and linkage disequilibrium can be detected in populations much larger than the one originally investigated. At the same time, a large number of meioses is being observed implicitly, thus narrowing the region in which the gene must be located. This is well illustrated from the consistent overlap shown in and between Figs. 1 and 2. In The Netherlands, the A455E data come from an unselected population, which shows that, for rare genes such as the A455E mutation, founder effects are present at the population level. A complication in haplotype comparison is marker allele mutation. In this dataset, we probably observe two such mutations (Dutch patients nos. 7 and 9 show an allele mutation in the microsatellites D7S486 and IVS17BTA, respectively) where the haplotype for surrounding markers seems to be conserved.

The present study shows that the suitability of populations for gene mapping through direct haplotype comparison can be investigated using some of the well-known recessive disease mutations that can easily be detected at the level of carriers. These mutations generally have a high frequency, as would be expected for alleles of genes involved in multifactorial diseases. By relating the gene frequency to the size of the shared haplotypes, an impression can be obtained regarding whether haplotype-sharing analysis can be used to find the map location of other genes.

Acknowledgements The authors thank Dr. J. P. Heyerman (Leyenborg Hospital, The Hague) for sending DNA samples of the parents of Dutch CF patients. This study was made possible by a grant from The Netherlands Organization for Scientific Research (NWO).

References

Bétard C, Raeymaekers P, Ouellette G, Jomphe M, Labuda M, Glorieux F, Mathieu J, Laberge C, Cassiman J-J, Sandkuijl LA, Gauvreau D (1995) The validation of a novel linkage disequilibrium mapping technique on Steinert myotonic dystrophy and pseudovitamin D deficient rickets in a founder population (abstract). Med Genet 2:238

Cannon-Albright LA, Goldgar DE, Gruis NA, Neuhasen S, Anderson DE, Lewis CM, Jost M, Tran TD, Nyguen K (1994) Localization of the 9p melanoma susceptibility locus to a 2 cM region between D9S736 and D9S171. Genomics 23:265

Chapelle A de la (1993) Disease gene mapping in isolated human populations: the example of Finland. J Med Genet 30:857–865

Cystic Fibrosis Genetic Analysis Consortium (1994) Population variation of common cystic fibrosis mutations. Hum Mutat 4:167–177

Gruis N, Sandkuijl LA, Bergman W, Frants RR (1994) Common 9p haplotype in Dutch FAMMM families. Genetics of the familial atypical multiple mole-melanoma syndrome. PhD thesis, Leiden

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, Bernardi G, Lathrop M, Weissenbach J (1994) The 1993–1994 Généthon human genetic linkage map. Nat Genet 7:246

Hastbacka J, Chapelle A de la, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Triveldi B, Weaver A, Coloma A, Lovett M, Buckler A, Kaitila I, Lander ES (1994) The diastrophic dysplasia gene encodes a novel sulphate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Heyer E, Tremblay M (1995) Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet 56:970–978

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8:380–386

Kate LP ten (1977) Cystic fibrosis in The Netherlands. Int J Epidemiol 6:23–35

Kerem B, Zielenski J, Markiewicz D, Bozon D, Gazit E, Yahav J, Kennedy D, Riordan JR, Collins FS, Rommens JM, Tsui L-C (1990) Identification of mutations in regions corresponding to the two putative nucleotide (ATP)-binding folds of the cystic fibrosis gene. Proc Natl Acad Aci USA 87:8447–8451

Meerman GJ te, Meulen MA van der, Sandkuijl LA (1995) Perspectives of identity by descent (IBD) mapping in founder populations. Clin Exp Allergy 25 (Suppl 2):97–102

Meyers DA, Postma DS, Panhuysen CIM, Xu J, Amelung PJ, Levitt RC, Bleecker ER (1994) Evidence for a locus regulating total serum IgE levels mapping to chromosome 5. Genomics 23:464

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Giménez J, Reis A, Varon-Mateeva R, Macek Jr M, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo MP, Garnerone S, Restagno G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwartz M, Novelli G, Ferec C, Arce M de, Nemeti M, Kere J, Anvret M, Dahl N, Kadasi L (1994) The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nat Genet 7:169–175

Nystrom-Lahti M, Sistonen P, Mecklin J-P, Pylkkanen L, Aaltonen LA, Jarvinen H, Weissenbach J, Chapelle A de la, Peltomaki P (1994) Close linkage to chromosome 3p and conservation of ancestral founding haplotypes in hereditary nonpolyposis colorectal cancer families. Proc Natl Acad Sci USA 91: 6054–6058

Ommen GJ van (1995) A foundation for limb-girdle muscular dystrophy. Nat Med 1:412–414

Rozen R, Schwartz RH, Hilman BC, Stanislovitis P, Horn GT, Klinger K, Daigneault J, Braekeleer M de, Kerem B-S, Tsui L-C, Fujiwara TM, Morgan K (1990) Cystic fibrosis mutations in North American populations of French ancestry: analysis of Quebec French Canadian and Louisiana Acadian families. Am J Hum Genet 47:606–610

Rozen R, De Braekeleer M, Daigneault J, Ferreira-Rajabi L, Gerdes M, Lamoureux L, Aubin G, Simard F, Fujiwara TM, Morgan K (1992) Cystic fibrosis mutations in French Canadians: three CFTR mutations are relatively frequent in a Quebec population with an elevated incidence of cystic fibrosis. Am J Med Genet 42:360–364

Scheffer H, Verlind E, Penninga D, Meerman G te, Kate LP ten, Buys CHCM (1989) Rapid screening for ΔF508 deletion in cystic fibrosis. Lancet II:1345–1346

Sulisalo T, Francomano CA, Sistonen P, Maher JF, McKusik VA, Chapelle A de la, Kaitila I (1994) High-resolution genetic mapping of the cartilage-hair hypoplasia (CHH) gene in Amish and Finnish families. Genomics 20:347–353

Vogel F, Motulski AG (1986) Human genetics, 3rd edn. Springer, Berlin Heidelberg New York

Zielenski J, Bozon D, Kerem B, Markiewicz D, Durie P, Rommens JM, Tsui L-C (1991a) Identification of mutations in exons 1 through 8 of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Genomics 10:229–235

Zielenski J, Markiewicz D, Rininsland F, Rommens JM, Tsui L-C (1991b) A cluster of highly polymorphic dinucleotide repeats in intron 17b of the CFTR gene. Am J Hum Genet 49:1256–1262

I-4 Association and haplotype sharing due to Identity by Descent, with an application to genetic mapping.

Martin A. van der Meulen* and Gerard J. te Meerman*

In: Genetic mapping of disease genes, edited by JH Edwards, IH Pawlowitzki and E Thompson, Academic Press, to appear April 1997.

*Department of Medical Genetics
University of Groningen
A. Deusinglaan 4
9713 AW Groningen
The Netherlands

phone: +31 50 3632925
fax: +31 50 3632947
email: M.A.van.der.Meulen@med.rug.nl G.J.te.Meerman@med.rug.nl

Running title: Haplotype Sharing Analysis

Key words:
Identity by Descent,
haplotype sharing,
genomic screening
genetic drift
gene mapping
complex diseases

Abstract

By tracing individual alleles drifting through many generations and simulating recombination, the degree of genomic similarity surrounding alleles that are Identical by Descent (IBD) can be estimated. It appears that carriers of alleles that coalesce up to 60 generations ago will share in average about 5 cM of DNA, with a standard deviation of 8 cM. This largely explains the numerous reports on large shared genomic areas surrounding a rare mutated allele.

Genomic sharing can be detected both by association between genetic markers, as single alleles or haplotypes, and disease. To locate genes we propose as Haplotype Sharing Statistic the standard deviation of the length of the shared haplotype segments between all pairs. Methodological difficulties due to small founder populations are outlined. Effects of different population parameters on the power of Identity by Descent (IBD) mapping are investigated by simulation.

A generally applicable design for genomic mapping is proposed in two stages: initial finding of promising genomic locations using patients from a small geographic area, with subsequent confirmation in a larger population.

Introduction

Fishers seminal book on the genetical theory of natural selection has been written already in 1930. Fisher gives an account of the variability involved in genetic drift and deduces upper bounds for the number of copies present of mutations, depending on their age and the population growth. There are two important conclusions with respect to genetic mapping through association analysis that can be drawn from his work. The first aspect is that recent mutations cannot be present in a population numerous enough for associations on the scale of a larger population. If, due to multifactorial inheritance, low penetrance and strong selection against mutations, only very few copies of predisposing alleles can be observed through patients, it will be very difficult to detect them by any form of association analysis. The second aspect is that when mutations are old, they can be very numerous on the scale of a population, but they will descend from a very early predecessor and, due to recombination, only a small area surrounding a mutation will be conserved. This type of linkage disequilibrium has been widely observed between rare disease alleles and anonymous genetic markers, e.g. in Cystic Fibrosis (Maciejko et al., 1989, Morral et al., 1994, Kerem

et al., 1989) and Huntingtons Chorea (MacDonald et al., 1992). This has apparently led to the expectation of many authors (e.g. Plomin et al., 1994) that a very fine mesh of genetic markers, at the 1 cM level or less, will be required to find low-penetrance alleles by studying association between markers and disease. Recently however a few claims have been presented that genomic sharing surrounding a mutated allele that is Identical by Descent, can be detected with initial genomic screens that are much wider than one or two centimorgan (Houwen et al., 1994). Empirical support for this idea comes from various sources. Linkage disequilibrium in the form of haplotype sharing over large genomic distances (5-15 cM) has been reported several times e.g. Melanoma (Cannon-Albright et al., 1994, Gruis et al., 1994), Polyposis Coli (Nystrom-Lahti et al., 1994), Bric (Houwen et al., 1994), Diastrophic dysplasia (Hastbacka et al., 1994), Cartilage-Hair Hypoplasia (Sulisalo et al., 1994), IgE (Meyers et al., 1994) and the A455E CF mutation (De Vries et al., 1996)

The purpose of this contribution is twofold:

1. Analysis of the combined effect of genetic drift and recombination with regard to genomic sharing surrounding alleles that are Identical by Descent (IBD),
2. Presentation of a statistical method to assist in systematic genome wide haplotype comparison as required for IBD mapping.

Simulation method for genetic drift and recombination.

We have shown elsewhere that the conserved area surrounding a common allele of a gene has a quite large variance, leading to a sizeable probability of extended haplotype sharing between carriers of the same disease allele (Te Meerman et al., 1995). Nevertheless, the empirically observed size of the shared area between apparently unrelated individuals is so large that it can only be explained as hidden consanguinity. We have investigated the process of genetic drift and recombination by following alleles descending from 60 generations ago (about 1500 years) as this is a realistic time frame for the introduction of alleles in many populations, that have remained relatively stable since then. Assuming a population growth of 5.5% per generation, a population increase of a factor 20 is present. Family size has been simulated by assuming a quasi-geometrical distribution, according to Lotka (1930), as cited by Feller (1957). This distribution fits the number of offspring in the american population, as cited by Feller (page 130, problem 11, 1957). This leads to a probability distribution with the probability to have N copies in the next generation P(N)

defined as:

$$P(k) = \frac{2*\alpha*p^k}{(2-p)^{k+1}} \quad , K>1$$
$$P(0) = 0.56, \alpha = 0.2, p = 0.736$$

This distribution has a larger variance than the poisson distribution, which implies that genetic drift in human populations will be stronger than expected under the poisson model used by Fisher (1930) for plants. There is an extinction probability after 60 generations of 96%, hardly different from the expectation according to Fisher when the population is stationary and the number of offspring is poisson distributed: 96.7%. Conditional on survival of an allele, this leads to an expected number of copies of 440, as the combined effect of drift and population growth. The distribution of the number of copies is almost exponential, as predicted by Fisher (1930).

The coalescence time is the number of generations we have to go back in order to find a common predecessor. The larger this coalescence time is, the less genomic sharing surrounding an allele IBD can be expected between carriers of an identical allele. We have statistically evaluated the coalescence time in two ways:
as the time to first coalescence and as the total number of meioses connecting all observed copies. We have further calculated the expected size of genomic overlap between all individuals sharing the same allele IBD, from the number of meioses between each pair. The results are summarized in table 1.

Table 1.
Number of generations to first coalescence, meiotic count per individual and expected sharing, as function of selection ratio. Simulation assumptions: 60 generations, 5,1% growth per generation, quasi exponential distribution for the number of offspring. 10622 simulations to find 440 replicates of surviving alleles.
Expected standard deviations are computed as the mean of standard deviations within simulations. The results are weighted for the number of copies present in the population, as we expect to sample proportional from sets of alleles proportional to that number.

selection ratio	generations to first coalescence (Exp. St. dev)	meiotic count per individual	expected sharing all pairs of alleles (cM)
100%	1.4 (1.2)	3.1 (.28)	4.7 (8.4)
20%	3.3 (3.8)	7.9 (1.1)	4.7 (8.4)
10%	5.4 (5.7)	11.6 (1.8)	4.7 (8.3)
5%	8.7 (8.0)	16.5 (3.2)	4.7 (7.9)

It appears (data not shown) that the meiotic count per individual is almost constant irrespective of the number of copies to which an allele actually drifted. This is also apparent from the fact that the meiotic count per individual does not change when weighted for the number of copies present and has a low standard deviation. The standard deviation for the number of generations to first coalescence is very high, within and between simulations, indicating that there are generally quite early common predecessors and much older ones.

The number of generations to first coalescence is directly related to the largest genomic overlap found between individuals with the same allele IBD. If two alleles coalesce in a predecessor 10 generations ago, there is an expected genomic overlap of 2x100/20=10 cM (for a derivation see the appendix). The size of this overlap has a large standard deviation: 5.8 cM (theoretical calculation, see appendix). The expected mean sharing when weighted for the number of alleles IBD from a predecessor is almost constant (4.7 cM), but with a very high expected variance (standard deviation about 8.0 cM).

An explanation for the perhaps unexpectedly low meiotic count is the following. At a selection ratio of 5%, the most recent common founder is present in average at generation 18.5, with a standard deviation of 13.4. The growth therefore takes place in on average 42 generations. If 5% of the alleles are observed, this amounts to a growth from 1 to 440 * 0.05 = 22 alleles. Assuming that growth has been proportional over all generations, the meiotic count is part of an infinite series, with multiplication factor $0.93=(1/22)^{(1/42)}$, and as first term 22. The sum of this infinite series equals 288, which gives a meiotic count per observed individual of 288/22=13.1. This is even less than what is actually observed (16.5). Assuming 60 generations of growth we expect an average meiotic count per individual of 18.9, larger than observed.

Haplotype drift-recombination equilibrium.

Haplotypes inherit as rare alleles, subject to random extinction due to genetic drift and creation because of recombination. This has the consequence that although alleles generally do not disappear from the population through genetic drift, the number of observed combinations of alleles in haplotypes varies considerably compared to the number of expected combinations. This phenomenon can be described as drift-recombination equilibrium, because eventually there will be an equilibrium between haplotypes that disappear because of drift but are also created due to recombination. We therefore agree with Kaplan et al. (1995), that

equilibrium models are almost certainly not appropriate for rare human diseases.

The haplotype drift-recombination equilibrium can be observed by simulation of gene drop in a founder population and computing chi-square values for the observed number of haplotypes, versus the expected number on the basis of linkage equilibrium for observed allele frequencies. Figure 1 shows the average value of chi-square, divided by the number

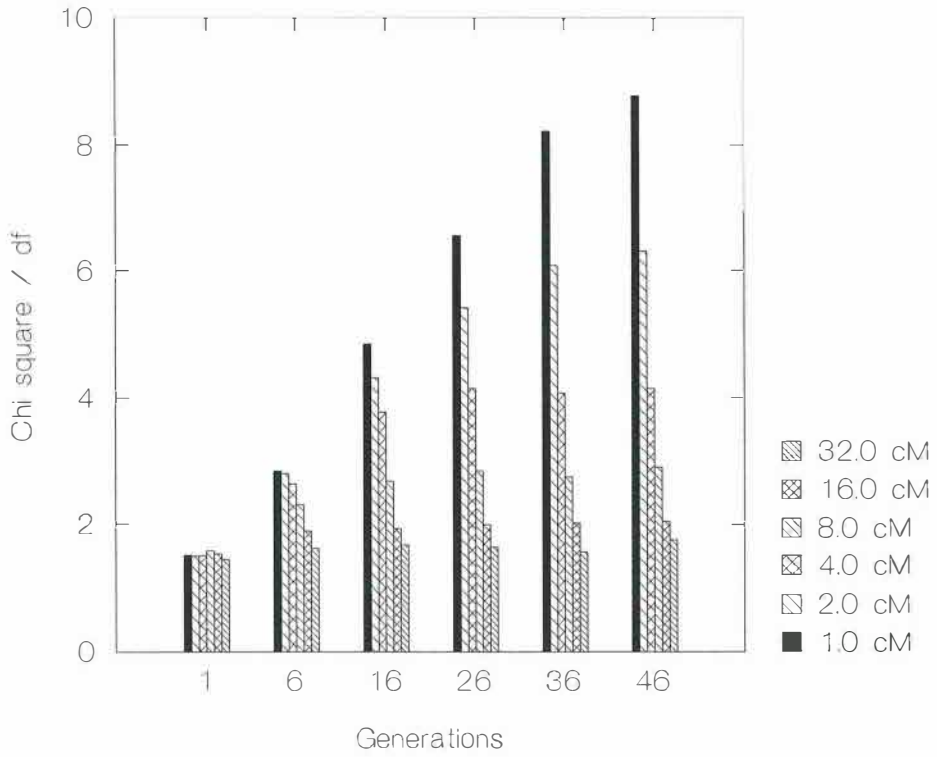


Figure 1.
Linkage disequilibrium measured by chi square divided by the degrees of freedom for variable number of generations random breeding and variable recombination distances in a sample of 150 persons. The original populations had 1000 founders.

of degrees of freedom as observed in a sample of 150 individuals, drawn from a growing population of 1000 founders depending on the number of generations random breeding. For a recombination frequency of 8% an equilibrium situation is reached after 16 generations, while the chi-square value does still increase after 50 generations for a recombination frequency of 2% and lower. These results are obtained under the same conditions as described later in detail in what is called 'the standard simulation'. In much larger

populations (data not reported here), the chi-square in a comparable sample is still higher than expected, but the effect becomes smaller. Analysis of linkage disequilibrium is therefore suitable to show the presence of founder and drift effects. Empirical support for this phenomenon comes from Peterson et al. (1995). When analyzing haplotype sharing there is no method to discriminate between sharing due to common disease alleles and due to random Identity by Descent. The test for excess haplotype sharing will use linkage equilibrium as null hypothesis, but in human populations this assumption may not be justified. The consequence is that P values are too low.

Definition of the Haplotype Sharing Statistic.

Haplotype overlap is computed starting from each marker locus, between all pairs of haplotypes. As long as alleles are the same, comparison between haplotypes is continued to either side. The Haplotype Sharing Statistic (HSS) is calculated as the standard deviation of the shared distance between haplotypes, because we expect that excess haplotype sharing will be observed as a few very large elements among many that are zero or almost zero. This criterion is an alternative for a previously proposed discrete criterion of sharing of three or more 3 locus haplotypes (Te Meerman et al., 1995). The HSS includes comparison of the two independent haplotypes within a person.

$$HSS=\sqrt{\frac{\sum_{i*j} dist_{ij}^2-((\sum_{i*j} dist_{ij})^2/(N*(N-1)))}{((N*(N-1))-1)}}$$

dist = calculated shared distance between two independent haplotypes

N = number of haplotypes

(N * (N-1)) = number of entries in matrix of calculated shared distance between two haplotypes

The absolute value of the HSS statistic described above is sensitive to the degree of heterozygosity of markers and the map position of the marker on the chromosome: at the telomere we expect lower values, because haplotype sharing can extend only in one direction. The HSS in the data is compared to the HSS distribution from random redistributions of the observed marker alleles over haplotypes. We have studied many distributions as they arise from multiple distributions of observed alleles and found that a normal distribution gives a

very good fit to the tail of the distribution, as can be expected according to the central limit theorem when a large set of weakly correlated stochastic variables is added. A fairly good impression of the tail of the distribution and therefore of the significance of the observed HSS in the data can already be obtained from a minimal number of times redistributing the observed alleles. 10 distributions is generally quite adequate for simulation studies where only average values are important and where a slight underestimation of statistical power is not very relevant for obtaining an impression of the effect of variations in simulation conditions. There is a correlation of .95 between results obtained from 10 random distributions compared to those from 100 distributions. For analysis of empirical data about 100-400 distributions are sufficient for accurate ordering of the genomic areas of interest. Because P values are computed by using a Monte-Carlo randomization test, the absolute value of the HSS statistic is not directly relevant. The absolute amount of sharing is however relevant to evaluate spurious significances that may occur with minimal excess overlap. The log of these probabilities is useful to display the results as a kind of lod-score statistic.

The computed probability of the Haplotype Sharing Statistic is biased due to the earlier described drift-recombination equilibrium, because the random distribution of alleles causes linkage equilibrium in the Monte-Carlo computation of P values. Therefore the probabilies cannot be interpreted absolutely. The average probability of all markers reflects the linkage disequilibrium in the population. We are however not interested in the absolute probability for a marker, but we are interested in the comparison between marker intervals. We achieve this by ranking of intervals between markers on the basis of the geometric mean of the two P values of the HSS for adjacent markers. The power of the HSS method is evaluated by ranking the probability for the area where the disease is located among other genomic areas. This is a strict criterion, as it often occurs that intervals adjacent to the real position of the disease allele have also low P values. The rank position can be interpreted as proportional to the fraction false positive results, because first areas with a lower P value would be chosen for confirmation. We assume that such further investigations would result in rejecting areas that do not contain the gene, because what initially may be identical by descent appears to be identical by state. This is not necessarily sufficient however, and investigation of another, larger sample, may be required to discriminate reliably between genomic regions.

Description of simulation parameters.

The main properties of the standard simulation from which all variants are simulated has the following characteristics:

- 1000 non-related founders, with assigned marker haplotypes in complete linkage equilibrium, using 10 alleles with frequencies proportional to 10 drawings from a homogeneous distribution. The disease locus is located approx. in the middle of chromosome 1, in between two markers.
- Chromosomes are simulated with a length of 100 cM. Markers are assigned every 5, 10, 15 or 20 cM, resulting in 20, 10, 7 or 5 markers on each chromosome.
- After generating the founder population, random breeding occurred for 10 generations
- The number of children per couple was poisson distributed, with a mean of 2.25 and a maximum of 10. About 20 % of couples had no offspring.
- 20 persons, which were not related for at least the most recent 4 generations (the segregation is not used), were selected from the last generation. In order to be able to detect the most frequent allele at the disease locus, all alleles at the disease locus in the founder population were uniquely numbered and counted in the last generation. 10 of the 20 selected persons were carrying the allele with the highest allele frequency at the disease locus which was declared to be the disease allele. The other 10 patients are randomly selected persons, without the disease allele (phenocopies). We did not evaluate the effect of using non-transmitted haplotypes as controls, which would be indicated in actual empirical investigations, because it offers additional information without much cost. For diseases that manifest themselves at advanced age, such control haplotypes would however need to come from spouse controls, which brings some methodological problems.

The low number of generations, may seem surprising, in view of the previously reported simulations with 60 generations since the introduction of the disease allele. This assumption may however be not too unrealistic when patients are sampled from a very small geographic area. Besides this, the results can be scaled in the sense that the results can be generalized to hold for older populations provided that the map distance is reduced. The simulations reported here should primarily be seen as a methodological exercise, to obtain a better understanding of the problems involved in applying haplotype sharing as a statistical method. In the analysis, only phase known haplotypes of all markers in patients are used. To investigate the factors influencing the power of IBD mapping we have simulated variants of

the standard simulation.

Results of the Haplotype Sharing Mapping method.

1. Data analysis

As an example of empirical data analysis, figure 2 shows results of the analysis of one simulation under standard simulation conditions, with a marker spacing of 10 cM. In figure 2 the HSS in the data, the average HSS in the random distribution of observed alleles and the final result the - 10 base logarithm of the P value are shown. Markers 1 to 10 are on the first chromosome, while the disease gene is located between markers 5 and 6 (figure 2a). Markers 11 to 20 and 21 to 30 are on chromosomes 2 and 3 respectively (figures 2b and c). Other chromosomes are not shown. Note the regular shape of the HSS of the random distributions over all chromosomes, which is the mean HSS of the specific marker over 100 random distributions of observed alleles. In contrast to the - log P value, the HSS in the data and in the random distribution of alleles are not invariant for chromosomal position. The standard deviation of the 100 random distributions of alleles is almost constant over markers. Figure 2 shows the results for markers, but we are interested, especially in the ranking of the interval between markers 5 and 6, on chromosome 1. This area is the most promising area.

2. Sensitivity of the results for changes in assumptions

Figures 3 to 8 show the results of the sensitivity of the HSS statistic to changes in the parameters used for simulation of the data, by showing the rank position and the standard deviation of the estimate. Except for the parameter varied all parameters are as in the standard simulation. Probability values are computed from 10 Monte Carlo distributions of observed alleles. Each simulation is repeated 100 times. In the figures the standard simulation is always given in black.

Figure 3 shows the results when the number of generations of random breeding is varied from 8-14 generations.

As expected the size of the shared haplotypes decreases when the number of generations increases (Te Meerman et al., 1995) and therefore the rank position increases.

In figure 4 the size of the founder population is varied from 100-2000 persons.

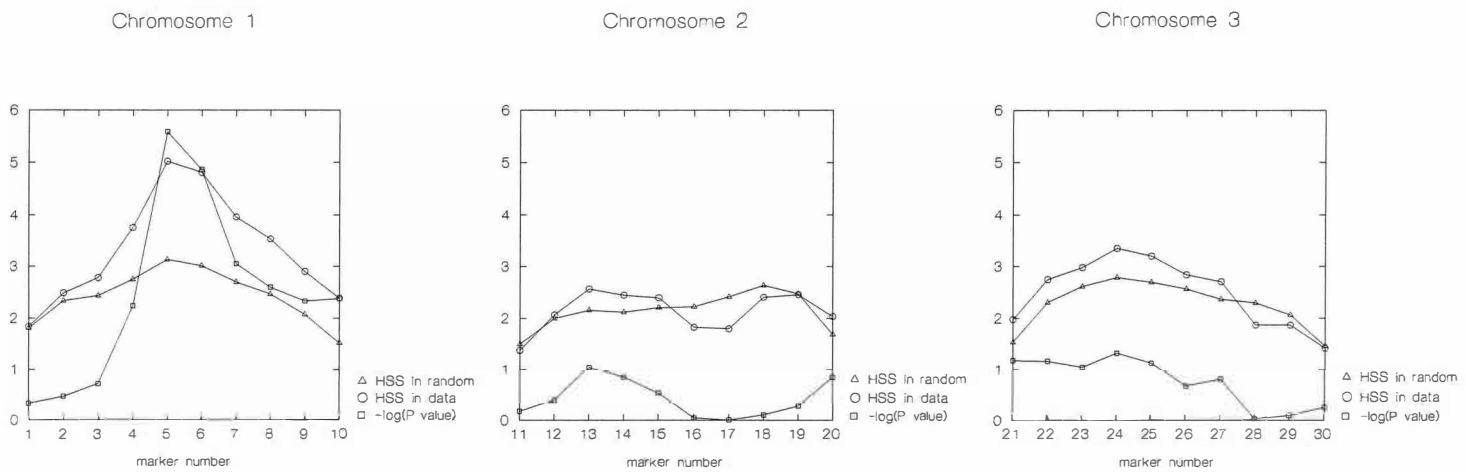


Figure 2a, b and c.

Standard deviation HSS of shared length between haplotypes for data and mean HSS, for 100 random distributions, in cM and the resulting -base 10 log P value. Results are shown for markers on chromosome 1 (markers 1 to 10, figure 2a), 2 (markers 11 to 20, figure 2b) and 3 (markers 21 to 30, figure 2c). The disease gene is located between markers 5 and 6 on chromosome 1. Data from a simulation under standard conditions (see text), with marker spacing 10 cM.

In case of small founder populations, random drift causes (random) IBD. Such sharing is indistinguishable from sharing due to common disease alleles, except that sharing of an allele implies systematic overlap of haplotypes.

In figure 5 the homogeneity level is varied between 0% (no systematic genetic effect) to 100% (all diseased individuals have the same allele).

Extreme heterogeneity causes power to drop excessively, the average ranking of a locus when there is no gene equals 50 %. Starting from 50% homogeneity, the power increases rapidly.

In figure 6 the number of patients used is varied from 10-40.

More patients give a better rank position for the disease locus. In figure 7 the initially present number of alleles per marker: 4, 6 and 9 equiprobable alleles, and the standard 10 random distributed alleles

Random drift causes an exponential distribution of allele frequencies. Apparently the initial number of alleles per marker and their distribution is relatively unimportant.

In figure 8 the average number of children per couple is 2.0-2.25-2.5.

The effect of the rate of increase of the population on false positives is negligible.

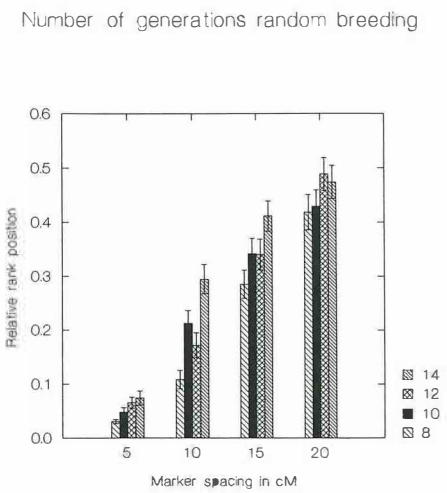


Figure 3.
Average relative rank position and st. dev. after variable number of generations random breeding for marker spacing of 5, 10, 15 and 20 cM.

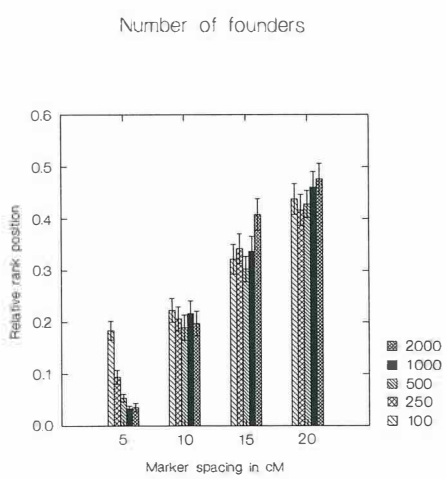


Figure 4.
Average relative rank position and st. dev. dependent on the number of founders of the population for marker spacing of 5, 10, 15 and 20 cM.

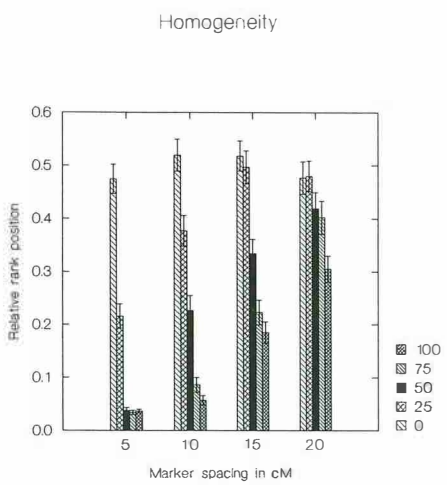


Figure 5.
Average relative rank position and st. dev. dependent on the homogeneity level of the disease for marker spacing of 5, 10, 15 and 20 cM.

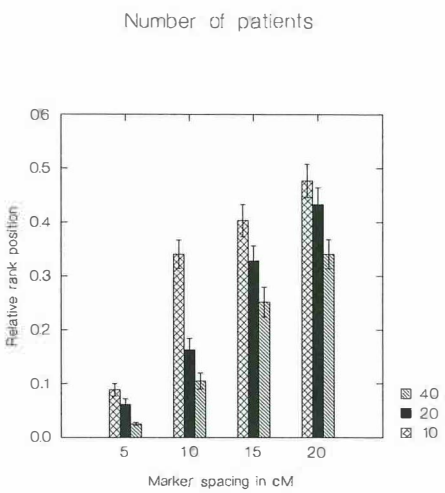


Figure 6.
Average relative rank position and st. dev. dependent on the number of selected patients (50% heterogeneity) for marker spacing of 5, 10, 15 and 20 cM.

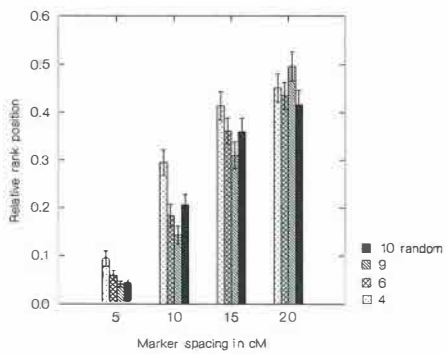


Figure 7.
Average relative rank position and st. dev. dependent on the number of alleles per marker: 4,6 and 9 equiprobable alleles, and the standard 10 random distributed alleles, for marker spacing of 5, 10, 15 and 20 cM.

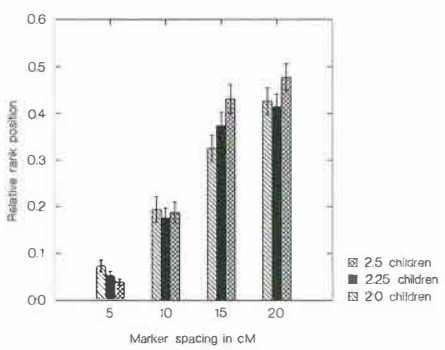


Figure 8.
Average relative rank position and st. dev. dependent on the average number of children per couple (2.0, 2.25 and 2.5) for marker spacing of 5, 10, 15 and 20 cM.

We also investigated (not shown) the effect of the number of generations that selected persons are not related (3 to 6 generations). Influence of this factor on rank position was minimal. Results from phase unknown data are also not shown, because the power drops very much, even taking into account that the patients sample can then be doubled to compensate for not having to type a relative to determine phase.

Discussion

There are three effects of genetic drift which determine the success of IBD mapping. The first effect of genetic drift is to cause reduction of heterogeneity by random elimination of genes predisposing to disease. Secondly genetic drift may also result in elimination of most independent descendency lines from the founder and thus cause more recent predecessors for the disease gene(s). Thirdly haplotype drift may cause apparent excess haplotype sharing at genomic locations unrelated to the disease. The first two effects should be as strong as possible, while the third effect should be as weak as possible. The best situation can be expected for rare disease alleles in large populations, which give a high relative risk.

The expected genomic sharing and its variance between carriers of the same copy of an allele is considerable. This applies to the situation that can still be found in many places of the world where populations have not mixed very much for at least 2000 years. In selecting populations to study diseases, one should concentrate on populations where substantial genetic drift will have contributed to high expected numbers of surviving copies of alleles. Population growth is as important as the probability for a gene to disappear because of genetic drift, because a long period of population growth may have caused alleles that confer a selective disadvantage still to be present in sizeable numbers. Loss of alleles due to genetic drift is highly influenced by population bottlenecks: periods with floods, famine, epidemics and wars.

By comparing all pairs of individuals for marker and haplotype sharing, indications of genomic regions where a disease gene is located can be obtained. Statistical analysis should be seen as complementary to genetic analysis. If markers appear identical, additional investigation is sometimes required to make sure that identity by descend rather than by state is present.

The crucial aspect for statistical power is the number of alleles that is identical by descent and the number of meioses between them. This number is negatively related to the

complexity of a disease (low penetrance, more genetic and environmental factors), mixture of populations, incomplete ascertainment, sampling of patients from a large geographic area and uncertain phenotypes. Increasing the number of patients from one small geographic region will be very advantageous for initial mapping, because geographic and genetic drift concur and lead to shorter coalescence times.

The size of the expected genomic overlap is so large that genomic screens that are presently feasible will be able to detect it, given a sufficient number of alleles that is IBD. Because the variance is so high, a strategy with an initial screen of even 10 cM could already be successful, although false signals will be present. Identity by descent mapping is an extension of statistical mapping methodology, because it offers the possibility to identify genomic regions involved in disease development, using apparently unrelated individuals, to whom segregation analysis cannot be applied. This makes this method suitable for multifactorial diseases and/or diseases with low penetrance.

The fact that stochastic factors, that cannot be controlled and that can only indirectly be measured, are important for the success of IBD mapping, makes the interpretation of data analysis difficult. It may be that genomic regions show up, apparently shared above chance level by affected individuals, but that subsequent verification in larger populations does not imply these regions. In founder populations genes that are marginally involved in the causation of disease, may play a more important role, because the population is not segregating for other risk factors.

The strategy we propose for studies using IBD mapping, includes in the first step the identification of 30-100 patients in a founder population from a small geographic region. Genomic screening at the 5-10 cM level will reveal if a weak level of linkage disequilibrium over large genomic distances exists. Those genomic regions that appear to display excess sharing of haplotypes, are investigated with interpolating markers. If the observed sharing appears due to sharing of genomic regions, a larger set of patients and controls from the surrounding population can determine if the relative risk for disease is associated with the presence of specific risk haplotypes in a genomic region. Haplotypes of patients should then show systematic overlap at only one location, pointing to a gene position. Final confirmation is obtained when gene mutations can be identified, associated with elevated risks.

Our results are obtained for single genes, but the high degree of genetic heterogeneity that is introduced, makes the results for a single gene comparable to what can be expected

for a gene contributing to multifactorial disease, especially when genetic drift has resulted in making multifactorial disease less multifactorial because several risk alleles of genes have drifted out the population.

Haplotype sharing analysis is to some degree comparable to association analysis with a very high level of polymorphism. The difference is however that haplotype sharing analysis uses all information with respect to the length of sharing, and is therefore a multilocus method.

Our results indicate that phase information contributes very much to the power of haplotype sharing analysis. Consequently, the power of methods not using phase, as is the case in affected pedigree member methods, should be very reduced compared to variants where phase is used. The reason is that especially in large pedigrees, sharing of large haplotypes is the rule, rather than the exception. The most powerful method of analysis in pedigrees is of course multilocus linkage analysis, using the pedigree relations between affected individuals. Such analysis is often very computing intensive. A rapid prescreening can, however, be performed with the Haplotype Sharing Statistic.

The actual relevance of IBD methods is not only that it offers a new perspective on the possibility of mapping disease genes using an affected only approach, but that it also gives an additional statistical tool to analyze data obtained in other designs, most notably sib-pair designs as used in Asthma (Meyers et al., 1994, Shirakawa et al., 1994), Diabetes (Davies et al., 1994) and Multiple Sclerosis (Wood et al., 1994). The emphasis is then on comparing data between affecteds in contrast to within sibs or small pedigrees.

With respect to the problem whether significant statistical tests give a reliable indication of the degree of involvement of a genomic region in genetic disease, we are quite pessimistic. It is clear that due to haplotype drift-recombination equilibrium computed significance levels go up considerably as the interval between markers is shortened. Statistical tests of the association type, especially those applied to small samples, will in many cases be insufficient to establish beyond doubt involvement of a genomic region in a disease. Increasing probability thresholds, as proposed by Lander and Kruglyak (1995) gives no answer, as the haplotype drift-recombination process may lead to significant haplotype sharing and association. On the other hand, comparison of significance levels between genomic loci, gives a rational approach for further confirmation on larger populations, even though the significance levels as such are not convincing.

Acknowledgements

We are grateful to Lodewijk A. Sandkuijl for useful discussion and comments. This work is supported by the Netherlands Organisation for Scientific Research.

References

Boehnke, M. (1994). Limits of Resolution of Genetic Linkage Studies: Implications for the Positional Cloning of Human Disease Genes. Am J Hum Genet 55:379-390.

Cannon-Albright, L.A., Goldgar, D.E., Gruis, N.A., Neuhasen, S., Anderson, D.E., Lewis, C.M., Jost, M., Tran, T.D. and Nyguen, K. (1994). Localization of the 9P Melanoma susceptibility locus to a 2cM region between D9S736 and D9S171, Genomics 23:265

Davies, J.L., Kawaguchi, Y., Bennett, S.T., Copeman, J.B., Cordell, H., Pritchard, L.E., Reed, P.W., Gough, S.C.L., Jenkins, S.C., Palmer, S.M., Balfour, K.M., Rowe, B.R., Farral, M., Barnett, A.H., Bain, S.C. and Todd, J.A. (1994). A genome wide screen for human type 1 diabetes susceptibility genes. Nature 371:130-136

De la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of Finland. J Med Genet 30:857-865

De Vries, H.G., Van der Meulen, M.A., Rozen, R., Halley D.J.J., Scheffer, H., Ten Kate, L.P., Buys, C.H.C.M. and Te Meerman, G.J. (1996) Haplotype identity between individuals who share a CFTR mutation allele Identical By Descent: demonstarion of the usefulness of the haplotype sharing concept for gene mapping in real populations. Hum Genet 98:304-309.

Donnelly, K.P. (1983). The probability that related individuals share some section of genome identical by descent. Theor Pop Biol 23:34-63

Feller, W. (1957). An introduction to probability theory and its applications. Volume 1, 2nd edition John Wiley and Sons, New York.

Fisher, R.A. (1930). The genetical theory of natural selection. Dover publications, 1st ed. (2nd rev. ed. 1958) New York.

Gruis, N., Sandkuijl, L.A., Bergman, W. and Frants, R.R. (1994). Common 9P Haplotype in Dutch FAMMM Families: PhD thesis N. Gruis, Genetics of the Familial Atypical Multiple Mole-Melanoma Syndrome, Leiden University

Hastbacka, J., De la Chapelle, A., Mahtani, M.M., Clines, G., Reeve-Daly, M.P., Daly, M., Hamilton, B.A., Kusumi, K., Triveldi, B., Weaver, A., Coloma, A., Lovett, M., Buckler, A., Kaitila, I. and Lander, E.S. (1994). The Diastrophic Dysplasia Gene Encodes a novel Sulfate Transporter: Positional Cloning by Fine-Structure Linkage Disequilibrium Mapping. Cell 78:1073-1087

Houwen, R.H.J., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L.A. and Freimer, N.B. (1994). Genome Screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. Nature Genet 8:380-386

Hill, W.G., Weir, B.S. (1994). Maximum-Likelihood Estimation of Gene Location by Linkage Disequilibrium. Am J Hum Genet 54:705-714

Kaplan, N.L., Hill, W.G., and Weir, B.S. (1995). Likelihood Methods for Locating Disease Genes in Nonequilibrium Populations. Am J Hum Genet 1995;56:18-32

Kerem, B.S., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C. (1989). Identification of the Cystic Fibrosis gene: Genetic analysis. Science 245:1073-1080

Lander, E.S. and Botstein, D. (1986). Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 51:49-62

Lander, E.S. and Botstein, D. (1987). Homozygosity Mapping: A way to Map Human Recessive Traits with the DNA of Inbred Children. Science 236:1567-1570

Lander, E.S. and Schork, N.J. (1994). Genetic dissection of Complex Traits. Science 265:2037-2048

Lander, E.S. and Kruglyak, L. (1995) Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat Genet 11:241-247

MacDonald, M.E., Novelletto, A., Lin, K., Tagle, D., Barnes, G., Bates, G., Taylor, S., Alitto, B., Altherr, M., Myers, R. et al (1992). The Huntington's disease candidate region exhibits many different haplotypes. Nature Genet 1:99-103

Maciejko, D., Bal, J., Mazurczak, T., Te Meerman, G., Buys, C., Oostra, B. and Halley, D. (1989). Different haplotypes for cystic fibrosis-linked DNA polymorphisms in Polish and Dutch populations. Hum Genet 83:220-222

Meyers, D.A., Postma, D.S., Panhuysen, C.I.M., Xu, J., Amelung, P.J., Levitt, R.C. and Bleecker, E.R. (1994) Evidence for a locus Regulating Total Serum IgE Levels Mapping to chromosome 5. Genomics 23:464

Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon-Mateeva, R., Macek, Jr M., Kalaydjieva, L., Angelicheva, D., Dancheva, R., Romeo, G., Russo, M.P., Garnerone, S., Restagno, G., Ferrari, M., Magnani, C., Claustres, M., Desgeorges, M., Schwartz, M., Novelli, G., Ferec, C., De Arce, M., Nemeti, M,. Kere, J., Anvret, M., Dahl, N. and Kadasi, L. (1994). The origin of the major cystic fibrosis mutation (DF508) in European populations. Nature Genet 7:169-175

Nystrom-Lahti, M., Sistonen, P., Mecklin, J.-P., Pylkkanen, L., Aaltonen, L.A., Jarvinen, H., Weissenbach, J., De la Chapelle, A., and Peltomaki (1994). Close linkage to chromosome 3p and conservation of ancestral founding haplotypes in hereditary nonpolyposis colorectal cancer families. Proc Natl Acad Sci USA 91:6054-6058;

Olson, J.M. and Wijsman, E.M. (1994). Design and Sample-Size Consideration in the Detection of Linkage Disequilibrium with a Disease Locus. Am J Hum Genet 55:574-580

Peterson, A.C., Rienzo, A.D., Lehesjoki, A.E., De la Chapelle, A., Slatkin, M. and Freimer, N.B. (1995) The Distribution of Linkage Disequilibrium over anonymous genome regions. Hum Mol Genet 4:887-894

Plomin, R., Owen, M.J., McGuffin, P. (1994). The genetic basis of complex human behaviors. Science 264:1733-1739

Puffenberger, E.G., Kauffman, E.R., Bolk, S., Matise, T.C., Washington, S.S., Angrist, M., Weissenbach, J., Garver, K.L., Mascari, M., Ladda, R., Slaugenhaupt, S.A. and Chakravarti, A (1994). Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 3:1217-1225

Risch, N. (1990). Linkage Strategies for Genetically Complex Traits. II. The Power of Affected Relative Pairs. Am J Hum Genet 46:229-241

Risch, N. (1991). A note on Multiple Testing Procedures in Linkage Analysis. Am J Hum Genet 48:1058:1064

Romeo, G., Devoto, M. and Galietta, L.J. (1989). Why is the Cystic Fibrosis gene so frequent? Hum Genet 84:1-5

Sherrington, R., Melm8er, G., Dixon, M., Curtis, D., Mankoo, B., Kalsi, G. and Gurling, H. (1991). Linkage Disequilibrium between two highly Polymorfic Microsatellites. Am J Hum Genet 49:966-971

Shirakawa, T., Li, A., Dubowitz, M., Dekker, J.W., Shaw, A.E., Faux, J.A., Ra, C., Cookson, W.O. and Hopkin, J.M. (1994). Association between atopy and variants of the beta subunit of the high-affinity immunoglobulin E receptor. Nature Genet 7:125-129

Slatkin, M. (1994). Linkage Disequilibrium in Growing and Stable Populations. Genetics 137:331-336

Sulisalo, T., Francomano, C.A., Sistonen, P., Maher, J.F., McKusick, V.A., De la Chapelle, A., and Kaitila, I. (1994). High-Resolution Genetic Mapping of the Cartilage-Hair Hypoplasia (CHH) Gene in Amish and Finnish Families. Genomics 20:347-353

Te Meerman, G.J., Van der Meulen, M.A., and Sandkuijl, L.A. (1995). Perspectives of Identity by Descent (IBD) mapping in Founder Populations. Clin and Exp Allergy 25:(suppl 2)97-102

Thomson, G. and Klitz, W. (1987). Disequilibrium Pattern Analysis. I. Theory. Genetics 116:623-632

Vogel, F., Motulski, A.G. (1986). Human Genetics, 2'ed. Springer.

Wood, N.W., Holmans, P., Clayton, D., Robertson, N., Compston, D.A. (1994). No linkage or association between multiple sclerosis and the myelin basic protein gene in affected sibling pairs. Neurol Neurosurg Psychiatry 57:1191-1194

Appendix.

The probability that no recombination occurs to the centromeric or telomeric side of an allele in N meioses, at a genomic distance where the probability of recombination is X, is equal to

$$(1-X)^N$$

The corresponding probability distribution for the size of the area where no recombination occurs is

$$N\times(1-X)^{N-1}$$

The expected size of this area is, using partial integration

$$\int_0^1 X\times N\times(1-X)^{N-1}dX=1/N$$

The variance of the size can be computed from the integral

$$\int_0^1 X^2\times N\times(1-X)^{N-1}dX-\left(\frac{1}{N}\right)^2$$

Partial integration gives as result, valid for N > 3:

$$var(X)=\frac{N^2-3\times N-2}{N^2\times(N+1)\times(N+2)}$$

The expectation and variance of the addition of the telomeric and centromeric part is twice that computed above. Interference has been neglected, because recombinations may take place at different meioses.

I-5 General discussion and Summary

Mapping the genes involved in diseases with a complex mode of inheritance is complicated. Identity by Descent (IBD) mapping is an extension of the methods available for mapping of genes involved in complex disease inheritance and can be used when affecteds with or without proven common ancestry in a founder population are observed. An advantage of mapping disease genes in a founder population is the expected reduced complexity of disease inheritance due to drift, which is likely to reduce heterogeneity and possibly higher penetrance due to fixation of mutation(s) involved in disease expression.

Sharing over extended segments of DNA surrounding predisposing genes inherited from a common ancestor can be expected. Confirmation of the prediction that the length of sharing around IBD mutations can be quite large is obtained from a study where the surrounding haplotypes are studied in known carriers of the A455E CF mutation. Similar results have been shown in many other studies. In IBD mapping studies distinction must be made between coalescence time, representing the number of meioses connecting two patients, and the total meiotic count between all patients. The first is a measure of the expected haplotype overlap between two affecteds and is therefore an indication of the marker spacing necessary in the initial genome screen. Short coalescence time results in large genomic overlap. The total meiotic count is a measure for the expected genome overlap between all carriers surrounding the shared mutation.

In simulation studies is shown that the proposed Haplotype Sharing Statistic (HSS) is effective in localizing the map position of disease genes under different conditions. Most notable is that the HSS is still effective when a dominant disease, with 50% homogeneity (or 50 % phenocopies), is simulated. Under this condition only 1 out of 4 haplotypes is IBD. The varied simulation conditions show that IBD mapping still works, when conditions are encountered which can be expected in empirical studies.

In many linkage and sibpair studies there is no consistency of marker allele scoring between families. This is not surprising, because markers are used in these studies as a tool to follow segregation, not as a tool for comparison between families. However, the

high informativity of markers make them an excellent tool for following both segregation (heterozygosity of markers is high and heterozygosity of markers is in IBD studies important to determine phase) and comparison between families (or affecteds) for IBD mapping. In IBD mapping studies it is important that all markers are typed for all affecteds and relatives for phase determination. Algorithms to handle unknown phase for some markers and for missing markers for some persons are necessary for empirical studies.

Many recent papers have addressed the problem of multiple testing and the effect on false positive rates in mapping studies. The problem of false positives can easily be handled by confirmation of the positive score in another population. This is often not possible, because all available patients are included in the initial study. In some mapping methods the false positives can also be the result of haplotype drift-recombination equilibrium. Note that in IBD mapping studies false positives can be divided in Identity by State (IBS), which can easily be detected by adding intervening markers, and IBD. Multiple testing is in our opinion not a big problem, as long as positive results are confirmed in other populations or by adding more patients (distant relatives) from the same population. Confirmation of positive results is easier in IBD mapping studies, because many more affected individuals are available, since single affecteds can be used. Confirmation in another population might lead to the same mutation, with the same surrounding haplotype, a different mutation at the same locus, or no confirmation, possibly due to heterogeneity and therefore does not lead to rejection.

In the search for genes involved in (complex) diseases, the power of a study can be calculated given certain assumptions on the etiology of a disease. The number of patients necessary to have enough power can be calculated and therefore the effort can be minimized. This is especially important for studies where complete genome screens are performed at the 10 cM level (or less). In IBD studies the power and the optimal spacing of markers can be estimated assuming a coalescence time between affecteds. Minimizing of the number of patients is also effective in IBD studies, because affecteds with the same phenotype can be selected and the geographical region from which affecteds are selected can be minimized.

Comparable to IBD mapping studies are linkage studies, where affecteds are known to be related in a certain degree. Calculations of lodscores are very elaborate when untyped persons over multiple generations are included in the calculation. In these studies it is assumed that the pedigree found is the only and true connection between the affecteds and therefore the mutation(s) leading to affection status have segregated along these pedigree lines. This assumption can lead to erroneous results, when unaffected parents are typed for markers and the affected person inherited the disease mutation through the parent, who is not the connecting parent in the pedigree. A study in the founder population of French Canada has shown that many common ancestors are found for affected individuals (Heyer and Tremblay, 1995).

In studies to map disease genes the consistency of phenotypes over patients is important, because genetic heterogeneity between patients, possibly involved to explain the different phenotype, will reduce the power of the study. On the other hand variability of phenotype in diseases with a complex inheritance is likely, because affecteds who are selected from a founder population may share one of the mutated genes involved in the disease, while they do not share another mutated gene involved in the disease, this may lead to a (slightly) different phenotype. In comparison to affected sibpair studies, selection of affecteds for IBD studies is easier, because single affecteds will be more frequent.

After successful mapping of disease genes in IBD studies, the mode of inheritance must be determined with care. For instance when our study for haplotypes surrounding the A455E CF mutation had been performed in order to map the CF gene using a 10 cM genome screen, the gene would probably have been found, because large overlap of haplotypes was present between affecteds, and would have resulted in fine mapping, using intervening markers, because enough recombinations are available. The mode of inheritance, however, would in first instance erroneously have been determined as dominant. When relatives carrying also the mutation are typed, reduced penetrance is an easy explanation. Careful comparison of the other (non A455E) haplotype in affecteds would have show overlap between affecteds and the mode of inheritance would change to recessive. Note that homozygosity mapping (Lander and Botstein, 1986, 1987) does not work in this mapping study of a recessive disease. Homozygosity mapping can only be

used in studies for recessive diseases where affecteds are likely to be homozygous due to close consanguinity. Because the number of meioses observed in homozygosity mapping is low, the power for fine mapping is reduced in comparison to IBD studies.

The IBD mapping technique can also be used to identify haplotypes surrounding low penetrant mutations for diseases where gene locations are known. This is shown in independent Norwegian breast cancer patients, where some carriers of an identical haplotype surrounding BRCA1 were found to be carrier of the same mutation (Dorum et al, submitted). Estimation of penetrance of low penetrance mutation can be performed correctly only on non affected proven carriers of the mutation, rather than on carriers of the haplotype. The same haplotype as the haplotype surrounding the low penetrant mutation may be present in the population without the mutation. Proof of IBD status of this haplotype is not proof of the carrier status for this mutation, because although haplotypes are IBD this copy might be a copy of this haplotype before the mutation occurred. Note that it is most likely that a mutation occurs at the most frequent haplotype.

When genes involved in complex diseases have been mapped, their function can lead to understanding of the mechanism which causes the affection status. This might eventually lead to some form of therapy and/or prevention.

Scoring of point mutations in patients leads to the question of the involvement of the mutation in the disease status. If nonsense mutations are found then there will be little doubt on a functional effect of the mutation. Missense mutations, however, might be a neutral genetic variant and is hard to distinguish from low penetrance mutations.

The modern lab techniques available in genetics lead to large amounts of data. In IBD mapping studies where complete genome screens are performed handling of data becomes very important, because the information contained in haplotype sharing can only be detected after phase determination, looking at multiple marker haplotypes. The HSS is a convenient tool to show the location of areas where haplotype sharing is increased. In the HSS no genetic model is incorporated, therefore regions showing increased sharing must be checked manually.

Conclusion

Identity by Descent mapping is a promising new method to map disease genes involved in diseases with a complex mode of inheritance. Empirical evidence proving that mutations leading to diseases are inherited through common ancestors are numerous in the recent literature. However, IBD mapping can only be performed on disease genes with low mutation rates in founder populations, where mutations must be old to be numerous enough to play a major role in a disease with a complex mode of inheritance.

References

Dorum A, Moller P, Kamsteeg EJ, Scheffer H, Burton M, Heimdal KR, Mehle LO, Hovig E, Trop CG, Van der Hout AH, Van der Meulen MA, Buys CHCM and Te Meerman GJ: Haplotype analysis, a strategy for identifying prevalent mutations, demonstrating a Norwegian BRCA1 founder mutation, submitted.

Heyer E and Tremblay M. Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet 1995; 56:970-978

Lander ES and Botstein D: Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 1986; 51:49-62

Lander ES and Botstein: Homozygosity Mapping: A way to Map Human Recessive Traits with the DNA of Inbred Children. Science 1987; 236:1567-1570

Part II

Mosaicism, a problem in recurrence risk calculation ?

II-1 Introduction

Mosaicism

The origin of mosaicism is defined by Murphy and Chase (1975) as: *The current view is that the mutation is a 'copying error' occurring when replication of DNA is taking place. If so, then such an accident could occur not only during meiosis but during mitosis as well. Once the faulty chromosome has been produced, there would be a strong tendency for it to breed true, so that all cells descended from the mutant line also would be abnormal. But since the development of the body cells is a branching process, it seems evident that the further down the 'tree' the accident occurs, the smaller the proportion of mutant gametes that should be produced.* From this definition a general definition of mosaicism follows: the existence of differing genetic information between cells within an individual as a result of events such as mutation or chromosomal rearrangements (Wijsman, 1991). In recurrence risk calculations, we will focus on the mutations leading to (a proportion of) affected gonadal cells, since only these result in affected offspring.

Although germinal mosaicism had not been conclusively demonstrated before 1971 to occur in humans, it had been demonstrated in laboratory mammals (Fisher, 1930) and is well known in Drosophila melanogaster. Therefore Hartl (1971) concluded that it appears logical to attribute the lack of examples of germinal mosaicism in man to the technical difficulties to prove mosaicism rather than to the nonoccurrence of premeiotic mutation in human gonadal cells. The literature at that moment revealed several examples of diseases which could be interpreted as involving mosaicism. Increasingly molecular analysis and clinical observations suggest that mosaicism may be relatively common and account for a substantial proportion of what is in the context of X-linked lethals called an apparent 'new' mutation (Hall, 1989). Young (1991) gives an overview of 5 implications of mosaicism for genetic counselling. 1) For a condition which is known to show fully regular mendelian dominant inheritance, healthy unaffected parents, who have had an affected child can no longer be confidently reassured of a negligible recurrence risk.
2) when the index person shows a relatively mild form of an autosomal dominant disorder, it may be that this individual is a somatic mosaic, with or without gonadal mosaicism. Therefore the risk to children may be less than 50%, but offspring will have

the disorder in non-mosaic form and might therefore be more severely affected. 3) if a disorder shows heterogeneity, with well documented autosomal dominant and recessive forms, the finding of affected individuals in only one sibship, does not automatically imply that the condition in that family follows autosomal recessive inheritance. 4) If two or more siblings have what appears to be a 'new' autosomal recessive disorder, autosomal dominant inheritance resulting from gonadal mosaicism in one parent cannot be excluded. Parental consanguinity would however be a strong indicator in favour of autosomal recessive inheritance. 5) in sex-linked disorders for which precise carrier tests are not available, ignoring the possibility of gonadal mosaicism is unlikely to result in major errors as the mother of an isolated case will generally be considered to have a relatively high prior probability of being a carrier. However great caution must be exercised if a totally reliable somatic carrier test is available, as a negative result in the mother of an isolated case could suggest a spuriously low recurrence risk given that gonadal mosaicism may be a relatively common occurrence (Bakker et al., 1989).

In the definition of mosaicism by Murphy and Chase the major point of the mosaicism model is introduced: the branching process of cell division. A human being is defined as the result of exponential cellgrowth, starting from one cell. Hartl showed that more complex models than this simple, so called symmetric synchronous dichotomous proliferation, model together with the assumption of a constant mutation rate per cell division do not change the recurrence risks as long as the number of gonadal cell generations is high enough.

Description of the model

When a mutation occurs in a mitosis before the formation of the germline stem cell a somatic mosaic will result, given that the germline stem cell is a descendent from the branch in which the mutation originated. In a somatic mosaic all diploid cells in the germline will carry the mutation, leading to 50% mutation carrier gametes due to the final meiotic cell division. When the mutation happens after the formation of the germline stemcell a germline mosaic will result. This leads to a variable percentage of carrier germline cells, dependent on the germinal generation or 'depth of the tree' in which the mutation has happened. Figure 1 shows a schematic representation of the symmetric

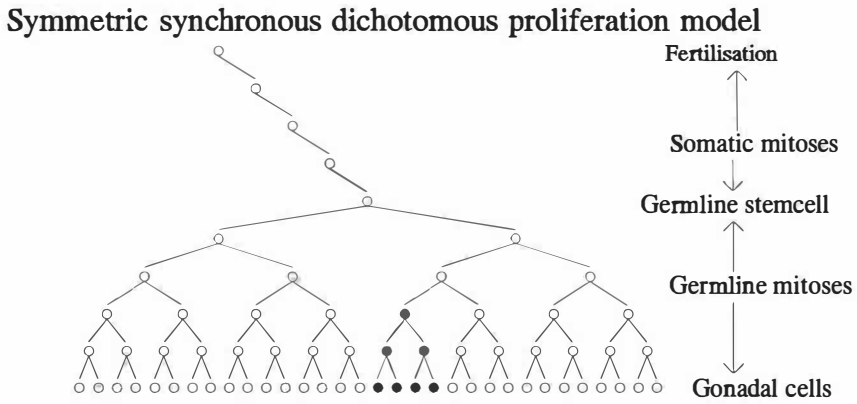


Figure 1

synchronous dichotomous proliferation model. In the germline the branching model is symmetric, because all gonadal cells are the result of the same number of mitoses, synchronous, because all cell divisions occur at the same time, and dichotomous, because each cell division results in two cells. From figure 1 can be seen that a mutation in an early cell generation leads to a higher carrier probability for gonadal cells than a late gonadal cell generation mutation. An example of a late gonadal cell generation mutation is shown in figure 1, where cells that carry the new mutation are given in black. In this model is assumed that oocytes and spermatocytes are randomly drawn from the pool of gonadal cells, that there is no selection against or in favour of cells that carry the mutation and that there is no reverse mutation.

Mode of inheritance

Differentiation between recessive inheritance and a new mutation is hard. However, a condition which is present only once in a large sibship offers some evidence against recessivity. But even in very large sibships the evidence against recessivity derived from a segregation ratio is weak: for example in sibships of 11 the proportion of those who have only one affected child is about 1/7 (Edwards, 1989). When the map location of the disease gene is known, unaffected sibs carrying the same chromosomes give evidence for mosaicism, but are sometimes erroneously seen as non-penetrants, although this is rare in recessive diseases. When unaffected sibs in dominant diseases where mosaicism is

possible are seen as non-penetrant cases, they are falsely seen as carriers. In X-linked recessive diseases both the mode of inheritance and the origin of a new mutation are relatively easy to establish, because of low probability of two independent mutations and high probability of expression in affected boys.

Direct evidence of mosaicism

In principle germline mosaicism can directly be recognized by examination of the germ cells for the mutation found in the affected child. These studies would benefit the consultand, because germline mosaics can be distinguished from somatic mosaics and in germline mosaics a good estimate of the recurrence risk can be obtained, but it would also give more insight in the correctness of models used for recurrence risk calculations. However, these studies are difficult, and rarely ethical, in the female (Edwards, 1989).

Mutation selection equilibrium

In diseases where the mutation frequency is high, the question after the birth of an affected child is: is the disease inherited through a carrier, or is it due to a new mutation? From population genetics can be deduced that for X-linked recessive lethal diseases there will be mutation selection equilibrium, as first described by Haldane (1935). From this mutation selection equilibrium it can be deduced that the mother whose only son is affected, and knows nothing of her family history, has, assuming equal mutation rates in males and females, a probability of 2/3 of being a carrier and 1/3 of not being a carrier. In Haldane's model new mutations occur during meiosis and therefore there is no recurrence risk to future offspring for that specific mutation. Bell and Haldane (1937) point out, however, that new mutations can either happen during meiotic or mitotic cell division, the latter leading to a somatic and/or gonadal mosaic situation. Such mosaics, if male, would probably not be affected, but might transmit the mutant gene to all, some or none of their daughters, depending on the proportion of his testicular cells carrying the mutant gene. A mosaic woman would also be not affected, but might transmit the defect to half or less than half of her children. Grimm et al. (1990) extended the mutation selection equilibrium to include the mitotic origin of the mutation in both the parents and the child. When the mutation is supposed to be of mitotic origin, the probability of a mutation in the parent is much bigger than in the child, since the number of mitoses in

the parents is much higher than in the child, when one realizes that the mutation must have happened early enough, for the child to get the affected phenotype. Figure 2 shows the mutation selection equilibrium for an X-linked recessive lethal disease, where all mutation occur in the parents and mutation frequency for males and females are supposed to be equal to μ for each chromosome. In mutation selection equilibrium the frequency of the disease in the parent generation must be equal to the frequency in the childrens generation. In X-linked lethal disease, the affected males will not produce gametes (offspring), while unaffected males can produce gametes carrying a new mutation. At the level of gametes can be seen that affected males allways receive the mutation from their mother, while there is a 1/3 probability that the mutation is new. Females (heterozygotes), have an equal probability of receiving a new mutation from one of her parents or receiving a mutation from her carrier mother.

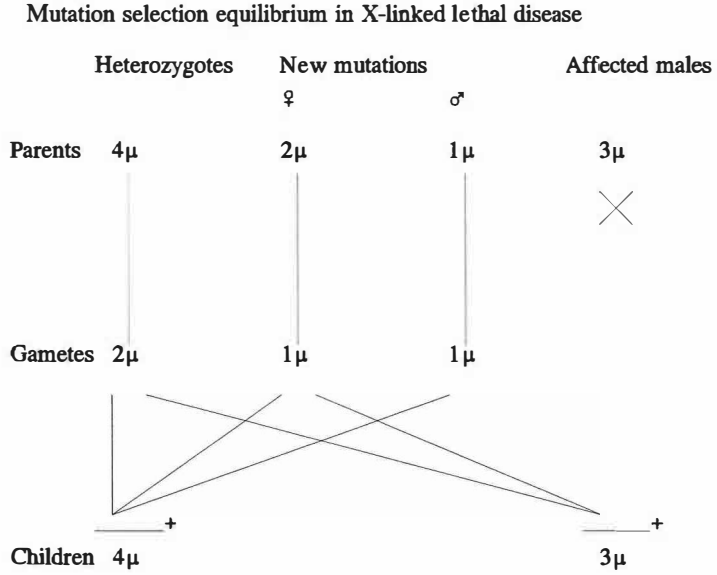


Figure 2

The assumption that all mutations occur in the parents can also be seen as a worst case scenario, in which all observed sporadic cases, previously seen as new mutations, lead to increased recurrence risks. The evidence for the assumption of equality of male and female mutation frequency is weak, although they are not demonstrable on the limited

data available. A priori oogenesis and spermatogenesis are so different that they would not be expected to be equal (Edwards, 1986). Edwards (1989) pointed out that almost all women will have several oocytes with mutations for the commoner genetic disorders, because the number of mitoses necessary to develop the approx. 8 million ova, present at birth, greatly exceeds the individual mutation rate, where the mutation frequency is estimated in the order of 10^{-6}, for most of the common disorders. It will be clear that only early mitoses will lead to a significant proportion of mutant cells. The same principle applies to males, where the number of mitoses in spermatogenesis is larger, but late mutations will lead to low frequencies. A much wider assortment of mutations can therefore be expected to be present in males, which are thought to increase with advancing male age.

Mapping disease genes

Mapping genes involved in diseases with high mutation rates is complicated, because due to the existence of mosaicism there may not be consistent segregation of the transmitted chromosome and the disease status throughout the pedigree. Mosaicism thus leads to falsely assuming double recombinants and therefore to low likelihoods, when mosaicism is not taken into account. Diseases which segregate normally in some of the pedigrees and unusual in others may erroneously be interpreted as heterogeneous, while they are in fact germline mosaics.

Assumptions for recurrence risk calculations

Before recurrence risk calculations can be performed, some general assumptions are necessary, apart from the assumptions made for modelling purposes: 1) no mistaken paternity or other identity problems exist, 2) a mutation in a transmitting parent will not lead to a disease phenotype in that parent, even when it concerns an early somatic mutation; the disease phenotype only occurs in the child, 3) meiotic mutations are not considered in the calculations. 4) No heterogeneity or other problems concerning disease status occur.

Duchenne Muscular Dystrophy: example for X-linked recessive lethal disease

In recurrence risk calculations concerning mosaicism, Duchenne Muscular Disease (DMD) is often taken as an example. DMD is a complete recessive X-linked lethal disease. The following description of DMD only gives the necessary details for recurrence risk calculation. Discussion on severity or clinical aspects of DMD are outside the scope of this thesis.

Location - short arm X-chromosome (Xp21, Davies et al, 1983), gene approx. 12 cM long, 79 exons (Roberts et al, 1993)

Incidence - 1:3500 (Bakker et al., 1989), 1:4200 (Van Essen et al, 1992)

Mutation freq - $7.1 * 10^{-5}$ (Moser, 1984)

Mutations - Deletion hotspots: major: exon 45-55, minor: exon 1-20 (Emery, 1988, Den Dunnen et al., 1989, Kunkel, 1986, Koenig et al., 1987, Wapenaar et al., 1988, Lindlof et al., 1989)

- point mutations throughout the whole gene (Prior et al., 1994)

Mutation - detection possible in about 2/3 of cases (Monaco et al., 1987), 1/3 of cases probably due to small deletions/insertions or point mutations (Roberts et al., 1994)

- detection of 98 % of DMD deletions by PCR (Beggs et al., 1990).

CK values - carrier females can be recognized by bloodtesting for CK levels. Only 5% of non-carrier females has increased CK values, while 1/3 of carrier females shows normal CK values. (Thompson et al., 1967, Zatz et al., 1976, Lange et al., 1979)

- Prenatal CK determination by foetal blood sampling was found to give false negatives (Golbus et al., 1979)

After the successful mapping of the DMD gene, the detection of new mutants led to the determination of their parental origin and provided evidence that both somatic and germline mosaicism occur in DMD (Bakker et al, 1987,1989). Bakker et al. (1989) describe the overall recurrence risk in case of germinal mosaicism to be 7% (14% conditional on inheriting the risk-chromosome). Van Essen et al. (1992) calculate that the recurrence risk due to inherit the new mutation together with the risk-chromosome after the birth of an affected boy equals 20%.

Other diseases

An overview of diseases, suggesting that somatic mosaic mutations in the germ line may be present in phenotypically normal individuals is given by Hall (1988). In all genetic diseases, new mutations can lead to mosaicism. When the mutation frequency is low and/or no selection against carriers of mutations exist, the probability of disease inheritance through carriers becomes high in comparison to the probability of a new mutation. Incorporation of mosaicism in risk/carrier calculations seems only useful in diseases with high mutation frequency and strong selection against mutations in the population.

Outline of thesis

In chapter 2 the recurrence risk formula for germline mosaicism from Hartl (1971) is extended with the possibility to incorporate information on genetic markers in the recurrence risk calculation due to germline mosaicism. The recurrence risk tables as calculated with the extended formula and published in chapter 2 are used to include in recurrence risk calculations the risk due to mosaicism in nuclear families and in extended families in chapter 3. It is also shown how phenotypic information and results from mutation screening can be incorporated in these manual calculations. Although manual recurrence risk calculations including the recurrence risk due to mosaicism becomes possible using the recurrence risk tables by Van der Meulen et al (1995), these calculations are still complicated and results are often surprising and even counter-intuitive. In chapter 4 a computer program, based on a linkage program (Te Meerman, 1991), is presented, in which recurrence risks, including the risk due to germline and somatic mosaicism, can be calculated. This program is also capable of handling recombination and multipoint analysis. The reason to develop this computer program is to have an additional level of checking calculations and the possibility of calculating recurrence risks on pedigrees in which manual calculations are not feasible.

References

Bakker E, Veenema H, Den Dunnen JT, Van Broeckhoven C, Grootscholten PM, Bonten EJ, Van Ommen GJB and Pearson PL. Germinal mosaicism increases the recurrence risk for 'new' Duchenne muscular dystrophy mutations. J Med Genet 1989; 26:553-559

Bakker E, Van Broeckhoven C, Bonten EJ, Van de Vooren MJ, Veenema H, Van Hui W, Van Ommen GJB, Vandenberghe A and Pearson PL. Germline mosaicism and Duchenne muscular dystrophy mutations. Nature 1987; 329:554-556

Barbujani G, Russo A, Danielli GA, Spiegler AWJ, Borbokowska J, Hausmanova Petruswicz I. Segregation analysis of 1885 DMD families: Significan departure from the expected proportion of sporadic cases. Hum Genet 1990; 84:522-526

Bell J and Haldane JBS. The linkage between the genes for colour blindness and heamophilia in man. Proceedings of the Royal Society of London, Series B 1937; 123:119-150

Bunyan DJ, Robinson DO, Collins AL, Cockwell AE, Bullman HMS, Whittaker PA. Germline and somatic mosaicism in an female carrier of Duchenne Muscular Dystrophy Hum Genet 1994 93:541-544

Edwards JH. The population genetics of Duchenne: natural and artificial selection in Duchenne Muscular Dystrophy. J Med genet 1986; 23:521-530

Edwards JH. Familiarity, recessivity and germline mosaicism. Ann Hum Genet 1989; 53:33-47

Fisher RA. Note on a tricolour (mosaic) mouse. J. Genet 1930; 23:77-81

Grimm T, Muller B, Muller CR and Janka M. Theoretical considerations on germline mosaicism in Duchenne muscular dystrophy. J Med Genet 1990; 27:683-687

Grimm T, Meng G, Liechti-Gallati S, Bettecken T, Muller CR and Muller B. On the origin of deletions and point mutations in Duchenne muscular dystrophy: most deletions arise in oogenesis and most point mutations result from events in spermatogenesis. J Med Genet 1994; 31:183-186

Haldane JBS. The rate of spontaneous mutations of a human gene. J Genet 1935; 31:317-326

Hall JG. Somatic mosaicism: observations related to clinical genetics. Am J Hum Genet 1988; 43:355-363

Hartl DL. Recurrence risk for germinal mosaics. Am J Hum Genet 1971; 23:124-134

Jeanpierre M. Germinal mosaicism and risk calculation in X-linked diseases. Am J Hum Genet 1992; 50:960-967

Jeanpierre M. A simple method for calculating risk before DNA analysis. J Med Genet 1988 25:663-668

Lathrop GM, Lalouel JM. Easy calculation of lod scores and genetic risks on small computers. Am J Hum Genet 1984; 36:460-5

Muller B, Dechant C, Meng G, Liechti-Galatti S, Doherty RA, Hejtmanchik JF, Bakker E et al. Estimation of the male and female mutation rates in Duchenne muscular dystrophy (DMD) Hum Genet 1992; 89:204-206

Murphy EA and Chase GA. *Principles of genetic counselling*. Chicago: Year Book Medical Publishers, 1975.

Murphy EA, Cramer DW, Kryscio RJ, Brown CC and Pierce ER. Gonadal mosaicism and genetic counselling for X-linked recessive lethals. Am J Hum Genet 1974; 26:207-222

Ott J. *Analysis of Human Genetic Linkage*. Rev ed. Baltimore and London: The Johns Hopkins University Press, 1991.

Passos-Bueno MR, Bakker E, Kneppers ALJ, Takata RI, Rapaport D, Dunnen JT, Zatz M and Ommen GJB. Different Mosaicism Frequencies for Proximal and Distal Duchenne Muscular Dystrophy (DMD) Mutations Indicate Difference in Etiology and Recurrence Risk. Am J Hum Genet 1992; 51:1150-1155

Passos-Bueno MR, Lima MABO, Zatz M. Estimate of germinal mosaicism in Duchenne muscular Dystrophy. J Med Genet 1990; 27:727-728

Te Meerman GJ. A logic programming approach to pedigree analysis. Thesis publishers Amsterdam, 1991.

Van der Meulen MA, Van der Meulen MJP and Te Meerman GJ. Recurrence risk for germinal mosaics revisited. J Med Genet 1995; 32:102-104

Van der Meulen MA, Te Meerman GJ and Sandkuijl LA. Calculation of Recurrence Risk in case of possible mosaicism, submitted.

Van Essen J, Busch HFM, Te Meerman GJ, Ten Kate LP. Birth and population prevalence of Duchenne muscular dystrophy in the Netherlands. Hum genet 1992 88:258-266

Van Essen J, Abbs S, Baiget M, Bakker E, Boileau C, Van Broeckhoven C et al. Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: a european study. Hum Genet 1992; 88:249-257

Wijsman EM. Recurrence risk of a new dominant mutation in children of unaffected parents. Am J Hum Genet 1991; 48:654-661

Williams WR, Thompson MW, Morton NE. Complex Segregation Analysis and computer-assisted genetic risk assessment for Duchenne Muscular Dystrophy Am J Med genet 1983 14:315-333

Workshop on germinal mosaicism, 9 March 1994 London, Organizers: Caroline Berry and Andrew Wilkie, Invited Speakers: Marc Jeanpierre, Bertram Muller and Bert Bakker, Clinical Genetics Society, problems and proceedings.

Young ID. *Introduction to risk calculation in genetic counselling*. Oxford: Oxford University Press, 1991.

References Duchenne Muscular Dystrophy

Bakker E, Veenema H, Den Dunnen JT, Van Broeckhoven C, Grootscholten PM, Bonten EJ, Van Ommen GJB and Pearson PL. Germinal mosaicism increases the recurrence risk for 'new' Duchenne muscular dystrophy mutations. J Med Genet 1989; 26:553-559

Bakker E, Van Broeckhoven C, Bonten EJ, Van de Vooren MJ, Veenema H, Van Hui W, Van Ommen GJB, Vandenberghe A and Pearson PL. Germline mosaicism and Duchenne muscular dystrophy mutations. Nature 1987; 329:554-556

Beggs AH, Koenig M, Boyce FM, Kunkel LM. Detection of 98% of DMD/BMD gene deletions by polymerase chain reaction. Hum Genet 1990; 86:45-48.

Davies KE, Pearson PL, Harper PS et al. Linkage analysis of two cloned DNA sequences flanking the Duchenne muscular dystrophy locus on the short arm of the human X chromosome. Nucleic Acids Res 1983; 11:2303-2312.

Den Dunnen JT, Grootscholten PM, Bakker E et al. Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. Am J Hum Genet 1989; 45:835-847.

Emery AE. Duchenne Muscular Dystrophy. Revised edition. Oxford: Oxford University Press, 1988.

Golbus MS, Stephens JD, Mahoney MJ, Hobbins JC, Haseltine FP, Caskey CT and Banker BQ. Failure of fetal creatine phosphokinase as a diagnostic indicator of Duchenne muscular dystrophy. New Engl J Med 1979; 300:860-861

Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 1987; 50:509-517.

Kunkel LM. Analysis of deletions in DNA from patients with Becker and Duchenne muscular dystrophy. Nature 1986; 322:73-77.

Lange K, Zatz M. A new method for the analysis of age trends in CPK levels with application to Duchenne muscular dystrophy. Hum Hered 1979; 29:154-160.

Lindlof M, Kiuru A, Kaariainen H, et al. Gene deletions in X-linked muscular dystrophy. Am J Hum Genet 1989; 44:496-503.

Monaco AP, Bertelson CJ, Colletti-Feener C, Kunkel LM. Localization and cloning of Xp21 deletion breakpoints involved in muscular dystrophy. Hum Genet 1987; 75:221-227.

Murray JM, Davies KE, Harper PS, Meredith L, Mueller CR, Williamson R. Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. Nature 1982; 300:69-71.

Prior TW, Papp AC, Snyder PJ, et al. Heteroduplex analysis of the dystrophin gene: application to point mutation and carrier detection. Am J Med Genet 1994; 50:68-73.

Roberts RG, Gardner RJ, Bobrow M. Searching for the 1 in 2,400,000: a review of dystrophin gene point mutations. Hum Mutat 1994; 4:1-11.

Roberts RG, Bentley DR, Bobrow M. Infidelity in the structure of ectopic transcripts: a novel exon in lymphocyte dystrophin transcripts. Hum Mutat 1993; 2:293-299.

Thompson MW, Murphy EG, McAlpine PJ. An assessment of the creatine kinase test in the detection of carriers of Duchenne muscular dystrophy. J Pediatr 1967; 71:82-93.

Van Essen AJ, Abbs S, Baiget M, et al. Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: a European study. Hum Genet 1992; 88:249-257.

Van Essen AJ, Busch HF, Te Meerman GJ, Ten Kate LP. Birth and population prevalence of Duchenne muscular dystrophy in The Netherlands. Hum Genet 1992; 88:258-266.

Wapenaar MC, Kievits T, Hart KA, et al. A deletion hot spot in the Duchenne muscular dystrophy gene. Genomics 1988; 2:101-108.

Zatz M, Frota Pessoa O, Levy JA, Peres CA. Creatine-phosphokinase (CPK) activity in relatives of patients with X-linked muscular dystrophies: a Brazilian study. J Genet Hum 1976; 24:153-168.

II-2. Recurrence risk for germinal mosaicism revisited.

Martin A. van der Meulen, Meine J.P. van der Meulen and Gerard J. te Meerman.

J Med Genet 1995;32:102-104

Department of Medical Genetics,
University of Groningen,
A. Deusinglaan 4,
9713 AW Groningen,
The Netherlands.

phone : +31 50 3632925
fax : +31 50 3632947
email : m.a.van.der.meulen@med.rug.nl

J Med Genet 1995;32:102–104

Recurrence risk for germinal mosaics revisited

Martin A van der Meulen, Meine J P van der Meulen, Gerard J te Meerman

Abstract

A formula to calculate recurrence risk for germline mosaicism published by Hartl in 1971 has been updated to include marker information. For practical genetic counselling new, more elaborate tables are given.

(*J Med Genet* 1995;32:102–104)

In 1974 Murphy *et al*[1] concluded "From the clinical standpoint the main implication is that it provides reassurance that, in any realistic size of family, ignoring the effect of gonadal mosaicism will have little effect on the estimate of the risk for the next child". This paper has been referred to many times in papers about mosaicism. A paper by Hartl,[2] entitled "Recurrence risk for germinal mosaics", gives formulae to calculate recurrence risks, correcting for the number of affected and unaffected sibs. This paper will discuss the impact of modern molecular genetic techniques on the recurrence risk for gonadal mosaicism.

The theory

In all humans some degree of mosaicism is the norm for the more common genetic disorders, since the mutation rate multiplied by the number of mitoses necessary to form the 5–7 million[3] or 8 million oocytes[4] in females and the many more spermatozoa in males is much higher than one. In mosaics distinction has to be made between somatic and germline mosaics. In somatic mosaics the germ stem cell is mutated. This mutation happens in one of the 46 chromosomes (2n) in a mitosis between fertilisation and the mitoses leading to the germ stem cell. In somatic mosaics all gonadal cells before meiosis will have the same mutation on one chromosome, and because of the meiosis in which a 2n chromosome cell is split into two cells with n chromosomes, there is a 50% recurrence risk. In germline mosaics a fraction of the gonadal cells will have a mutation. This is dependent on the gonadal generation in which the mutation occurred, since all descendent cells of the mutated cell are carriers of the mutation. For a more detailed description of this model see Murphy *et al.*[1] The recurrence risk is dependent on the gonadal generation in which the mutation occurred to a maximum of 50% (first gonadal generation mutation).

The recurrence risk owing to germline mosaicism can be explained through the development of all 2^n oocytes from one healthy wild type cell. In 2^n, n is the number of cell generations to get from one healthy cell to the total amount of oocytes. Hartl[2] showed that more complex models than this simple model do not change recurrence risks as long as the number of gonadal generations is high enough. If we assume that during gametogenesis the probability of mutation in each gonadal generation, conditional on the occurrence of a mosaic genotype, is the same for all cell generations, the recurrence risk owing to germline mosaicism, only taking into account the information that one parent is a mosaic, can be calculated from the number of cell generations n to be

$$\text{Recurrence risk} = \Sigma_{i=1}^{n} \frac{1}{n} * (2)^{-i} \approx \frac{1}{n}.$$

According to Hartl,[2] in human females the effective number of gonadal generations n is 10–12. This will give a recurrence risk of 8·3%. If, however, it is assumed that during fetal life the number of oocytes reaches 5–7 million,[3] n has to be at least 22 in females. This gives a recurrence risk of 4·5%. In males n has to be at least 30.[3,4] Note that the recurrence risk is independent of the mutation rate. The recurrence risk is conditional on the fact that one of the parents is a germinal mosaic, owing to a single mutation.

Bakker *et al*[5] estimated recurrence risks for X linked Duchenne muscular dystrophy mosaics from empirical data to be 7%. Since families are not collected at random, proper correction for ascertainment bias is essential. Proper ascertainment bias correction is only possible if the correct model is known, which is rarely the case.[6] Prenatal diagnoses based on X haplotype information either leads to the exclusion of the at risk X chromosome or a 14% recurrence risk for male pregnancies. Van Essen *et al*[7] give an overview of the pooled results from 25 European centres.

Hartl[2] gives a formula to calculate the recurrence risk from the number of affected and unaffected sibs. He does not, not surprisingly in 1971, correct for marker information, confirming that sibs actually inherited the at risk chromosome from their mosaic parent. Of course, his calculations are still valid, if there are no marker data available, for instance because the location of the disease is unknown or because there is no genetic material from affected sibs, so the at risk chromosome can not be determined. To include marker information in recurrence risk calculations the formula has to be extended to take in typed and untyped sibs, making more accurate risk calculations possible.

If we define the percentage of affected chromosomes R, the following classes of sibs can be distinguished.

Department of Medical Genetics, University of Groningen, A Deusinglaan 4, 9713 AW Groningen, The Netherlands
M A van der Meulen
M J P van der Meulen
G J te Meerman

Correspondence to:
Dr M A van der Meulen.

Received 12 April 1994
Revised version accepted for publication 11 October 1994

Table 1 Recurrence risk for gonadal mosaics with one affected child (a = 1) and a variable number of untyped unaffected children (b) and typed unaffected children with the at risk chromosome (c)

Recurrence risk (a = 1)		*No of typed unaffected children with the at risk chromosome (c)*				
		0	*1*	*2*	*3*	*4*
No of untyped unaffected children (b)	0	0·048	0·018	0·013	0·010	0·008
	1	0·033	0·015	0·012	0·009	0·008
	2	0·025	0·013	0·010	0·008	0·007
	3	0·019	0·012	0·009	0·008	0·007
	4	0·016	0·011	0·009	0·007	0·006

Table 2 Recurrence risk for gonadal mosaics with two affected children (a = 2) and a variable number of untyped unaffected children (b) and typed unaffected children with the at risk chromosome (c)

Recurrence risk (a = 2)		*No of typed unaffected children with the at risk chromosome (c)*				
		0	*1*	*2*	*3*	*4*
No of untyped unaffected children (b)	0	0·333	0·143	0·120	0·100	0·084
	1	0·286	0·133	0·111	0·093	0·079
	2	0·240	0·124	0·103	0·087	0·074
	3	0·200	0·115	0·096	0·081	0·069
	4	0·168	0·107	0·089	0·075	0·065

Table 3 Recurrence risk for gonadal mosaics with three affected children (a = 3) and a variable number of untyped unaffected children (b) and typed unaffected children with the at risk chromosome (c)

Recurrence risk (a = 3)		*No of typed unaffected children with the at risk chromosome (c)*				
		0	*1*	*2*	*3*	*4*
No of untyped unaffected children (b)	0	0·429	0·200	0·183	0·164	0·145
	1	0·400	0·194	0·175	0·156	0·138
	2	0·366	0·187	0·168	0·149	0·131
	3	0·328	0·179	0·160	0·141	0·124
	4	0·290	0·172	0·153	0·134	0·118

Table 4 Recurrence risk for gonadal mosaics with four affected children (a = 4) and a variable number of untyped unaffected children (b) and typed unaffected children with the at risk chromosome (c)

Recurrence risk (a = 4)		*No of typed unaffected children with the at risk chromosome (c)*				
		0	*1*	*2*	*3*	*4*
No of untyped unaffected children (b)	0	0·467	0·226	0·216	0·203	0·188
	1	0·452	0·222	0·211	0·197	0·181
	2	0·431	0·218	0·206	0·191	0·175
	3	0·406	0·214	0·200	0·185	0·168
	4	0·375	0·209	0·195	0·178	0·161

Untyped sibs:

Affected: 100% chance of having the risk chromosome, R% of the sibs;

Unaffected: (50-R)% chance of having the risk chromosome, 50% of having the not risk chromosome. Total (100-R)% of the sibs.

Typed sibs: first of all to be used to determine if the grandmaternal or the grandpaternal chromosome is the mosaic chromosome.

Affected: has the at risk chromosome, informative to increase recurrence risk, R% of the sibs.

Unaffected: having the at risk chromosome, informative to decrease a priori risk, (50-R)% of the sibs; not having the at risk chromosome, not informative, except for determination on which chromosome the mutation occurred (grandpaternal or grandmaternal), 50% of the sibs.

From all the classes mentioned above, the three informative classes have to be included in a risk calculation: the affected (A), the untyped unaffected (B), and the typed unaffected with the at risk chromosome (C). The unaffected sibs who are typed and have the non-mosaic chromosome are not informative, because they do not provide information about the frequency of the mosaic chromosome and therefore of the recurrence risk R.

Extending the recurrence risk formula as given by Hartl[2] with the third class of typed unaffected children with the at risk chromosome leads to the straightforward Bayesian formula, where P(A/F) is the probability that the next child born to a mosaic parent is affected, given the distribution F of the number of affected children (a), untyped unaffected children (b), and typed unaffected children with the risk chromosome (c). Details of derivation of the formula are shown in the appendix.

$$P(A \mid F) = \frac{\sum_{i=0}^{n} \frac{1}{2}^{(i+1).a} \cdot (1 - \frac{1}{2}^{i+1})^{b} \cdot (\frac{1}{2} - \frac{1}{2}^{i+1})^{c}}{2.\sum_{i=0}^{n} \frac{1}{2}^{(i+1).a} \cdot (1 - \frac{1}{2}^{i+1})^{b} \cdot (\frac{1}{2} - \frac{1}{2}^{i+1})^{c} \cdot 2^{i}}$$

Results

Since the formula is quite complex, recurrence risks are calculated and shown in tables 1 to 4 for the situations which appear most often in practical genetic counselling. The recurrence risks are calculated given at least one affected child. Wijsman[3] and Edwards[4] estimated that the number of gonadal generations n must be approximately 23 in females and approximately 30 in males. Recurrence risks are almost constant if n is greater or equal to 20 and because we do not want to underestimate the recurrence risk, n is set to 20. The risks in the last column of table 1 from Hartl[2] can then be found in the tables. The risks in the tables are the recurrence risks for the next child given the pedigree with at least one affected child and no chromosomal information on the "newborn". Prenatal information on the inherited chromosome of the "newborn" reduces the risk to 0 if the not at risk chromosome is encountered or doubles the risk as stated in the tables otherwise.

Discussion

One of the major problems in risk calculation for germline mosaics is the impact on the recurrence risk in comparison to the impact of somatic mosaics and, in X linked recessive cases, the carrier probability of the mother. In Duchenne muscular dystrophy (X linked) one-third of the cases are thought to be the result of a new mutation.[8 9] If parents are non-carriers, as can be seen from their phenotype in dominant cases, the mutation must be either somatic or germinal. Although it is unclear how to divide probabilities over somatic and germline mosaics, the impact of somatic mosaics is higher because in this case 50% of the children are affected against approximately 5% in germline mosaics. However, one of the advantages of modern genetic technology is the ability to determine which chromosome is inherited from the parents. This way it can be proven that one of the parents is a germline mosaic if an affected and an unaffected child, carrying the same chromosome from one of the parents, is en-

countered. Germline mosaicism is proven if at least two children have the same genotype, but different phenotype (affected and unaffected). Still, care has to be taken to exclude phenocopies, mutation in the affected child itself leading to somatic mosaicism, incomplete penetrance, or intragenic recombinations.

The formula presented in this paper has an equal conditional probability of mutation in each gonadal generation. Passos-Bueno *et al*[10] concluded that different mosaicism frequencies for proximal and distal DMD mutations exist. The proximal mutations in the DMD families are thought to have a higher recurrence risk through occurring very early in embryonal development, therefore increasing the proportion of mutated cells. The formula can be extended to take the different mutation probabilities of the different gonadal generations into account as soon as more information is available.

In practical genetic counselling, there is often no chromosomal material available from an affected diseased child. Since it is therefore unknown which chromosome of the parents is the mosaic chromosome, special care must be taken. If, for instance, there is a pedigree available with an X linked disease, one affected son genotype unknown, one unaffected son with the mother's paternal chromosome and two sons with the mother's maternal chromosome, the situation can be analysed as:

(1) Treating all sons as unknown genotype, since it is not known what the at risk genotype is. In table 1, one affected and three unaffected sons can be found with a recurrence risk of 1·9% (A=1, X=3, G=0).

(2) Analysing the situation in two steps: if the affected son has the mother's father's chromosome, the recurrence risk is, given one unaffected brother (table 1, A=1, X=0, G=1) 1·8%; if the affected son has the mother's mother's chromosome, the recurrence risk is, given two unaffected brothers (table 1, A=1, X=0, G=2), 1·3%. If prenatal information is available the risk of the appropriate chromosome can be doubled.

Linkage programs currently do not routinely allow for germline mosaicism. This effect may be reduced by allowing for a rather high mutation rate.[6] Grimm *et al*[9] described an approach to estimating parameters at the level of the population. This results in much easier algebra and could therefore be incorporated into existing computer programs like LINKAGE.[11] Jeanpierre[12] devised a computer program to calculate the probability of a possible carrier, in order to settle the origin of a mutation of a given family.

This work is supported by the Netherlands Organisation for Scientific Research.

Appendix

In the derivation of the formula presented in this paper, the same assumptions are made as Hartl[2] did for the synchronous, symmetric, dichotomous model. Therefore, in the Bayesian formula all terms are the same as derived by Hartl,[2] except the term $P(F \mid M=i)$. This term has to be extended with the third informative class: the typed unaffected sibs with the at risk chromosome (class C, with c sibs).

$$P(F \mid M=i) = K.\tfrac{1}{2}^{(i+1).a}.(1-\tfrac{1}{2}^{(i+1)})^{b}.(\tfrac{1}{2}-\tfrac{1}{2}^{(i+1)})^{c}.$$

As in the formula of Hartl[2] the birth order K of the sibs in the sibship is irrelevant, because this term appears in the Bayes formula in both the numerator and denominator and therefore cancels out. Note that the number of unaffected sibs (b) in this extended formula are the untyped unaffected sibs only. The number of affected sibs (a) stays the same as in the original formula. Substituting the term above and the terms derived by Hartl[2] into the Bayes formula leads to the recurrence risk formula $P(A \mid F)$ as presented in this paper.

1 Murphy EA, Cramer DW, Kryscio RJ, Brown CC, Pierce ER. Gonadal mosaicism and genetic counseling for X linked recessive lethals. *Am J Hum Genet* 1974;**26**:207–22.
2 Hartl DL. Recurrence risk for germinal mosaics. *Am J Hum Genet* 1971;**23**:124–34.
3 Wijsman EM. Recurrence risk of a new dominant mutation in children of unaffected parents. *Am J Hum Genet* 1991; **48**:654–61.
4 Edwards JH. Familiarity, recessivity and germline mosaicism. *Ann Hum Genet* 1989;**53**:33–47.
5 Bakker E, Veenema H, Den Dunnen JT, *et al*. Germinal mosaicism increases the recurrence risk for 'new' Duchenne muscular dystrophy mutations. *J Med Genet* 1989; **26**:553–9.
6 Ott J. *Analysis of human genetic linkage*. Revised ed. Baltimore: The Johns Hopkins University Press, 1991.
7 Van Essen J, Abbs S, Baiget M, *et al*. Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: a European study. *Hum Genet* 1992;**88**: 249–257.
8 Bakker E, Van Broeckhoven C, Bonten EJ, *et al*. Germline mosaicism and Duchenne muscular dystrophy mutations. *Nature* 1987;**329**:554–6.
9 Grimm T, Muller B, Muller CR, Janka M. Theoretical considerations on germline mosaicism in Duchenne muscular dystrophy. *J Med Genet* 1990;**27**:683–7.
10 Passos-Bueno MR, Bakker E, Kneppers ALJ, *et al*. Different mosaicism frequencies for proximal and distal Duchenne muscular dystrophy (DMD) mutations indicate difference in etiology and recurrence risk. *Am J Hum Genet* 1992; **51**:1150–5.
11 Lathrop GM, Lalouel JM. Easy calculation of lod scores and genetic risks on small computers. *Am J Hum Genet* 1984;**36**:460–5.
12 Jeanpierre M. Germinal mosaicism and risk calculation in X linked diseases. *Am J Hum Genet* 1992;**50**:960–7.

II-3.1 Calculation of Recurrence Risk in case of possible Mosaicism.

Martin A. van der Meulen[1], Gerard J. te Meerman[1] and Lodewijk A. Sandkuijl[123].

Submitted

[1]Department of Medical Genetics,
University of Groningen,
A. Deusinglaan 4,
9713 AW Groningen,
The Netherlands.

[2]Institute of Clinical Genetics,
Erasmus University,
Rotterdam,
The Netherlands.

[3]Department of Human Genetics,
Leiden University,
Leiden,
The Netherlands.

phone : +31 50 3632925
fax : +31 50 3632947
email : m.a.van.der.meulen@med.rug.nl
keywords: mosaicism, recurrence risk, Duchenne muscular dystrophy

Abstract

A general framework for recurrence risk calculation in case of possible mosaicism is presented. This work is an extension of the work presented by Van der Meulen et al. in 1995, who calculated recurrence risks in nuclear families under germinal mosaicism. Here it is shown how the possibility of somatic or germinal mosaicism can be taken into account in manual risk calculations for pedigrees of arbitrary structure. Examples of carrier risk calculations are given for a pedigree with the X-linked recessive disease Duchenne Muscular Dystrophy. The suggested approach is suitable for implementation in a computer programme.

Introduction.

The possible presence of mosaicism complicates the accurate calculation of recurrence risks in genetic counselling situations. A model to account for germline mosaicism was initially presented by Hartl (1971), and has been extended by Van der Meulen et al. (1995) to include DNA marker information. The latter authors provide detailed tables with recurrence risks under germline mosaicism for various sibship compositions. Their tables are given in the appendix. These tables are restricted to nuclear families and to the situation that the marker genotype of the index patient is known. Here we present a general method to calculate the recurrence risk or the carrier probability for a specific member of a pedigree of arbitrary structure, taking into account all available phenotypic information and allowing for the possibility of somatic and germline mosaicism.

Theoretical aspects of mosaicism.

The following considerations refer to families with at least one affected individual, the family situation which is most relevant for risk calculation. The transmitting parent of the index patient can either be a carrier, or a mosaic due to the occurrence of a mutation in one of the cell divisions between fertilisation and the formation of germ cells. We assume, like Hartl (1971), Wijsman (1991), Jeanpierre (1992), Van der Meulen et al. (1995), and most others, a synchronous, symmetric, dichotomous model with a constant mutation rate per cell division. Disadvantages and advantages of this model were discussed in detail by Wijsman (1991). If the mutation in the transmitting parent occurred in one of the mitoses that lead to the germ stem cell, all germline cells and 50 % of the

oocytes and spermatocytes will carry the mutation: somatic mosaicism. Mutations that occur later will lead to a smaller proportion of gametes to be affected: germline mosaicism (Bakker et al., 1987, 1989). Ergo, germline mosaics are defined as being able to transmit the risk chromosome with and without a mutation, while somatic mosaics can only transmit the risk-chromosome with the mutation. Somatic mosaics will by definition have the mutation in their germline, because the calculation is conditional on the index patient. Germline mosaics will never show the affected phenotype or can be recognized as carrier in blood samples, because they only have the mutation in their germ cells.

Although the exact number of mitoses after fertilisation and before the germline stem cell is formed in humans is unknown, this number must be in the range of 10 to 15 mitoses. This is based on embryogenesis, in which very early in development germline stemcells can be recognized by their large size and prominent nuclei (Longo and Anderson, 1974). Up to this stage male and female development of genitalia has been similar (Beck et al, 1973). Starting from the germ stem cell, at least 20 gonadal generations, but possibly up to 30, are effectively required in the germline to form all oocytes or all sperm cells. For a more detailed description of this model, see Murphy et al. (1974). Van der Meulen et al. (1995) calculated recurrence risks due to germline mosaicism based on 20 gonadal generations. They argue that the risks calculated assuming 20 gonadal generations will hardly change, and only decrease, if one or a few extra mitoses occur in reality (see also Wijsman, 1991).

When an equal mutation rate per mitosis is assumed the number of mitoses before and after the germline stemcell is formed can be used to approximate the 10:20 or 15:30 $\approx$ 1/3:2/3 ratio between somatic mosaics and germline mosaics. Based on the expected recurrence risk for a somatic mosaic (50 %) and the expected recurrence risk for a germline mosaic (4.8 %, Table I, appendix, one affected child, no other children), the recurrence risk due to a new mutation can be calculated as

$1/3 * .50 + 2/3 * .048 = .20$

Next to the assumptions made for modelling purposes, some general assumptions are necessary: 1) no mistaken paternity or other identity problems exist, 2) a mutation in a transmitting parent will not lead to a disease phenotype in that parent, even when it concerns an early somatic mutation; the disease phenotype only occurs in the child (see discussion for implications), 3) oocyte and spermatocyte are randomly drawn from the

pool, there is no selection against or in favour of cells that carry the mutation, 4) the possibility of multiple independent mutation events is not considered: risks are calculated for inheriting the mutation that is already present in the index patient, 5) meiotic mutations are not considered in the calculations.

Probability calculations

If the intragenic marker genotype of the affected sibling(s) is known and there is at least one unaffected sib with the same genotype, there is no ambiguity in the segregation pattern and the recurrence risk can be taken from the lookup tables (appendix). In cases of ambiguity, however, such as uncertain carrier status in females or unknown marker status for the index patient, all possible segregation patterns have to be evaluated, and the total risk is a weighted average of the risks obtained for the individual segregation patterns.

Example / Results

To indicate how a recurrence risk can be calculated considering the above model and assumptions, we present a pedigree as shown in figure 1. The index patient had Duchenne Muscular Dystrophy (DMD). In X-linked recessive lethals the mutation-selection equilibrium implies that a woman with an affected son has a probability 2/3 of being a carrier (Haldane , 1935, Murphy and Chase, 1975), assuming an equal mutation frequency in males and females.

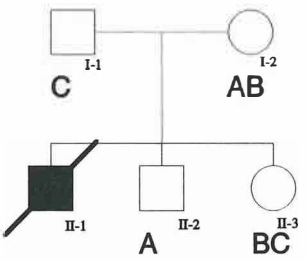


Figure 1: Pedigree

Table 1 shows the recurrence risk calculation for consultand II-3 in the case that no marker information is available. This calculation is comparable to classical calculations that account for mutation (Young, 1991). The genotype probabilities for I-2 are conditional on the occurrence of the disease in her son II-1, and are based on the ratio between mutations in the somatic phase and the germline phase (1/3 : 2/3) and on the ratio between gene frequency and new mutations (given equilibrium between mutation and selection and given equal mutation frequency in males and females, $q=2\mu$).

The conditional probability for II-2 to be unaffected if the mother is a germline mosaic is taken from table I, appendix: no unaffected siblings. The probability to be unaffected is obviously equal to (1 - the probability to be affected). When more siblings are present, a separate line should be added for each sibling, listing the probabilities for this sibling's disease or carrier status conditional on the parts of the pedigree that were accounted for on the preceding lines. Accordingly, the conditional genotype probabilities for II-3 (carrier or not), given the presence of germline mosaicism in the mother and given the information on II-2, are obtained from table I, appendix: 1 untyped sibling. Birth order is irrelevant in these calculations (Hartl, 1971, Van der Meulen et al., 1995). The likelihood for a given column is calculated by multiplication of all entries in the column. The risk for a consultand to be carrier can be calculated by summing the likelihoods for all columns in which the consultand was entered as carrier, and dividing by the sum of the likelihoods for all columns. Note that the entire calculation is conditional on the first affected sibling; this index patient determines the priors for person I-2. Therefore no separate line is required for this patient.

When marker information is available, it is no longer possible to apply the tables given by Van der Meulen et al. (1995) directly. Those tables were all constructed for the situation that it is not known which maternal chromosome is inherited by the at-risk person: there is a 50 % probability to inherit the chromosome without the mutation (risk=0), and a 50 % probability to inherit the risk-chromosome which carries the mutation (in some or all instances). The risks given in the tables therefore represent an average between these two alternatives, never exceeding a total risk of 50 %. When marker information can be used to decide which maternal chromosome was inherited, there are two possible outcomes. When the risk-chromosome is inherited, the risk to inherit the mutation is twice as high as given in the tables, while there is zero risk for children who inherit the other maternal chromosome.

Figure 1 also shows the genotypes of the persons in the pedigree for an intragenic marker (to simplify this example, we do not allow for recombination). The marker genotype for the index patient is unknown. Table 2 shows the calculation of the carrier risk using the marker information. The number of columns is doubled, because the mutated

chromosome in the mother can carry either allele A or B at the marker, and these two linkage phases have to be considered separately. The probability for II-2 to be unaffected given that this person inherited the A allele from the mother now depends on the linkage phase that we assume for the mother. If the mother is a germline mosaic and if her risk-chromosome carries the B allele (table 2, last column), the conditional probability for II-2 to be unaffected is 1, while for the opposite maternal phase it is equal to 1 minus the expected frequency of the mosaicism (twice the risk from table I, appendix).
In case mother I-2 is a germline mosaic on the chromosome with marker allele B the conditional probability for consultand II-3 to be carrier given that she inherited the chromosome with allele B is taken from table I, appendix, no unaffected siblings with the risk chromosome. This value must also be doubled, because it is known that the maternal risk-chromosome is inherited. When consultand II-3 inherits the chromosome with marker allele B from her mother and the risk-chromosome is the chromosome with marker allele A, II-3 will not be carrier of the mutation.

In DMD CK values are often used as additional information to infer carrier status in females. In Table 3 we calculated the carrier risk given that mother I-2 and consultand II-3 have normal CK values. In this calculation we assumed that all non-carriers and 1/3 of carriers have normal CK values. Table 4 shows that the calculations can be extended to include information obtained from screening the relevant gene for mutations in possible carriers. Carrier status becomes less likely when the presence of some of the more common mutations has been excluded in a possible carrier. For table 4 we assumed that in 60 % of families the responsible mutation can be detected via screening. Note that if both the mother and the consultand are checked for mutations, the probability of detecting the mutation needs to be entered in the table only once, even if the mutation screening is carried out on both individuals. In table 4 both the mother and the consultand have no detectable mutations. We assumed that in case of somatic mosaicism the mutation cannot be detected in the mother. This assumption has no influence on the result of the calculation, because all columns in which there could be influence have already become impossible.

Discussion

We illustrate that in case of possible mosaicism recurrence risks can be calculated for pedigrees of arbitrary structure, using the tables by Van der Meulen et al. (1995). The framework for manual calculations as presented here is analogous to the method of presentation followed by Young (1991). Although the examples as given are dependant on several assumptions, the general framework is more generally applicable.

The example calculations require some basic assumptions regarding the origin and dynamics of gonadal mosaicism in DMD. It is important to realize that the risk to inherit a mutation from a mosaic person is only considerably increased over the population risk when that mosaic person carries the mutation in a high proportion of gametes. Mutations in the early mitoses will lead to a high frequency mosaicism, while late events are practically indistinguishable from an isolated mutation in a gamete and will not lead to a seriously increased risk for offspring.

The most important assumption in our model relates to the number of cell divisions prior to germ stem cell formation. We deduce a ratio of 1/3:2/3 between somatic and germline mosaics from the expected number of mitoses in the somatic and germline phase. Van Essen et al. (1992) concluded from empirical data that sibs of a DMD patient with a detectable, apparently new, DMD mutation have a 20% probability of inheriting the mutation together with the at risk chromosome. This estimated recurrence risk of apparent new mutations leads to lower recurrence risks, because they define a mosaic when the mutation can not be detected in the parent, while this is not the case in our definition, where early somatic mosaics will be undistinguishable from carriers.

Our assumption that a new mutation always occurs in the transmitting parent and never in the child is a worst case scenario: if a mutation in one of the first mitoses after fertilisation has led to the affected phenotype in the index patient then the recurrence risk for siblings of the affected is zero.

We assume that mosaic persons never express the disease phenotype. Some reports suggest that cases of early somatic mosaicism for autosomal dominant diseases or for X-linked recessive disease (males) may show the disease. This implies that in case of an unaffected male mosaic parent, we might have to decrease the number of mitoses (and thereby the opportunity for mutation) in the somatic phase. In females an early mutation leading to a somatic mosaic might influence the CK value.

No assumption was needed regarding the frequency of mutation, since all calculations are conditional on the presence of disease in the index patient. Differences in mutation frequency between males and females will only be relevant when the grandparental origin of a risk-chromosome can be traced, or when the type of mutation reveals the grandparental origin (Grimm et al., 1994), because both haplotypes of the transmitting parent will have identical carrier probabilities as long as the parental origin of the X chromosome of the index patient is unknown. Differences in mutation frequency between males and females will also influence the mutation selection equilibrium.
In the examples given here, we only included sibs of the index patient, while unaffected or affected sons of sisters may also contribute important information. Such information can be taken into account following the principles as indicated, but the calculations will become increasingly complex, because for each person with genotypic ambiguity the number of columns in the calculation table doubles. We are currently preparing a general computer programme to handle more complicated pedigrees with multiple generations. Even when such a program is available, it is important that risks for relatively simple pedigree structures can be calculated or verified manually via the procedure presented here.

Calculation of recurrence risk in case of possible mosaicism is based on theoretical models and assumptions. The models and assumptions are based on possible biological mechanisms. Insight in biological mechanisms, through experimental data, will change the model parameters, but the framework as described will still be valid. Robustness of the model to parameter changes can be tested by variation of the input parameters. Using extreme values for parameters gives insight in the influence of the specific parameter on the calculated carrier/recurrence risk.

Acknowledgements

This work is supported by the Netherlands Organisation for Scientific Research (NWO). We wish to thank Dr. A. J. van Essen for helpful and critical comments on the manuscript.

References

Bakker E, Van Broeckhoven C, Bonten EJ, Van de Vooren MJ, Veenema H, Van Hui W, Van Ommen GJB, Vandenberghe A and Pearson PL (1987): Germline mosaicism and Duchenne muscular dystrophy mutations. Nature 329:554-556

Bakker E, Veenema H, Den Dunnen JT, Van Broeckhoven C, Grootscholten PM, Bonten EJ, Van Ommen GJB and Pearson PL (1989): Germinal mosaicism increases the recurrence risk for 'new' Duchenne muscular dystrophy mutations. J Med Genet 26:553-559

Beck F, Moffat DB, Lloyd JB (1973): Human embryology and Genetics. Oxford London Edinburgh Melbourne: Blackwell Scientific Publikations, pp 248-249.

Grimm T, Meng G, Liechti-Gallati S, Bettecken T, Muller CR and Muller B (1994): On the origin of deletions and point mutations in Duchenne muscular dystrophy: most deletions arise in oogenesis and most point mutations result from events in spermatogenesis. J Med Genet 31:183-186

Haldane JBS (1935): The rate of spontaneous mutations of a human gene. J Genet 31:317-326

Hartl DL (1971): Recurrence risk for germinal mosaics. Am J Hum Genet 23:124-134

Jeanpierre M (1992): Germinal mosaicism and risk calculation in X-linked diseases. Am J Hum Genet 50:960-967

Longo FJ and Anderson E (1974): Gametogenesis. In Lash J and Whittaker JR (eds): Concepts of Development. , Stamford Connecticut: Sinauer Associates, pp 3-47.

Murphy EA and Chase GA (1975): Principles of genetic counselling. Chicago: Year Book Medical Publishers.

Murphy EA, Cramer DW, Kryscio RJ, Brown CC and Pierce ER (1974): Gonadal mosaicism and genetic counselling for X-linked recessive lethals. Am J Hum Genet 26:207-222

Van der Meulen MA, Van der Meulen MJP and Te Meerman GJ (1995): Recurrence risk for germinal mosaics revisited. J Med Genet 32:102-104

Van Essen AJ, Abbs S, Baiget M, Bakker E, Boileau C, Van Broeckhoven C et al. (1992): Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: A European study. Hum Genet 88:249-257

Wijsman EM (1991): Recurrence risk of a new dominant mutation in children of unaffected parents. Am J Hum Genet 48:654-661

Young ID (1991): Introduction to risk calculation in genetic counselling. Oxford: Oxford University Press, 1991.

Table 1:
Carrier probability calculation for person II-3 in Figure 1, without using information of the intragenic marker. Car+ and Car- refers to carriers and non-carriers respectively.

I-2	Carrier		Som. Mos.		Germl. Mos.	
Probability						
Prior:	2/3		1/3 * 1/3		1/3 * 2/3	
Conditional:						
II-2 unaffected	1/2		1/2		1-0.048	
II-3:	Car+	Car-	Car+	Car-	Car+	Car-
	1/2	1/2	1/2	1/2	0.033	1-0.033
Joint	1/6	1/6	1/36	1/36	0.007	0.205

The carrier probability for II-3 equals $(1/6+1/36+0.007)/(1/6+1/6+1/36+1/36+0.007+0.205)=0.336$

Table 2:

Carrier probability calculation for person II-3, using the intragenic marker as shown in Figure 1. Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

I-2	Carrier				Som. Mos.				Germl. Mos.			
Mutation on chr.	A		B		A		B		A		B	
Probability												
Prior:	2/3*1/2		2/3*1/2		1/3*1/3*1/2		1/3*1/3*1/2		1/3*2/3*1/2		1/3*2/3*1/2	
Conditional:												
II-2 unaffected & allele A	0		1		0		1		1-2*0.048		1	
II-3:	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-
allele B	X	X	1	0	X	X	1	0	0	1	2*0.048	1-2*0.048
Joint	0	0	1/3	0	0	0	1/18	0	0	0.1	0.011	0.1

The carrier probability for II-3, if she has the intragenic marker genotype BC, equals (1/3+1/18+0.011)/(1/3+1/18+0.1+0.011+0.1)=0.667.

Table 3:
Carrier probability calculation for person II-3, using the intragenic marker as shown in Figure 1 and I-2 and II-3 have normal CK values. Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

I-2	Carrier				Som. Mos.				Germl. Mos.			
Mutation on chr.	A		B		A		B		A		B	
Probability												
Prior:	1/3		1/3		1/3*1/3*1/2		1/3*1/3*1/2		1/3*2/3*1/2		1/3*2/3*1/2	
Conditional:												
I-2:												
normal CK	**1/3**		**1/3**		**1**		**1**		**1**		**1**	
II-2 unaffected & allele A	0		1		0		1		1-2*0.048		1	
II-3:	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-
allele B	X	X	1	0	X	X	1	0	0	1	2*0.048	1-2*0.048
normal CK	**X**	**X**	**1/3**	**X**	**X**	**X**	**1/3**	**X**	**X**	**1**	**1/3**	**1**
Joint	0	0	1/27	0	0	0	1/54	0	0	0.1	0.0036	0.1

The carrier probability for II-3, if she has the genotype BC and I-2 and II-3 have normal CK values, equals (1/27+1/54+0.0036)/(1/27+1/54+0.1+0.0036+0.1)=0.228.

Table 4:

Carrier probability calculation for person II-3, using the intragenic marker as shown in Figure 1 and I-2 and II-3 have normal CK values and I-2 and II-3 are screened for all known mutations (60 %). Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

I-2	Carrier				Som. Mos.				Germl. Mos.			
Mutation on chr.	A		B		A		B		A		B	
Probability												
Prior:	1/3		1/3		1/3*1/3*1/2		1/3*1/3*1/2		1/3*2/3*1/2		1/3*2/3*1/2	
Conditional:												
I-2:												
normal CK	1/3		1/3		1		1		1		1	
no mutation	**2/5**		**2/5**		**1**		**1**		**1**		**1**	
II-2 unaffected & allele A	0		1		0		1		1-2*0.048		1	
II-3:	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-
allele B	X	X	1	0	X	X	1	0	0	1	2*0.048	1-2*0.048
normal CK	X	X	1/3	X	X	X	1/3	X	X	1	1/3	1
no mutation	**X**	**X**	**1**	**X**	**X**	**X**	**2/5**	**1**	**X**	**1**	**2/5**	**1**
Joint	0	0	2/135	0	0	0	2/270	0	0	0.1	0.0014	0.1

The carrier probability for II-3, if she has the genotype BC and I-2 and II-3 have normal CK values and no detectable mutation, equals (2/135+2/270+0.0014)/(2/135+2/270+0.1+0.0014+0.1)=0.106.

Appendix

Tables I - IV are reprints of Van der Meulen et al. (1995)

Table I. Recurrence risk for gonadal mosaics with one affected child (a=1) and a variable number of untyped unaffected children (b) and typed unaffected children with the risk chromosome (c).

Recurrence Risk a = 1		Number of typed unaffected children with the risk chromosome (c)				
		0	1	2	3	4
Number of untyped unaffected children (b)	0	0.048	0.018	0.013	0.010	0.008
	1	0.033	0.015	0.012	0.009	0.008
	2	0.025	0.013	0.010	0.008	0.007
	3	0.019	0.012	0.009	0.008	0.007
	4	0.016	0.011	0.009	0.007	0.006

Table II. Recurrence risk for gonadal mosaics with two affected children (a=2) and a variable number of untyped unaffected children (b) and typed unaffected children with the risk chromosome (c).

Recurrence Risk a = 2		Number of typed unaffected children with the risk chromosome (c)				
		0	1	2	3	4
Number of untyped unaffected children (b)	0	0.333	0.143	0.120	0.100	0.084
	1	0.286	0.133	0.111	0.093	0.079
	2	0.240	0.124	0.103	0.087	0.074
	3	0.200	0.115	0.096	0.081	0.069
	4	0.168	0.107	0.089	0.075	0.065

Table III. Recurrence risk for gonadal mosaics with three affected children (a=3) and a variable number of untyped unaffected children (b) and typed unaffected children with the risk chromosome (c).

Recurrence Risk a = 3		Number of typed unaffected children with the risk chromosome (c)				
		0	1	2	3	4
Number of untyped unaffected children (b)	0	0.429	0.200	0.183	0.164	0.145
	1	0.400	0.194	0.175	0.156	0.138
	2	0.366	0.187	0.168	0.149	0.131
	3	0.328	0.179	0.160	0.141	0.124
	4	0.290	0.172	0.153	0.134	0.118

Table IV. Recurrence risk for gonadal mosaics with four affected children (a=4) and a variable number of untyped unaffected children (b) and typed unaffected children with the risk chromosome (c).

Recurrence Risk a = 4		Number of typed unaffected children with the risk chromosome (c)				
		0	1	2	3	4
Number of untyped unaffected children (b)	0	0.467	0.226	0.216	0.203	0.188
	1	0.452	0.222	0.211	0.197	0.181
	2	0.431	0.218	0.206	0.191	0.175
	3	0.406	0.214	0.200	0.185	0.168
	4	0.375	0.209	0.195	0.178	0.161

II-3.2 Calculation of Recurrence Risk in case of possible Mosaicism: multiple generation pedigrees.

Martin A. van der Meulen[1], Gerard J. te Meerman[1] and Lodewijk A. Sandkuijl[123].

[1]Department of Medical Genetics,
University of Groningen,
A. Deusinglaan 4,
9713 AW Groningen,
The Netherlands.

[2]Institute of Clinical Genetics,
Erasmus University,
Rotterdam,
The Netherlands.

[3]Department of Human Genetics,
Leiden University,
Leiden,
The Netherlands.

phone : +31 50 3632925
fax : +31 50 3632947
email : m.a.van.der.meulen@med.rug.nl
keywords: mosaicism, recurrence risk, Duchenne muscular dystrophy

Abstract

As an extension to Van der Meulen et al (submitted) manual multiple generation recurrence risk calculations are presented including the risk due to mosaicism, in multiple persons.

Introduction

Van der Meulen et al. (submitted) describe how in nuclear families a recurrence risk can be calculated, taking into account the risk due to somatic and germline mosaicism.

The main difference between recurrence/carrier risk in nuclear pedigrees and multiple generation pedigrees is that next to the possibility of a mutation entering the pedigree through a carrier, a new mutation can occur not in one but in multiple predecessors of an affected individual. Calculations presented are based on the same assumptions as made by Van der Meulen et al. (submitted). Some of the important assumptions are mentioned briefly:

- Figure 1 shows the mutation selection equilibrium in which all mutation occur in the parent(s) and none in the child. The mutation selection equilibrium is based on an equal male and female mutation frequency, which is also assumed.
- Ratio between origin of somatic versus germline mosaicism equals 1/3 vs 2/3, based on an equal mutation rate in each mitosis (Luria and Delbruck, 1943) and a recurrence risk of 20% for a new mutation.
- Recurrence risk of germline mosaicism is based on Van der Meulen et al. (1995), who calculate recurrence risk for different compositions of nuclear families, including the use of (intragenic) DNA-marker information.
- Recurrence of the observed mutation is calculated, taking into account different origins of the mutation within persons (carrier founder or new mutation) and between persons.

Example / Results

All calculations are based on a pedigree where the sister of the consultand has a son, who has the X-linked recessive disease Duchenne Muscular Dystrophy (DMD). Genotypes of the intragenic marker vary between the examples, recombination is excluded.

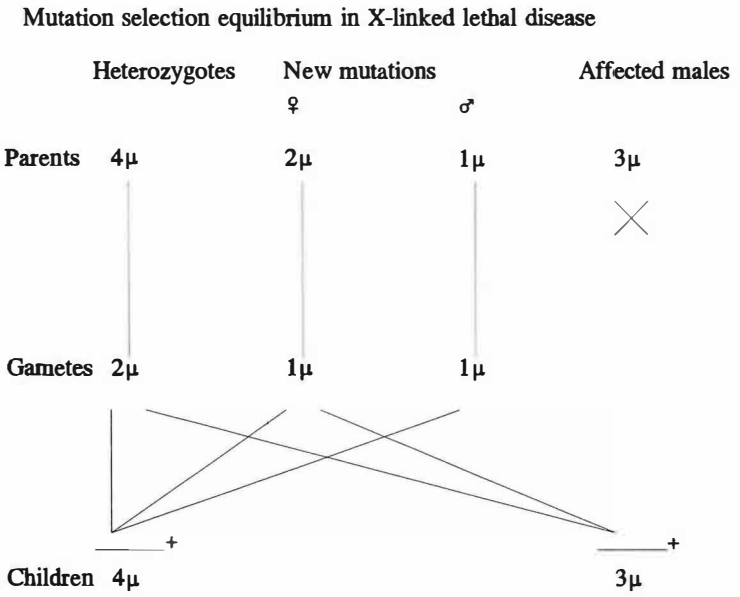


Figure 1 Mutation selection equilibrium

1-Possible grandmaternal mosaicism

Table 1 shows the recurrence risk calculation in case the grandmaternal chromosome is inherited by the affected grandson. The carrier probability according to mutation selection equilibrium for a founder female in a X-linked recessive disease equals 4μ and the probability of a new mutation equals 2μ (Figure 1). In table 1 the prior carrier probability for I-2 equals 2μ and the total new mutation probability equals μ, due to the knowledge that only her carrier ship on the chromosome with the intragenic marker allele A and a new mutation on this specific chromosome is relevant for this calculation. The same rationale can be applied to the probability of a new mutation in II-2.

2-Possible grandpaternal mosaicism

Table 2 shows the recurrence risk calculation in case of possible grandpaternal origin of the observed mutation. In this example we are certain that a new mutation occurred in either the grandfather or the mother of the affected child.

In DMD can be argued that the probability of a male to be a somatic mosaic is reduced, because he is unaffected. When we assume that the unaffected male grandfather cannot be

a somatic mosaic, the carrier risk for II-3 is changed considerable. In this calculation the effect of the reduced somatic fraction in unaffected males on mutation selection equilibrium is ignored.

3-Possible grandpaternal and/or grandmaternal mosaicism

a-marker genotype of affected boy unknown

In table 3 the recurrence risk is calculated in case of unknown marker genotype for the affected boy. In the calculation in table 3 the probability of a new mutation in II-2 is not restricted to μ in contrast to table 1, as a mutation on both chromosomes can lead to the affected phenotype of III-1, because the origin of the mutation carrying chromosome is unknown. The transmission probability of the mutated chromosome equals 1/2 for all columns and is therefore left out of the table. The effect on the carrier risk for the consultand of the restricted mutation probability in the grandfather through restriction of his somatic mosaic probability is shown again.

In this calculation the consultand inherits a high and a low risk chromosome. In table 1 the risk for each chromosome is also calculated separately (no somatic mosaic in grandfather). Note that when somatic mosaicism is excluded for unaffected males, the ratio of paternal vs maternal origin of the mutation equals $2/3\mu : (2\mu + 1/3\mu + 2/3\mu) =$ 2:9 in stead of the usual 1:3 in mutation selection equilibrium. The carrier risk ratio of paternal vs maternal origin after one carrier daughter equals for the next daughter $2/3\mu*2*0.048 : (2\mu*0.5 + 1/3\mu*0.5 + 2/3\mu*0.048) = 2:37$. When this offspring is proven to be carrier of the maternal risk chromosome this ratio increases to $2/3\mu*2*0.048 : (2\mu + 1/3\mu + 2/3\mu*2*0.048) = 1:37$. This ratio is the same as in table 3. Although prenatal exclusion of disease inheritance is not possible, the risk of the grandpaternal chromosome is much smaller, so prenatal selection might be considered.

b-marker genotypes of grandparents unknown

The grandparental origin of the chromosome carrying the mutation and marker allele C is unknown in the calculation in table 4. In this example a non-affected boy II-4 with the other grandmaternal chromosome is added to exclude homozygosity of the grandmother I-2. Mendelian a priori probabilities for origin of the mutation carrying chromosome of the affected boy are used. Again the effect of reduced somatic mosaicism probability in

the grandfather is shown.
In contrast to the preceding example, prenatal screening is possible, because the risk chromosome can be recognized, although the grandparental origin of the risk chromosome is unknown.

Note that when both the genotypes of grandparents and the genotype of the affected boy are unknown, the total carrier risk for the consultand II-3 is the same as in the calculation in tables 3 and 4, only the risk chromosome is unknown.

All examples can easily be extended with phenotypic information as described in Van der Meulen et al (submitted).

Discussion

Although the presented calculations can be used for practical genetic counselling, the main purpose of this manuscript is the presentation of the methods used. Therefore the resulting carrier risks are not discussed. As in the preceding paper the presented framework is generally applicable, but improved insight in underlying biological mechanisms will probably change the model parameters. Therefore, whenever this method is used in genetic counselling, we advise not only to calculate the recurrence/carrier risk under the presumed model, but also investigate the sensitivity of the calculated risk to variation in model parameters. This is for instance shown in table 2, where the effect of reduction in the somatic mosaic probability of the grandfather is shown to result in a decreased carrier risk for the consultand.

In the presented calculations male and female mutation rates are assumed to be equal. Muller et al. (1992) calculates the ratio between male and female mutation rates in Duchenne Muscular Dystrophy (DMD) in a sample size 295 to be very close to 1. Later Grimm et al. (1994) show deletions and point mutations may have different origin in DMD. Passos-Bueno et al. (1992) hypothesize that different mosaicism frequencies in DMD may exist for proximal and distal mutations, indicating difference in etiology and recurrence risk. Karel et al. (1986) discuss the power of proving differences between male and female mutation frequencies. As has been shown for other phenotypic

observations (Van der Meulen et al, submitted) differences in male and female mutation rates can easily be implemented in the manual calculation by implementing different mutation frequencies for the male and female chromosomes carrying a mutation. In X-linked recessive diseases the reduced observed mutation frequency in unaffected males can possible be explained as a result of a reduced probability of mutation in the somatic phase. Note that the mutation selection equilibrium changes when different mutation frequencies are assumed in males and females.

All presented calculations are independent of the absolute mutation frequency, because the mutation frequency appears in both the nominator and the denominator of the a posteriori carrier risk and therefore cancels out. When multiple mutations are considered in the calculations, this will no longer be the case, but because the mutation frequency is very small this will hardly change the calculated recurrence/carrier risk.

Recurrence risk/carrier calculations in multiple generation pedigrees is elaborate, but possible, manually. Recurrence risk calculation on more complicated pedigrees is not practically feasible. The presented calculations are used to check the computer program in which the described model is implemented. More information concerning the computer program can be obtained through the first author. The computer program is based on a linkage program (Te Meerman, 1990) and therefore capable of handling recombination and multiple marker problems of any pedigree structure.

In practical genetic counselling, the probability of finding a recombination within the approximately 12 cM long DMD gene is far from unlikely. When a mutation can not be found in the affected and a non-mosaic and non (double) recombinant segregation pattern is available the non-mosaic interpretation will be most important, because this interpretation will lead to the biggest likelihood. On the other hand, when a non-mosaic model is used in case of a pedigree, where recombinants are needed to explain the observed segregation pattern, erroneous scoring of recombinants will result (Bakker et al., 1989). One must realize that the probability of observing a recombinant depends on the recombination fraction but is generally more likely then a multiple mutation explanation.

In DMD CK values are often elevated in female carriers. The affect on the number of mitosis in the unaffected male available for mutation is discussable. This leads to the question what to expect of the CK value of early somatic mosaics in females. The effect when a high CK value is encountered will be limited, but normal CK values will increase the probability of a new mutation and within the recurrence risk of a new mutation the main recurrence risk is due to the somatic mosaics.

In calculations concerning mosaicism, the X-linked recessive disease DMD has mainly been used as an example. Using the same framework, recurrence/carrier risk calculation can also be performed for autosomal diseases.

Acknowledgements

This work is supported by the Netherlands Organisation for Scientific Research (NWO).

References

Bakker E, Veenema H, Den Dunnen JT, Van Broeckhoven C, Grootscholten PM, Bonten EJ, Van Ommen GJB and Pearson PL. Germinal mosaicism increases the recurrence risk for 'new' Duchenne muscular dystrophy mutations. J Med Genet 1989; 26:553-559

Grimm T, Muller B, Muller CR and Janka M. Theoretical considerations on germline mosaicism in Duchenne muscular dystrophy. J Med Genet 1990; 27:683-687

Grimm T, Meng G, Liechti Gallati S, Bettecken T, Muller CR and Muller B. On the origin of deletions and point mutations in Duchenne muscular dystrophy; most deletions arise in oogenesis and most point mutations result from events in spermatogenesis. J Med Genet 1994; 31:183-6

Karel ER, Te Meerman GJ, Ten Kate LP. On the power to detect differences between male and female mutation rates for Duchenne muscular dystrophy, using classical segregation analysis and restriction fragment length polymorphisms. Am J Hum Genet 1986; 38:827-840

Luria SE and Delbruck M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 1943; 28:491-511

Muller B, Dechant C, Meng G, Liechti-Galatti S, Doherty RA, Hejtmanchik JF, Bakker E et al. Estimation of the male and female mutation rates in Duchenne muscular dystrophy (DMD) Hum Genet 1992; 89:204-206

Passos-Bueno MR, Bakker E, Kneppers ALJ, Takata RI, Rapaport D, Dunnen JT, Zatz M and Ommen GJB. Different Mosaicism Frequencies for Proximal and Distal Duchenne Muscular Dystrophy (DMD) Mutations Indicate Difference in Etiology and Recurrence Risk. Am J Hum Genet 1992; 51:1150-1155

Te Meerman GJ. A logic programming approach to pedigree analysis. Thesis publishers Amsterdam, 1991.

Van der Meulen MA, Van der Meulen MJP and Te Meerman GJ. Recurrence risk for germinal mosaics revisited. J Med Genet 1995; 32:102-104

Van der Meulen MA, Te Meerman GJ and Sandkuijl LA. Calculation of Recurrence Risk in case of possible mosaicism, submitted.

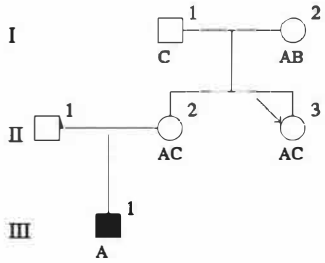
Van Essen J, Busch HFM, Te Meerman GJ, Ten Kate LP. Birth and population prevalence of Duchenne muscular dystrophy in the Netherlands. Hum genet 1992 88:258-266

Van Essen J, Abbs S, Baiget M, Bakker E, Boileau C, Van Broeckhoven C et al. Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: a european study. Hum Genet 1992; 88:249-257

Possible grandmaternal origin. Two daughters with the same genotype, II-2 has an affected son, II-3 is consultand.

Table 1:
Carrier probability calculation for person II-3, using the intragenic marker as shown in the pedigree. Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

I-2	No Carrier				Carrier		Som. Mos.		Germl. Mos	
Prior allele A:	1-3μ				2μ		1/3μ		2/3μ	
Conditional II-2 allele A	Som. Mos. 1/3μ		Germl. Mos. 2/3μ		Carrier 1		Carrier 1		Carrier 1	
II-3: allele A	Car+ 0	Car- 1	Car+ 0	Car- 1	Car+ 1	Car- 0	Car+ 1	Car- 0	Car+ 2*0.048	Car- 1-2*0.048
Joint	0	1/3	0	2/3	2	0	1/3	0	2/3*(2*0.048)	2/3*(1-2*0.048)



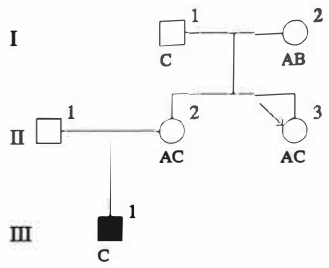
In the joint all columns are divided by μ, and 1-3μ is assumed to be equal to 1.

The carrier probability for II-3 equals:
(2+1/3+2/3*(2*0.048))/(1/3+2/3+2+1/3+2/3*(2*0.048)+2/3*(1-2*0.048))=2.397/4.0=0.5993

Possible grandpaternal origin. Two daughters with the same genotype, II-2 has an affected son, II-3 is consultand.

Table 2:
Carrier probability calculation for person II-3, using the intragenic marker as shown in the pedigree. Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

I-1	No Carrier				Som. Mos.		Germl. Mos	
Prior:	$1-\mu$				$1/3\mu$ (0)		$2/3\mu$	
Conditional II-2 allele C	Som. Mos. $1/3\mu$		Germl. Mos. $2/3\mu$		Carrier 1		Carrier 1	
II-3: allele C	Car+ 0	Car- 1	Car+ 0	Car- 1	Car+ 1	Car- 0	Car+ 2*0.048	Car- 1-2*0.048
Joint	0	1/3	0	2/3	1/3 (0)	0	2/3*(2*0.048)	2/3*(1-2*0.048)



In the joint all columns are divided by μ, and $1-\mu$ is assumed to be equal to 1.

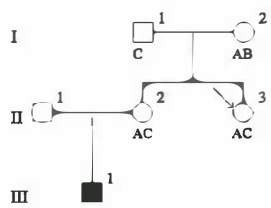
The carrier probability for II-3 equals:
$(1/3+2/3*(2*0.048))/(1/3+2/3+1/3+2/3*(2*0.048)+2/3*(1-2*0.048))=0.20$.

When we assume that I-1 can not be a somatic mosaic, the carrier probability for II-3 equals:
$(2/3*(2*0.048))/(1/3+2/3+2/3*(2*0.048)+2/3*(1-2*0.048))=0.038$.

Genotype of affected child unknown, maternal, grandmaternal and grandpaternal origin of mutation possible

Table 3:
Carrier probability calculation for person II-3, using the intragenic marker as shown in the pedigree. Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

I-1	No Carrier				No Carrier						Som.Mos		G.Mos.	
Prior allele C:	1-μ				1-μ						1/3μ (0)		2/3μ	
Conditional														
I-2	No Carrier				Carrier		Som. Mos		G.Mos.		No Carrier			
Prior allele A:	1-3μ				2μ		1/3*1μ		2/3*μ		1-3μ			
Conditional														
II-2	Som. Mos.		Germl. Mos.		Carrier						Carrier			
allele A and C	2*1/3μ		2*2/3μ		1						1			
II-3:	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-
allele A and C	0	1	0	1	1	0	1	0	2*0.048	1-2*0.048	1	0	2*0.048	1-2*0.048
Joint	0	2/3	0	4/3	2	0	1/3	0	2/3*M	2/3*(1-M)	1/3 (0)	0	2/3*M	2/3*(1-M)



M=2*0.048

In the joint all columns are divided by μ, and 1-μ and 1-3μ are assumed to be equal to 1.

The carrier probability for II-3 equals:
$\{2+1/3+2/3*M+1/3+2/3*M\}/\{2/3+4/3+2+1/3+2/3*M+2/3*(1-M)+1/3+2/3*M+2/3*(1-M)\}=2.79/6.0=0.465$

When we assume that I-1 can not be a somatic mosaic, the carrier probability for II-3 equals:
$\{2+1/3+2/3*M+2/3*M\}/\{2/3+4/3+2+1/3+2/3*M+2/3*(1-M)+2/3*M+2/3*(1-M)\}=\{2.79-1/3\}/\{6.0-1/3\}=0.434$

Carrier probability for II-3 if the mutation on chromosome with allele A (I-1 can not be somatic mosaic)
$\{2+1/3+2/3*M\}/\{2/3+4/3+2+1/3+2/3*M+2/3*(1-M)+2/3*M+2/3*(1-M)\}=\{2.39\}/\{6.0-1/3\}=0.423$

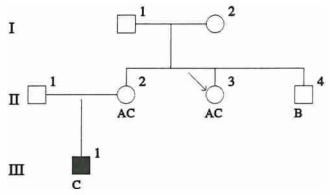
Carrier probability for II-3 if the mutation on chromosome with allele C (I-1 can not be somatic mosaic)
$\{2/3*M\}/\{2/3+4/3+2+1/3+2/3*M+2/3*(1-M)+2/3*M+2/3*(1-M)\}=\{2/3*2*0.048\}/\{6.0-1/3\}=0.011$

Genotype of affected child C. Maternal, grandmaternal and grandpaternal origin of mutation possible. Genotypes of grandparents unknown.

Table 4:
Carrier probability calculation for person II-3, using the intragenic marker as shown in the pedigree. Car+ and Car- refers to carriers and non-carriers respectively, X equals impossible or probability 0.

	II-2				I-2						I-1			
Origin of allele C	1				1/2						1/2			
I-1	No Carrier				No Carrier						Som.Mos		G.Mos.	
Prior allele A or C:	1-μ				1-μ						1/3μ (0)		2/3μ	
Conditional I-2	No Carrier				Carrier		Som. Mos		G.Mos.		No Carrier			
Prior allele A or C:	1-3μ				2μ		1/3*1μ		2/3*μ		1-3μ			
Conditional II-2	Som. Mos.		Germl. Mos.		Carrier						Carrier			
allele A and C	1/3μ		2/3μ		1						1			
II-3:	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-	Car+	Car-
allele A and C	0	1	0	1	1	0	1	0	2*0.048	1-2*0.048	1	0	2*0.048	1-2*0.048
Joint	0	1/3	0	2/3	1	0	1/6	0	2/6*M	2/6*(1-M)	1/6 (0)	0	2/6*M	2/6*(1-M)

M=2*0.048



In the joint all columns are divided by μ, and 1-μ and 1-3μ are assumed to be equal to 1.

The carrier probability for II-3 on allele C equals:
$\{1+1/6+2/6*M+1/6+2/6*M\}/\{1/3+2/3+1+1/6+2/6*M+2/6*(1-M)+1/6+2/6*M+2/6*(1-M)\}=1.40/3.0=0.465$

When we assume that I-1 can not be a somatic mosaic, the carrier probability for II-3 on allele C equals:
$\{1+1/6+2/6*M+2/6*M\}/\{1/3+2/3+1+1/6+2/6*M+2/6*(1-M)+2/6*M+2/6*(1-M)\}=\{1.40-1/6\}/\{3.0-1/6\}=0.434$

II-4 Risk calculation in the possible presence of mosaicism, a general computer program.

Martin A. van der Meulen and Gerard J. te Meerman

Department of Medical Genetics
University of Groningen
A. Deusinglaan 4
9713 AW Groningen
The Netherlands

phone : +31 50 3632925
fax : +31 50 3632947
email : m.a.van.der.meulen@med.rug.nl g.j.te.meerman@med.rug.nl

Abstract

A linkage program is presented with an extension for somatic and germline mosaicism. Recurrence and/or carrier risk calculation, including the risk due to inheritance of disease alleles through mosaicism, is possible for pedigrees of any structure and size.

Introduction

Recurrence risk calculation in disorders which are known to have high mutation frequencies, resulting in mosaicism, are complicated. Winter (1980) introduced a way of calculation of genetic risks in X-linked recessive conditions using programmable calculators. He emphasises on the calculation for a female to be carrier. Sarfarazi and Williams (1986) presented a computer programme for the calculation of the recurrence risk in X linked disorders, combining pedigree and DNA probe and other conditional information. In their discussion they state that the calculations become complex when DNA data are included and especially when new mutation at the disease locus is a possibility. An obvious solution is a computer programme. From their examples can be deduced that the recurrence risk due to a new mutation is assumed to be zero. Another computer programme was written by Clayton (1986), who discussed the problem of testing a computer program. With simple pedigrees, the results can be checked by hand. On the other hand, pedigrees of even moderate complexity can involve a thousand or more possible permutations of alleles in their full description. Each permutation can involve 20 or more separate calculations. It is not possible to check such a quantity by hand. The programs described above can deal with recombination and the recurrence risk due to a new mutation is assumed to be zero.

Te Meerman (1991) illustrates the technique to calculate recurrence risk in the presence of mutation. Germline mosaicism is a special case of mutation, because not all of the germ cells carry the mutation. He discussed also how minimum recombinant programs can be used to interpret pedigrees, where many recombinants including double recombinants can be explained by the mosaic interpretation.

Jeanpierre (1992) presented the first computer programme in which recurrence risk due to germline mosaicism is taken into account. He studied the effect of the model and the

effect of differences between male and female mutation frequencies. He concludes that germinal mosaicism should only be considered where a recent mutation is likely to have occurred, which is not an infrequent situation, since one-third of the mothers of DMD boys and two-third of the grandmothers, have not inherited a defective gene.

We presented recently extension of a model, originally presented by Hartl (1971), to include DNA marker information in the recurrence risk for germinal mosaicism (Van der Meulen et al., 1995). This was later followed by a method to use the preceding model in manual recurrence risk calculations in pedigrees of arbitrary structure (Van der Meulen et al., submitted). We indicated that recurrence risk calculation is virtually impossible in complex pedigrees, because the number of possibilities becomes to large. This is specially the case when multiple generations are involved or when information is available of extragenic DNA markers, making corrections for recombination necessary. We introduce here the computer program GRONLMOS, which is an extension of a general linkage program GRONLOD (te Meerman, 1991). This computer program is extended to calculate recurrence risk in case of possible mosaicism. The program calculates recurrence risk, taking also into account multiple independent mutations in multiple persons. The program is in principle capable of handling pedigrees of any structure and size.

From manual calculation to computer program

In manual calculations taking into account the recurrence risk due to mosaicism, all different origins of the mutation are evaluated in separate columns. All entries in a column are multiplied and this will give the joint likelihood for a specific column, or mutation origin. By dividing the likelihood of one or the sum of multiple columns by the sum of all column joint likelihoods, Bayesian probability can be calculated. In complex pedigrees, with multiple probable origins of a mutation, the manual calculations become elaborate and complex.

The same method as applied in the manual calculations can be applied in a computer program, by subsequently calculating the likelihood for all origins of the mutation and their segregation patterns, followed by calculation of the bayesian risk by summing of

appropriate likelihoods divided by the sum of all likelihoods. In complex pedigrees the number of likelihoods to evaluate increases rapidly, specially when also recombination is taken into account.

Te Meerman (1991) discusses in detail how in linkage programs a pedigree can be peeled. Peeling is the process in which a pedigree is subdivided in separate to analyze cutsets. Cutsets are connected through articulation persons. In a cutset the likelihoods for genotypes for the articulation person are calculated. These likelihoods are subsequently used in the next cutset, until the likelihoods of genotypes of the last cutset are calculated. Resulting likelihoods of genotypes of the consultand are used to calculate Bayesian probabilities. Peeling is performed because it divides the problem in easy to analyze pieces and it is memory efficient, because the storage of genotypes and attached results for articulation persons only increases linearly with the number of genotypes and individuals.

The mosaicism model defined by Van der Meulen et al (submitted) can easily be implemented in a way compatible with peeling. From allele frequencies the carrier probability can be calculated for founders. While assuming that the mutation frequency is equal for each mitosis and that the occurrence of an observed new mutation equals the mutation frequency μ, the probability of a mutation in each gonadal generation can be calculated. The number of mitoses in each generation is known and the mutation frequency per mitosis can be calculated as the total mutation frequency divided by the number of cell generations between conception and the final meiotic cell division, which result in gametes. Since the number of mitoses in late gonadal generations is high, the occurrence of multiple independent mutations is likely. A priori probabilities for genotypes are set using these mutation probabilities for normal alleles.
From the knowledge that the probability of observing a mutation which occurred in a specific generation equals 1 out of 2 to the power of the cell generation of mutation after generation of the germline stemcell, probabilities of encountering a new mutations which occurred in a specific gonadal generation can be calculated.

Disadvantage of this implementation is in comparison to the manual method, that the concept of recurrence risk due to the same mutation is abandoned, because multiple independent mutations become possible. When the mutation frequency is low the likelihood of multiple independent mutations is still low in comparison to the one mutation explanation. Edwards (1989) concludes from extensive observations available that it is unlikely that the average mutation rate per locus per generation exceeds 10^{-5} or even 10^{-6}.

When mosaicism is defined as above, the model can be seen as an extension of the Thompson model of probability of any genealogy with separate definitions for genotype at conception and the genotype passed on to the next generation at reproduction.

$$\sum_{\text{all genotype combinations}} \Big[\prod_{\text{founders}} P(\text{genotype}_{\text{conception}}) * \prod_{\text{nonfounders}} P(\text{genotype}_{\text{conception}} | \text{genotype parents}_{\text{reproduction}}) * \prod_{\text{parents}} P(\text{genotype}_{\text{reproduction}} | \text{genotype}_{\text{conception}}) * \prod_{\text{observed individuals}} P(\text{phenotype} | \text{genotype}) \Big]$$

The implemented model

In GRONLMOS the number of gonadal generations after formation of the germline stemcell is fixed to 20. A separate class is introduced for somatic mosaics. All other parameters can be chosen in the input file. When for instance mutation selection equilibrium is required, allele frequency for the disease allele and mutation frequency must be chosen accordingly. More information on the implementation of the model will be given after the description of the input of the program GRONLMOS

Program Standard Input

Knowledge of the standard input for GRONLOD, as described by te Meerman (1991) and in 'Introduction to computer methods for risk analysis in genetic counselling' (available from one of the authors), is presumed. Described statements are specific statements for GRONLMOS and should be added on top of a normal input file for GRONLOD.

Common errors in standard input:

- in X-linked diseases the male recombination frequency has to be set to 0.5 (*rec_male([0.5,...])*), the Y-chromosome is given as allele -1. The -1 allele does not have to be specified in the *allele_frequencies* statement.
- the disease locus can best be specified by *phen_gen* statements (probability of the phenotype given the genotype). The disease locus should be defined in the *fenotype* statement as 0,0. The male *fenotype* in X-linked diseases must be -1,0.

GRONLMOS specific input

- *mut_locus(i)*

The mutation locus statement defines the mutation to be located at the i-th locus of the fenotype statements. Strict rules have to be followed concerning the mutation locus as described in the fenotype statement. The normal allele has to be coded as allele 1, the disease allele as allele 2. The program codes the somatic mosaics as allele 3, the 100% germline mosaic as allele 4, the 50% germline mosaic as allele 5,...as allele 23. In mutation persons, the normal allele 1 is recoded as 25. The mosaic alleles are never observed, which means that they only exist as theoretical entities in the program and in the datafile. This means also that they are not defined in the allele_frequencies statement, as defined to fill in unknown alleles.

- *selection([locus1,locus2,...])*

The selection statement is described in the standard GRONLOD input. GRONLMOS is an extension of the linkage program GRONLOD. This gives the advantage that it is possible to include DNA markers in the calculation. However, if there is no information on DNA markers, we have to add a non-informative DNA marker with 1 allele. Recombination between the marker and the disease locus can be set to 0. The selection statement always has to contain at least two loci, including the *mut_locus(i)*.

- *mut_person(["Name1",...])*

For the persons Name1 and others who are in the list L of mutation persons all mosaic genotypes are evaluated. To keep calculation time as low as possible, people who can not be the cause of the mutation, are left out of the list L.

- *allele_frequencies(mut_locus,[normal,disease]) and mut_freq(μ)*

Allele frequencies for the mutation locus have to be defined. From the mutation selection equilibrium in X-linked recessive diseases follows, that the carrier frequency (2q) is equal to four times the mutation frequency (μ). From Hardy Weinberg equilibrium follows that the allele frequency (q) in case of low allele frequencies equals approximately half the carrier frequency (2q). Conclusion: the allele frequency q equals twice the mutation frequency μ. In the papers by Van der Meulen et al. (1995, submitted) is described, that the calculation is independent on the mutation frequency, because the calculation is conditional on the mutation. In the computer program, however, independent mutations in independent individuals are possible. The mutation frequency is used to compare multiple mutations against single mutation explanations. Note that it is not possible to use *vector_frequencies* for the *mut_locus(i)*.

- *soma_frac(i)*

The somatic fraction defines the fraction i of the mutations descending from a mutation in one of the mitoses before the germ line stem cell is formed. A mutation in the somatic phase (in the line of which the germline stem cell is formed) leads to 100% cells in the germ line carrying the mutation and therefore leading to a recurrence risk of 50% after meiosis. The somatic fraction is set to 1/3 by Van der Meulen et al. (submitted), considering 10 mitoses before the germline stem cell is formed and 20 mitoses in the germline. Reduction of the somatic fraction can be based on as well, an assumption of less then 10 somatic mitoses, as on more mitoses in the germline. As indicated by van der Meulen et al. (1995) the recurrence risk changes minimally when more then 20 mitoses are considered in the germline.

- *phen_gen(mut_locus,allel1,allel2,"phenotype",probability)*

The phen_gen statement is described in the basis manual. The phen_gen statements defines the probability of a phenotype given the genotype. It is not necessary to define different phen_gen and phenotype statements for persons in the mut_person list. Mosaic states are also possible genotypes and can be controlled with phen_gen statements. The phen_gen statements of the mosaic states are by default set to 1. In special cases, control over specific genotypes can be useful. For instance in Duchenne Muscular Disease

(DMD), a X-linked recessive lethal disease, a mutation in one of the first mitoses in males will probably lead to the affected phenotype. But in females a mutation during these mitoses will not change the phenotype, except maybe for the CK level. This can be handled by reducing the probability of a mutation in the somatic phase in males. For instance when we assume 10 mitoses in the somatic phase (*soma_frac(0.33)*) and we want to exclude mutation in the first 5 mitoses, the statement *phen_gen(mut_locus, -1,3, "unaffected male",0.5)* reduces the somatic probability of 'unaffected males' in an X-linked disease to 50% of the original value, without changing the model for other phenotypes. Consequence is that, in order to keep the model complete, the affected males can be affected due to a mutation in these first five mitoses. There are two ways of handling this.

1- An Affected male gets a probability of being affected while he inherits a normal allele from his mother of 5/10 * soma_frac * mut_freq. This approach is only possible when affecteds produce no offspring.

2- The affected male is defined as a mutation person. Phen_gen statements for all mosaic genotypes must be added to the input file. The phen_gen statement of the somatic mosaic state equals 0.5, all others are 0.

This phenomenon can be the cause of the difference in observed mutation frequency in males and in females.

- *risk_person("Name1")*

The risk person Name1 is the person for whom the bayesian risks of genotypes are calculated. When the consultand is a mutation person in the calculation, the program will give bayesian probabilities for all mosaic states. Probabilities of mosaic states are hard to interpret. Easier is to add an extra child to the calculation and defining him/her as the risk person. (remember that in X-linked diseases the sex of the child is determined by the given fenotype).

Modelling mosaicism, a 2 step procedure

In the manual calculations, recurrence risks for germline mosaicism is taken from the look up tables by Van der Meulen et al. (1995). The computer program calculates likelihoods for a mutation to have happened in a specific gonadal generation. The

calculation of these likelihoods is a 2 step procedure, in which first a priori likelihoods for a mutation in a specific gonadal generation is set (procedure mutation1) and the second step, in which the likelihood of a specific gonadal generation is subsequently multiplied for each offspring by the probability of inheriting the genotype from this gonadal generation (procedure mutation2). The probability to inherit a mutation from a specific gonadal generation is the frequency of mutated gametes, which results when the mutation occurred in that gonadal generation. The rationale of this approach can easily be understood with an example: the likelihood of a late gonadal generation mutation is low, when multiple affected offspring are observed, but will be high if one affected and multiple non-affected offspring are observed. As in the manual calculations will likelihoods of columns with impossible segregation patterns become zero. This is for instance the case for the likelihood of no mutation (normal alleles) for the parents, when affected offspring is observed. This does not mean that a normal allele can not be inherited, because a child will not be affected when a normal allele is transmitted from a gonadal generation, where there is as well a probability of receiving a normal allele as a mutated allele. The probability of transmission of a normal allele varies between zero (somatic mosaic, first germline gonadal generation mutation and carrier) and almost 1 (last (20) germline gonadal generation mutation). The final likelihoods of the gonadal generation mutations is used to weigh the origin of the transmitted allele.
Above for simplicity is assumed that each transmission involved the same chromosome and not mentioned that autosomes and X-chromosomes in females are available in duplicate. Therefore we can distinguish between risk chromosome and the other chromosome.

Description of the algorithm in PROLOG

When the program is started, the data file is read into the memory database. Using the mut_freq and other specific input from the data file, the probabilities for step1: mut_1 and step2: mut_2 are calculated and stored in the database (procedure make_mut_freq).
mut_1 is defined as the probability of a mutation to occur in gonadal generation i. μ is defined as the total mutation frequency. The probability of somatic mosaic (a mutation before the germline stem cell is formed) is

Prob = μ * soma_frac

The probability of a mutation to occur in one of the 20 gonadal cell generations is:

$$Prob_i = 2^i * 1/N * \mu \text{ (1-soma\_frac)}$$

The total number of gonadal generations N in the germline is fixed in the program at 20. The sum of all probabilities mut_1 equals 1, because the rest probability after mutation is the probability of a normal allele not to be involved in a mutation.

mut_2 is defined as probabilities of as well inheriting a normal allele as a mutated allele when a mutation occurred in that gonadal generation. The sum of these two probabilities equals 1 for each gonadal generation i.

$$Prob_{\text{mutated allele}} = 2^{-(i)}$$

$$Prob_{\text{normal allele}} = 1 - Prob_{\text{mutated allele}}$$

These database entries are calculated for all possible gonadal generation mutations.

Then the analysis of the pedigree can begin. The program will start with the generation of genotypes: for founders from phenotype and allele frequencies for unknown alleles, non founders for all possible genotypes. When likelihoods for genotypes are set, all possible genotypes are evaluated consecutively. For mosaic persons the normal allele at the disease locus will mutate (mutation1), using database entries mut_1. Then the pedigree is peeled, where the genotype of each offspring translates into a multiplication of the likelihood for each genotype with the probability of transmitting that genotype (mutation2, using database entries mut_2).

The cutset is peeled upon the articulation person, where there are two possibilities:
Peeling down: all likelihoods of possible genotypes (also the gonadal generation mutation genotypes) will be multiplied by the probability to transmit a mutated allele and to transmit a normal allele for that genotype (database entries mut_2), to get the final result: the total likelihood of transmitting a normal and the total likelihood of transmitting a mutated allele. These results can be used as input for the next part of the pedigree to analyze or to calculate the recurrence/carrier risk for the consultand.
Peeling up: new mutations are inherited from parents as normal alleles, so the likelihood of all mosaic genotypes and normal alleles can be added to get the likelihood of the reception of a normal allele from parents (procedure mutate_back).

Note that in a mut_person for both available chromosomes (autosomes or female X-chromosomes) all mosaic genotypes are formed. The mosaic genotypes will however have only an important influence on recurrence risk, when a mutation is observed. This approach makes testing of mutation selection equilibrium possible.

High mutation frequencies

When the mutation frequency per mitosis is high and there are many mitoses, which is the case in for instance the last gonadal generation in the germline, gametes with independent mutations will likely be present. This also means that there will not be individuals who are not a carrier for this disease in some of their gametes. The procedure mutation1 in the model used in GRONLMOS is based on the probability of a mutation happening in a specific cell generation. This probability doubles for each generation, since the number of mitoses in each gonadal cell generation doubles and the mutation frequency per mitosis is assumed to be constant. After $1/\mu_{\text{per mitose}}$ mitoses, the mutation leading to the highest frequency of mutated gametes has occurred. Other mutations are neglected in the calculation, so when high mutation frequencies are used, the number of germline gonadal generations is reduced. The effect on result of the calculation is neglectable, which could be expected, because very low grade mosaicism has always very low recurrence risks and therefore neglectable influence on results. However, this implies that recurrence risk is calculated assuming that the affected is due to the most frequent or earliest mutation and therefore recurrence risk due to the observed mutation might be overestimated.

Restriction of the model

The program generates for each mosaic person, likelihoods for all mosaic genotypes, next to having a normal allele or being carrier on the other chromosome. The possibility of being mosaic on both chromosomes is left out of the calculations, Since likelihoods for double mosaics are very small this will lead small mistakes and is acceptable.

Testing of the model

Testing of the model is possible in X-linked recessive lethals through:

- mutation selection equilibrium: When a female is mated to a non-affected male to

produce a daughter, which is subsequently mated to a non-affected male, etc, the carrier probability for the daughter after an arbitrarily number of generation must be equal to 4μ. When the last generation offspring is defined as a male, his risk of being affected equals 3μ.

- When in a pedigree one affected boy is observed, genotype probabilities for the new mutation to have happened in each gonadal generation must be the same.
- Recurrence risk for proven germinal mosaics must be equal to the recurrence risks calculated by Van der Meulen et al. (1995), because the same model is used. Only in the program 20 gonadal generation of germline mosaics are defined, where the first generation mutation is defined as 100% mutation, leading to 50% recurrence risk, and the formula has effectively 21 gonadal generations, resulting in slightly different results.
- Comparison of results with results of manual calculations in nuclear families (Van der Meulen et al., submitted) and in extended pedigrees.
- Code inspection

The output

The output of GRONLMOS, exists of the databases mut_1 and mut_2 to normal alleles, mut_2 to mutated allele, followed by likelihoods at the defined recombination fraction and at recombination 0.5 for genotypes for all articulation persons, who are the persons in the pedigree used for peeling upon in the pedigree. The output finally gives bayesian likelihoods for all genotypes off the consultand.

Example

The example is problem 3a from a workshop from the clinical genetics society on germinal mosaicism: risk calculation in Duchenne Muscular Dystrophy (London, march 1994). The consultand is a sister of a woman, who has an affected son. The affected son has inherited the grandpaternal X-chromosome. Figure 1 shows the pedigree and the observed alleles of an intragenic marker.

```
phen_gen(1,-1,1,"male aff",0)
phen_gen(1,-1,2,"male aff",1)
phen_gen(1,-1,1,"male unaff",1)
phen_gen(1,-1,2,"male unaff",0)
phen_gen(1,-1,3,"male unaff",0)
phen_gen(1,1,1,"female unaff low CK",0.95)
phen_gen(1,1,2,"female unaff low CK",0.33)
phen_gen(1,2,2,"female unaff low CK",0)
phen_gen(1,1,1,"female unaff high CK",0.05)
phen_gen(1,1,2,"female unaff high CK",0.67)
phen_gen(1,2,2,"female unaff high CK",0)
phen_gen(1,1,1,"female unaff",1)
phen_gen(1,1,2,"female unaff",1)
phen_gen(1,2,2,"female unaff",0)
phenotype("I-1","male unaff")
phenotype("I-2","female unaff")
phenotype("II-1","male unaff")
phenotype("II-2","female unaff")
phenotype("II-3","female unaff")
phenotype("III-1","male aff")
parents("I-1","I-2","II-2")
parents("I-1","I-2","II-3")
parents("II-1","II-2","III-1")
fenotype("I-1",[0,3],[-1,-1])
fenotype("I-2",[1,1],[1,2])
fenotype("II-1",[0,0],[-1,-1])
fenotype("II-2",[1,2],[0,3])
fenotype("II-3",[1,1],[0,3])
fenotype("III-1",[2,3],[-1,-1])
mut_locus(1)
rec_male([0.5])
rec_female([0.0])
set_for_analysis(["II-2"],["III-1"])
set_for_analysis(["II-3"],["II-2","II-3"])
set_for_analysis(["II-3"],[])
risk_person("II-3")
allele_frequencies(1,[0.999999,0.000001])
allele_frequencies(2,[0.4,0.3,0.3])
selection([1,2])
mut_person(["I-1","II-2"])
mut_freq(0.0000005)
soma_frac(0.3333333)
```

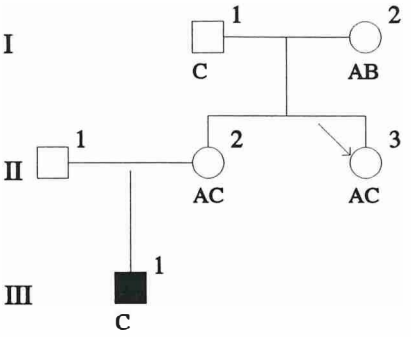


Figure 1: Pedigree

GRONLMOS specific Input

The mutation frequency and allele frequency of the disease allele are set in mutation selection equilibrium ($q=2\mu$). The somatic fraction is set to 1/3. The grandfather (I-1) and the sister of the consultand (II-2) are the mutation persons. Somatic mosaicism is made impossible for the phenotype 'unaffected male' by phenotype genotype relation *phen_gen(1,-1,3,"unaff male",0)*. In the input file phenotype_genotype relations are also defined for females with normal low CK levels and high CK levels. Peeling is performed using the utility program GRONTREE (Te Meerman, 1991).

Output file GRONLMOS.OUT

```
gronlmos (c) 1996 G.J. te Meerman & Martin A. van der Meulen
Medical Genetics University of Groningen
 datafile : vb3a.som Attributes : Date (Day/Month/Year/Hour/Min/Size):
5 7 1996 14 30 1212
allele mut_1 mut_21 mut_22
3  0.0000003333333 0 1
4  0.0000000333333335 0 1
5  0.00000006666667 0.5 0.5
6  0.00000013333334 0.75 0.25
7  0.00000026666668 0.875 0.125
8  0.00000053333336 0.9375 0.0625
9  0.00000106666672 0.96875 0.03125
10  0.00000213333344 0.984375 0.015625
11  0.00000426666688 0.9921875 0.0078125
12  0.00000853333376 0.99609375 0.00390625
13  0.00001706666752 0.998046875 0.001953125
14  0.00003413333504 0.9990234375 0.0009765625
15  0.00006826667008 0.99951171875 0.00048828125
16  0.00013653334016 0.99975585938 0.000244140625
17  0.00027306668032 0.99987792969 0.0001220703125
18  0.00054613336064 0.99993896484 0.00006103515625
19  0.0010922667213 0.99996948242 0.000030517578125
20  0.0021845334426 0.99998474121 0.000015258789063
21  0.0043690668851 0.99999237061 0.0000076293945313
22  0.0087381337702 0.9999961853 0.0000038146972656
23  0.01747626754 0.99999809265 0.0000019073486328
25  0.96504783159 1 0
 Recombination vector [0]
 Recombination vector [0.5]
-[1,2]
Problem: ["II-2"]["III-1"]
[3,2][1,3] II-2 0.0000E+00 1.6667E-07
[4,2][1,3] II-2 0.0000E+00 1.6667E-08
[5,2][1,3] II-2 0.0000E+00 1.6667E-08
[6,2][1,3] II-2 0.0000E+00 1.6667E-08
[7,2][1,3] II-2 0.0000E+00 1.6667E-08
[8,2][1,3] II-2 0.0000E+00 1.6667E-08
[9,2][1,3] II-2 0.0000E+00 1.6667E-08
[10,2][1,3] II-2 0.0000E+00 1.6667E-08
[11,2][1,3] II-2 0.0000E+00 1.6667E-08
[12,2][1,3] II-2 0.0000E+00 1.6667E-08
[13,2][1,3] II-2 0.0000E+00 1.6667E-08
[14,2][1,3] II-2 0.0000E+00 1.6667E-08
[15,2][1,3] II-2 0.0000E+00 1.6667E-08
[16,2][1,3] II-2 0.0000E+00 1.6667E-08
[17,2][1,3] II-2 0.0000E+00 1.6667E-08
[18,2][1,3] II-2 0.0000E+00 1.6667E-08
[19,2][1,3] II-2 0.0000E+00 1.6667E-08
[20,2][1,3] II-2 0.0000E+00 1.6667E-08
[21,2][1,3] II-2 0.0000E+00 1.6667E-08
[22,2][1,3] II-2 0.0000E+00 1.6667E-08
[23,2][1,3] II-2 0.0000E+00 1.6667E-08
[1,2][3,3] II-2 3.3333E-07 1.6667E-07
[1,2][4,3] II-2 3.3333E-08 1.6667E-08
[1,2][5,3] II-2 3.3333E-08 1.6667E-08
[1,2][6,3] II-2 3.3333E-08 1.6667E-08
[1,2][7,3] II-2 3.3333E-08 1.6667E-08
[1,2][8,3] II-2 3.3333E-08 1.6667E-08
[1,2][9,3] II-2 3.3333E-08 1.6667E-08
[1,2][10,3] II-2 3.3333E-08 1.6667E-08
[1,2][11,3] II-2 3.3333E-08 1.6667E-08
[1,2][12,3] II-2 3.3333E-08 1.6667E-08
[1,2][13,3] II-2 3.3333E-08 1.6667E-08
[1,2][14,3] II-2 3.3333E-08 1.6667E-08
[1,2][15,3] II-2 3.3333E-08 1.6667E-08
[1,2][16,3] II-2 3.3333E-08 1.6667E-08
[1,2][17,3] II-2 3.3333E-08 1.6667E-08
[1,2][18,3] II-2 3.3333E-08 1.6667E-08
[1,2][19,3] II-2 3.3333E-08 1.6667E-08
[1,2][20,3] II-2 3.3333E-08 1.6667E-08
[1,2][21,3] II-2 3.3333E-08 1.6667E-08
[1,2][22,3] II-2 3.3333E-08 1.6667E-08
[1,2][23,3] II-2 3.3333E-08 1.6667E-08
[3,2][2,3] II-2 3.3333E-07 3.3333E-07
[4,2][2,3] II-2 3.3333E-08 3.3333E-08
[5,2][2,3] II-2 6.6667E-08 5.0000E-08
[6,2][2,3] II-2 1.3333E-07 8.3333E-08
[7,2][2,3] II-2 2.6667E-07 1.5000E-07
[8,2][2,3] II-2 5.3333E-07 2.8333E-07
[9,2][2,3] II-2 1.0667E-06 5.5000E-07
[10,2][2,3] II-2 2.1333E-06 1.0833E-06
[11,2][2,3] II-2 4.2667E-06 2.1500E-06
[12,2][2,3] II-2 8.5333E-06 4.2833E-06
[13,2][2,3] II-2 1.7067E-05 8.5500E-06
[14,2][2,3] II-2 3.4133E-05 1.7083E-05
[15,2][2,3] II-2 6.8267E-05 3.4150E-05
[16,2][2,3] II-2 1.3653E-04 6.8283E-05
[17,2][2,3] II-2 2.7307E-04 1.3655E-04
[18,2][2,3] II-2 5.4613E-04 2.7308E-04
[19,2][2,3] II-2 1.0923E-03 5.4615E-04
[20,2][2,3] II-2 2.1845E-03 1.0923E-03
[21,2][2,3] II-2 4.3691E-03 2.1845E-03
[22,2][2,3] II-2 8.7381E-03 4.3691E-03
[23,2][2,3] II-2 1.7476E-02 8.7381E-03
[25,2][2,3] II-2 9.6505E-01 4.8252E-01
[1,2][2,3] II-2 1.0000E+00 5.0000E-01
[3,3][2,2] II-2 3.3333E-07 3.3333E-07
[4,3][2,2] II-2 3.3333E-08 3.3333E-08
[5,3][2,2] II-2 3.3333E-08 5.0000E-08
[6,3][2,2] II-2 3.3333E-08 8.3333E-08
[7,3][2,2] II-2 3.3333E-08 1.5000E-07
[8,3][2,2] II-2 3.3333E-08 2.8333E-07
[9,3][2,2] II-2 3.3333E-08 5.5000E-07
[10,3][2,2] II-2 3.3333E-08 1.0833E-06
[11,3][2,2] II-2 3.3333E-08 2.1500E-06
[12,3][2,2] II-2 3.3333E-08 4.2833E-06
[13,3][2,2] II-2 3.3333E-08 8.5500E-06
[14,3][2,2] II-2 3.3333E-08 1.7083E-05
[15,3][2,2] II-2 3.3333E-08 3.4150E-05
[16,3][2,2] II-2 3.3333E-08 6.8283E-05
```

```
[17,3][2,2] II-2 3.3333E-08 1.3655E-04
[18,3][2,2] II-2 3.3333E-08 2.7308E-04
[19,3][2,2] II-2 3.3333E-08 5.4615E-04
[20,3][2,2] II-2 3.3333E-08 1.0923E-03
[21,3][2,2] II-2 3.3333E-08 2.1845E-03
[22,3][2,2] II-2 3.3333E-08 4.3691E-03
[23,3][2,2] II-2 3.3333E-08 8.7381E-03
[25,3][2,2] II-2 0.0000E+00 4.8252E-01
[1,3][2,2] II-2 0.0000E+00 5.0000E-01
Problem: ["II-3"]["II-2","II-3"]
[1,1][1,3] II-3 3.2000E-06 2.6000E-06
[1,1][2,3] II-3 1.3333E-07 6.6667E-08

Problem: ["II-3"][]
 risk person II-3 [1,1][2,3] genotype probability 0.040000155053
 risk person II-3 [1,1][1,3] genotype probability 0.95999984495
 result from added likelihoods of articulation person
Elapsed time 6.76
 End gronlmos version date: 25 mar 1996
```

Recurrence risk in example

The consultand has in the example a carrier probability of 4%. However, if the somatic fraction of the grandfather (unaff male) is not reduced, the carrier probability equals 20%. This example illustrates the sensitivity of the results of the calculation to this parameter. Note that results are not exactly the same as in the manual calculations of the same problem (Van der Meulen et al., extended pedigrees), because the program has 20 gonadal generations and the formula 21 (Van der Meulen et al., 1995). The manual calculation give the same results when in stead of 0.048, 0.05 is used as recurrence risk due to mosaicism after one affected child.

Discussion

Incorporation of information on mutation screening has to be done with care: for instance when two affecteds/carriers from one pedigree are screened for mutations, the results are not independent.

The bayesian risks calculated for the risk_person, will not show in what percentage of the cases, the disease locus is inherited through the mosaic interpretation. By removal of the list ***mut_person*** from the input a check can be performed on the possibility of a non mosaic interpretation. Comparison of the mosaic likelihood with the non mosaic likelihood is a measure for the plausibility of the mosaic interpretation. Recombinants (single or double) will cause the overall likelihood of the non mosaic interpretation to go

down. In case of intragenic markers, with recombination 0 to the disease allele, the likelihood of the non-mosaic interpretation will even be 0. As mentioned by Bakker et al. (1989), analyses of mosaic families with non mosaic models will give erroneous scoring of recombinants.

In the output file can be seen that for person II-2 the likelihoods for all mosaic genotypes, which transmit the new mutation to the affected son are equal. This is expected due to the incorporated model, in which an equal probability of mutation for each mitosis is assumed.

Although computer programs make recurrence risk calculation possible in large and complicated pedigrees, the input of the program is still complicated and use is not possible without expertise. Advantage of a computer program is, that the size of the pedigree is irrelevant for the complexity of the input. It might be helpful to make input files first for small pedigrees.

Computer models are important in recurrence risk calculation, because they are the best you have. Recurrence risk calculations involving mosaicism are complicated and the effect of additional information on the recurrence risk are often difficult to estimate. A big advantage of the computer program in comparison to manual calculations is that it is easy to evaluate the effect of variation of the different model parameters on the recurrence risk.

Computer programs described are available to the public domain without any restriction on request from one of the authors.

Acknowledgements

We thank Dr. Lodewijk A. Sandkuijl for fruitful discussion and critical comments on the manuscript. This work is supported by the Netherlands Organisation for Scientific Research (NWO).

References

Bakker E, Veenema H, Den Dunnen JT, Van Broeckhoven C, Grootscholten PM, Bonten EJ, Van Ommen GJB and Pearson PL. Germinal mosaicism increases the recurrence risk for 'new' Duchenne muscular dystrophy mutations. J Med Genet 1989; 26:553-559

Clayton JF. A computer programme to calculate risk in X-linked disorders using multiple marker loci. J Med Genet 1986; 23:35-39

Edwards JH. Familiarity, recessivity and germline mosaicism. Ann Hum Genet 1989; 53:33-47

Hartl DL. Recurrence risk for germinal mosaics. Am J Hum Genet 1971; 23:124-134

Jeanpierre M. Germinal mosaicism and risk calculation in X-linked diseases. Am J Hum Genet 1992; 50:960-967

Sarfarazi M and Williams H. A computer programme for the estimation of genetic risk in X-linked disorders, combining pedigree and DNA probe data with other conditional information. J Med Genet 1986; 23:40-45

Te Meerman GJ. A logic programming approach to pedigree analysis. Thesis publishers Amsterdam, 1991.

Thompson EA. Pedigree Analysis in Human Genetics. 1986 The Johns Hopkins University Press Ltd, London

Van der Meulen MA, Van der Meulen MJP and Te Meerman GJ. Recurrence risk for germinal mosaics revisited. J Med Genet 1995; 32:102-104

Van der Meulen MA, Te Meerman GJ and Sandkuijl LA. Calculation of Recurrence Risk in case of possible mosaicism, submitted.

Winter RM. The calculation of genetic risk in X-linked recessive conditions using programmable calculators. Clin Genet 1980; 17:171-175

II-5 General discussion and Summary

Edwards (1986) says that in practice, with a disorder as severe as Duchenne, there is little consumer tolerance of uncertainty and its exact computation is of limited clinical value. Although this is certainly correct, it is also useful to have a complete understanding of the models used for recurrence risk calculation. A general framework for recurrence risk calculation in case of possible mosaicism is presented. This framework is used in manual recurrence risk calculations and for implementation in a computer program. The mutation selection equilibrium determines the ratio between transmitted and new mutations. In case of a new mutation, distinction is made between somatic mosaics, in whom the mutation had occurred before the germline stem cell is formed leading to 50% mutation carrying oocytes or spermatocytes, and germline mosaics, in whom the mutation occurred in one of the cell generations in the germline leading to a variable percentage of carrier gonadal cells.

The tables presented by Van der Meulen et al. (1995, chapter II-2), which resulted out of an extension of the formula originally presented by Hartl (1971), are useful, because they make incorporation of the family structure in the manual recurrence risk calculation in case of possible germline mosaicism possible. We show that the estimated recurrence risk changes considerably when obtainable phenotypic information is collected and used (Van der Meulen et al., submitted, chapter II-3.1). Although errors are likely, manual calculations are even possible on multiple generation pedigrees (chapter II-3.2). Recurrence risk calculation on extended pedigrees can, however, be performed more reliably using a computer program as GRONLMOS (chapter II-4). The computer program can be used for checking of manual calculations, but can deal with problems of virtually any structure and size and with recombination. An advantage of the computer program in comparison to manual calculations is, that input for the computer program for complex pedigrees is not more complex than for simple pedigrees, while the required calculations are.

When mutation search improves, recurrence risk calculation will be less required, but the recurrence risk due to new mutations and therefore germline mosaicism will remain

difficult, especially when no DNA is available from affected individual(s). Prenatal mutation screening would be a possibility in that case, but care has to be taken in the recurrence risk calculation, because when no mutation is found, observations on multiple pedigree members are not independent.

For better understanding of disease etiology, large studies are important to evaluate the genetic models used for recurrence risk calculation. Stec et al. (1995) present a study of 415 families. From this study follows that a proven mosaic female (deletion found in offspring and not in mother) had an empirical recurrence risk after two affected children of 0.34 +/- 0.12 when the at risk chromosome was inherited (personal communication Bertram Muller). Note that in this study, similar to the study of Van Essen et al (1992), mosaics are defined as non-carriers, when the mutation found in the affected child(ren) can not be found in blood of the parent. However in our model early somatic mutations might lead to detectable new mutations in parents. This even though the mutation is de novo. It would be interesting to search for the non-mutated variant in blood or other tissues in which a mutation was detected, in the study described by Van Essen et al. (1992) seen as carriers, to prove that somatic mosaics with a detectable new mutation really exist. Clinical distinction between somatic mosaics and carriers has direct impact on the carrier risk for her sibs, but in genetic counselling direct testing of sibs for the detectable mutation is a more logical approach.

Many studies have focused on the estimation of male and female mutation rates in Duchenne Muscular Dystrophy (DMD). Muller et al. (1992) conclude that the ratio between the two is very close to 1, thus giving evidence for equal mutation rates in males and females in DMD. This is consistent with the equal development in early embryology in males and females, which is the phase giving rise to the highest proportion of cells carrying the mutation, because only mutations in early development will lead to a significant proportion of mutant cells. Karel et al. (1986) describe that it requires large sample sizes to detect even large differences between male and female mutation rates. Ascertainment bias is shown to have a great effect on the outcome of the segregation analysis. Grimm et al. (1994) argue that, although mutation frequencies appear to be about equal in males and females, pointmutations stem mainly from spermatogenesis,

while most deletions arise from oogenesis. Passos Bueno et al. report of differences in mosaicism frequencies for proximal and distal mutations in DMD (1992). This result could not be confirmed by Grimm et al. (1994). Note that incorporation of knowledge about male versus female mutation origin or differences in mutation frequencies can easily be build in in recurrence risk calculations.

Distinction between somatic mosaics and carriers of a mutation is hard in DMD. When a new mutation happened early enough to make the new mutation detectable, CK values can be expected to behave comparable to CK levels in carriers. On the other hand when the mutation occurred one of the last mitoses before the germline stemcell is formed, the mutation might have no influence on the CK value. A correlation can be expected between mutation screening and CK values in somatic mosaics.

The validity of assumptions and robustness of models used in the recurrence risk calculation can easily be questioned. For genetic counselling purposes results have to be used carefully. Changes in assumptions used in the calculations can give insight in the sensitivity of the calculated risks to the parameter varied in the model.

In practice, this is a field in which molecular biology is advancing faster than methods of statistical analysis, which will become irrelevant as soon as direct diagnostic procedures are developed and perfected (Edwards, 1986). In cases where direct diagnostics are not conclusive, which is 10 years later still the case, recurrence risk calculations will still be necessary.

References

Edwards JH. The population genetics of Duchenne: natural and artificial selection in Duchenne Muscular Dystrophy. J Med genet 1986; 23:521-530

Grimm T, Muller B, Muller CR and Janka M. Theoretical considerations on germline mosaicism in Duchenne muscular dystrophy. J Med Genet 1990; 27:683-687

Karel ER, Te Meerman GJ, Ten Kate LP. On the power to detect differences between male and female mutation rates for Duchenne muscular dystrophy, using classical segregation analysis and restriction fragment length polymorphisms. Am J Hum Genet 1986; 38:827-840

Muller B, Dechant C, Meng G, Liechti-Galatti S, Doherty RA, Hejtmanchik JF, Bakker E et al. Estimation of the male and female mutation rates in Duchenne muscular dystrophy (DMD) Hum Genet 1992; 89:204-206

Passos-Bueno MR, Bakker E, Kneppers ALJ, Takata RI, Rapaport D, Dunnen JT, Zatz M and Ommen GJB. Different Mosaicism Frequencies for Proximal and Distal Duchenne Muscular Dystrophy (DMD) Mutations Indicate Difference in Etiology and Recurrence Risk. Am J Hum Genet 1992; 51:1150-1155

Stec I, Kress W, Meng G, Muller B, Muller CR, Grimm T. Estimate of severe autosomal recessive limb-girdle muscular dystrophy (LGMD2C, LGMD2D) among sporadic muscular dystrophy males: a study of 415 families. J Med Genet 1995; 32:930-933

Van der Meulen MA, Van der Meulen MJP and Te Meerman GJ. Recurrence risk for germinal mosaics revisited. J Med Genet 1995; 32:102-104

Van der Meulen MA, Te Meerman GJ and Sandkuijl LA. Calculation of Recurrence Risk in case of possible mosaicism, submitted.

Van Essen AJ, Abbs S, Baiget M, Bakker E, Boileau C, Van Broeckhoven C et al. Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: A European study. Hum Genet 1992; 88:249-257

Samenvatting voor leken

Algemeen

Een mens bestaat uit cellen. In de kern van de cellen zit het genetische materiaal (DNA), waarop alle erfelijke eigenschappen beschreven zijn. Een gen is een deel van het genetisch materiaal, dat een bepaalde functie heeft. Alle cellen in een mens bevatten hetzelfde genetische materiaal, maar door het activeren van bepaalde genen kunnen ze verschillende functies in het lichaam uitvoeren. Wanneer een verandering (=mutatie) optreedt in het genetisch materiaal kan het zijn dat een gen zijn functie niet meer (goed) kan uitvoeren.

Het genetische materiaal is verdeeld over chromosomen, waarvan mensen 23 paar bezitten. Van ieder paar is één chromosoom geërfd van de vader en één van de moeder. Eén paar bevat de chromosomen die het geslacht bepalen (=geslachtschromosomen). Vrouwen hebben van deze geslachtschromosomen twee dezelfde, die noemen we X, maar mannen hebben twee verschillende geslachtschromosomen, een X en een Y chromosoom.

Celdeling

De cellen waaruit het menselijk lichaam bestaat zijn ontstaan door delingen vanuit de bevruchte eicel. Voor iedere vermeerderingsdeling (=mitose) wordt het genetische materiaal gekopieerd, waarna door deling twee identieke cellen ontstaan met in beide cellen weer een complete set chromosomen.

Voor het vormen van de zaadcellen bij de man en de eicellen bij de vrouw (=voortplantingscellen) vindt de zogenaamde reductiedeling (=meiose) plaats. Bij deze deling ontstaan cellen met slechts één chromosoom van ieder paar. Tijdens de meiose gaan de van vader en moeder verkregen identieke chromosomen naast elkaar liggen. Er ontstaan dan regelmatig uitwisselingen tussen beide chromosomen. Wanneer er uitwisseling heeft plaatsgevonden is een zogenaamde 'recombinant' ontstaan. Een recombinant zorgt ervoor, dat iemand aan zijn kind niet een chromosoom van alléén de grootvader of van alléén de grootmoeder doorgeeft, maar een chromosoom dat bestaat uit een deel van een chromosoom van grootmoeder en het andere gedeelte van een chromosoom van grootvader. Tijdens iedere meiose komen gemiddeld 53 recombinaties voor, verspreid over alle chromosomen. Twee dicht naast elkaar gelegen genen op één

chromosoom zullen dus meestal gezamenlijk overerven (=gekoppelde kenmerken), behoudens wanneer een recombinatie precies er tussenin plaatsvindt. Doordat de chromosomen van vader en moeder onafhankelijk worden verdeeld over de voortplantingscellen, bestaat het genetisch materiaal van een voortplantingscel uit een mengeling van het genetisch materiaal van moeder en van vader. Genen op verschillende chromosomen zullen dus onafhankelijk van elkaar overerven (=niet gekoppelde kenmerken). Doordat mannen van het geslachtschromosoom een X en een Y chromosoom bezitten, zijn er na meiose X dragende en Y dragende zaadcellen. Het Y chromosoom hebben mannen dus van hun vader ontvangen.
Bij de bevruchting van een eicel door een zaadcel wordt het genetisch materiaal van beide cellen samengevoegd, zodat weer een cel ontstaat met twee exemplaren van ieder chromosoom.

Het vinden van de plaats van een eigenschap op het chromosoom

Bij een erfelijke ziekte zullen in sommige gevallen in een familie alle aangedane personen een bepaald stukje chromosoom identiek geërfd hebben, terwijl in de niet aangedane personen dit chromosomale stukje niet identiek is. De zoektocht naar de erfelijke basis, die leidt tot een bepaalde ziekte of eigenschap, is gericht op het vinden van deze chromosomale lokatie van een (ziekte)gen. Een ziektegen (=ziekte veroorzakend gen) is een mutatie/verandering in het gen dat normaliter op deze plaats aanwezig is, maar door de mutatie zijn functie niet meer goed uit kan voeren. De meest efficiënte methode om te zoeken naar de lokatie van een gen is het verzamelen van bloed, met daarin cellen met het genetisch materiaal, in families met een groot aantal aangedane personen. Gezocht wordt dan naar een identiek chromosomaal stuk(je), dat in de cellen van alle aangedane personen in de stamboom aanwezig is, of in andere woorden: samen met de ziekte overerft. Doordat recombinaties hebben plaatsgevonden kan de lokatie van een ziektegen nauwkeuriger bepaald worden op een chromosoom. Immers, wordt de persoon met de recombinant ziek, dan ligt het ziektegen op dat deel van het chromosoom dat door deze zieke persoon geërfd is. In een familie wordt de kans berekend, dat een bepaalde lokatie gekoppeld is aan een ziekte. Wanneer deze kans groter is dan bijvoorbeeld 1000 op 1, dan is het ziektegen hoogst waarschijnlijk gelokaliseerd. Vaak moeten, om zeker genoeg te zijn van de lokatie, verschillende families geanalyseerd worden.

Complex overervende ziektes

Veel genen, betrokken bij de ziektes waarbij het zo simpel is als boven beschreven, zijn inmiddels gelokaliseerd, zodat nu de ingewikkelder (=complex) overervende ziektes aan de beurt zijn. Een aantal van de moeilijkheden zijn: dat sommige ziekten veroorzaakt kunnen worden door verschillende genen op zelfs verschillende chromosomen (=heterogeniteit), dat mensen ziek zijn vanwege een niet erfelijke variant van de ziekte (=fenokopie), dat niet alle mensen die een mutatie bij zich dragen daadwerkelijk ziek worden (=onvolledige of verminderde penetrantie), dat meerdere genen betrokken zijn in één ziekte, dat omgevings faktoren invloed hebben op het wel of niet tot expressie komen van de ziekte, enz, enz.

Bij heterogeniteit en fenokopieën wordt het lokaliseren van genen bemoeilijkt, doordat de verschillende lokaties in verschillende families ervoor zorgen, dat het bewijs voor lokatie 1 in familie A teniet wordt gedaan door familie B, waarin de ziekte wordt bepaald door lokatie 2 of doordat de ziekte niet genetisch is. In familie B zal immers voor lokatie 1 geen overeenstemming bestaan tussen de aangedane familieleden, behalve door toeval. Het kan zelfs voorkomen dat heterogeniteit voorkomt binnen families. Heterogeniteit wordt veroorzaakt doordat verschillende genen betrokken kunnen zijn bij een celfunctie, waarbij in één van de betrokken genen een mutatie aanwezig is. Bij verminderde penetrantie kan het moeilijk zijn om families te vinden met meerdere aangedane personen. Meerdere aangedane personen zijn nodig, omdat pas door vergelijking van het erfelijke materiaal van de twee of meer aangedane personen het gemeenschappelijke stukje chromosoom waar het ziektegen ligt kan worden gevonden.

IBD mapping

In dit proefschrift wordt een methode beschreven, hoe de genlokatie van een ziekte met een lage penetrantie of een complexe overerving kan worden gevonden, ondanks het feit dat er weinig of geen familiaire gevallen bekend zijn. Bij deze methode zoeken we naar gelijke stukken van chromosomen tussen aangedane personen uit een bepaald geografisch gebied. Er bestaat dan een kans, dat deze aangedane personen de aanleg voor deze ziekte geërfd hebben van een gemeenschappelijke voorouder, alhoewel zij meestal niet weten dat zij deze gemeenschappelijke voorouder hebben. Bij het zoeken naar dit ziektegen maken we er gebruik van dat deze personen niet alleen het ziektegen identiek hebben door

afstamming (=identity by descent=IBD), maar ook een om het ziektegen gelegen deel van het chromosoom. Achter de veronderstelling dat de aangedane personen hetzelfde ziektegen zullen delen ligt een theoretische gedachte. Bij het doorgeven van het genetische materiaal aan het nageslacht treedt een toevallig proces op (=drift), waardoor bepaalde kopieën van een chromosoom verdwijnen en anderen veelvuldig gaan voorkomen. Een proces dat vergelijkbaar is met genetische drift is de verdeling van achternamen in Nederland. In een bepaald dorp kan een naam zeer veelvuldig voorkomen, terwijl deze naam in een ander dorp, dat niet eens ver daar vandaan hoeft te liggen, vrij weinig voorkomt. Het verdwijnen van bepaalde kopieën van chromosomen komt bijvoorbeeld doordat er personen zijn die geen nageslacht krijgen of doordat aan alle kinderen het andere chromosoom wordt doorgegeven. De veronderstelling van een of zelfs meerdere gezamenlijke voorouders voor mensen uit een bepaald gebied is niet zo onwaarschijnlijk, als bedacht wordt dat het aantal voorouders die een persoon heeft n generaties terug 2^n bedraagt, ofwel iedereen heeft $2^1=2$ ouders, $2^2=4$ grootouders, $2^3=8$ overgrootouders, enz. In een studie in Frans Canada bleek zelfs dat een grote groep patiënten zoveel gezamenlijke voorouders hadden, dat de inbrenger van het ziekte veroorzakende gen niet aangewezen kon worden. Zelfs als genealogisch wel een relatie tussen personen kan worden aangetoond is het niet zeker of dit wel de verbinding is waarlangs het ziektegen is gekomen. De IBD mapping methode is alleen geschikt voor ziektes met een lage mutatiefrequentie, omdat anders de aangedane personen niet (bijna) allemaal dezelfde mutatie bij zich dragen.

Hoge mutatie frequentie

Bij ziektes waarvan de lokatie op een chromosoom bekend is, maar de mutatiefrequentie hoog, is de kans groot dat de ziekte in de aangedane persoon veroorzaakt wordt door een nieuwe mutatie. Een voorbeeld is de Spierdystrofie van Duchenne; een ziekte waaraan patiënten als jong volwassene overlijden, gelegen op het X chromosoom, met een recessieve overerving. Een recessieve ziekte is een aandoening die niet tot uiting komt zolang één niet gemuteerd gen aanwezig is. Mannen hebben slechts één X-chromosoom en zullen dus ziek zijn wanneer zij een mutatie hebben in het ziektegen. Wanneer de ziekte op jonge leeftijd dodelijk is, zullen aangedane mannen zich niet voortplanten, waardoor deze gemuteerde genen uit de populatie verdwijnen. Vrouwen kunnen echter

draagster zijn van de mutatie, zonder dat zij daar last van hebben, omdat zij nog een ander (niet gemuteerd) gen hebben op het andere X-chromosoom. Voor deze ziekte kan de verhouding tussen nieuwe en geërfde mutaties worden berekend vanuit de populatie genetica: mutatie selectie evenwicht. Dat betekent dat het totaal aantal mutaties niet toeneemt, omdat er evenveel mutaties door selektie uit de bevolking verdwijnen (=overlijden vóór voortplanting) als er nieuwe bijkomen. In mutatie selectie evenwicht hebben vrouwen met 1 aangedaan kind met de Spierdystrofie van Duchenne een kans van 2/3 om draagster te zijn van deze aandoening op een X-chromosoom en 1/3 kans dat het een nieuwe mutatie betreft.

Nieuwe mutatie

Bij een nieuwe mutatie is de vraag wanneer deze mutatie heeft plaats gevonden. Het menselijk lichaam bestaat uit vele miljarden cellen, die allen afkomstig zijn uit celdelingen vanuit één enkele bevruchte eicel. Wanneer een kopieërfout (=mutatie) ontstaat in een celdeling, zullen alle cellen, die vanuit deze gemuteerde cel ontstaan, deze mutatie bevatten. Na ongeveer 10 celdelingen vanaf de bevruchte eicel ontstaat de germline stamcel, waaruit alle voortplantingscellen worden gevormd. Is een nieuwe mutatie, die niet in de eicel zat, al aanwezig in de germline stamcel, dan is de mutatie dus automatisch aanwezig in alle cellen in de germline en spreken we van een somatisch mozaïek. In onze definitie hebben somatisch mozaïeken een aangedane germline stamcel, omdat ze anders niet de mutatie door kunnen geven aan hun nageslacht. De meiose zorgt ervoor dat de kans voor somatisch mozaïeken op het doorgeven van de mutatie 50% is. Als de mutatie na de vorming van de germline stamcel plaatsvindt levert dit, afhankelijk van wanneer de mutatie plaats vindt, een bepaald deel van de voortplantingscellen op die de mutatie hebben (=germline mozaïek). In het gebruikte model, is het aantal celdelingen of celgeneraties na vorming van de germline stamcel voor het vormen van alle voortplantingscellen op 20 gesteld.

Herhalingsrisico berekeningen

In herhalingsrisico berekeningen bij ziektes met een hoge mutatie frequentie, moet dus rekening gehouden worden met alle mogelijke herkomsten van de geobserveerde mutatie. Wanneer in een X chromosomale aandoening een aangedane zoon wordt geboren, kan de

moeder draagster, een somatisch mozaïek, of een germline mozaïek zijn. Deze drie mogelijkheden hebben verschillende herhalingsrisico's. In de herhalingsrisico berekening kunnen ook waarnemingen meegenomen worden over de doorgegeven chromosomen van ouders naar kinderen, maar ook resultaten van onderzoek naar dragerschap van een mutatie en andere kenmerken, waarvan bekend is dat ze gekoppeld zijn aan de ziekte. Een voorbeeld van een kenmerk gekoppeld aan de Spierdystofie van Duchenne is het CK gehalte in het bloed, dat verhoogd is bij ongeveer 70 % van de draagsters van een mutatie.

Wanneer de stamboom uitgebreider wordt, wordt de berekening ingewikkelder, doordat verschillende herkomsten van de mutatie bekeken moeten worden. Het beschreven computer programma kan herhalingsrisico's berekenen voor uitgebreide stambomen, die handmatig niet meer uit te voeren zijn. Belangrijker echter is, dat het computerprogramma gebruikt kan worden als controle en voor het onderzoeken van het effect op het herhalingsrisico, wanneer het door de computer gebruikte erfelijkheids-model gevarieerd wordt. Tevens kan het computerprogramma recombinatie in de herhalingsrisico berekening betrekken.

Curriculum vitae

Ik, Martin Allert van der Meulen, ben geboren 6 december 1965 en getogen te Eindhoven. Vanaf de 5e klas Atheneum B tot en met het eindexamen voortgezet in Oss. In 1984 begonnen aan de studie zoötechniek aan de Landbouw Universiteit Wageningen, richting veefokkerij. Mijn eerste kennismaking met het doen van wetenschappelijk onderzoek op een universiteit was tijdens mijn stage aan de University of Guelph, Ontario, Canada, alwaar ik bezig ben geweest met de statistische verwerking van onderzoeksgegevens van een proef naar de vermeerdering van melkgift bij koeien door het verstrekken van een hormoon. In Wageningen terug gekomen meegewerkt in een onderzoek naar de genetische component in de immuunrespons bij kippen. Na het afstuderen in juni 1990 begonnen als wetenschappelijk medewerker aan het Proefstation voor de rundveehouderij, schapenhouderij en paardenhouderij te Lelystad. Na 2 en een half jaar een nieuwe uitdaging gezocht in een door de Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO) gefinancierd promotie onderzoek, aan de vakgroep Medische Genetica van de Rijksuniversiteit Groningen. Het proefschrift, als afronding van dit promotie onderzoek, ligt nu voor U.

List of publications

M.-H. Pinard, **M.A. van der Meulen**, M.B. Kreukniet, M.G.B. Nieuwland and A.J. van der Zijpp: Divergent selection for antibody production in chickens: differences in Major Histocompatibility Complex (MHC) haplotype distribution. Proceedings of the 4th World Congres for Genetics Applied to Livestock Production, Edinburgh, 1990

J.P. Gibson, **M. van der Meulen**, B.W. McBride and J.H. Burton: The effects of genetic and phenotypic production potential on response to recombinant Bovine SomatoTropin. Journal of Dairy Science 1992; 75:878-884

Gerard J te Meerman, Erik Mullaart, **Martin A van der Meulen**, Johanna HG den Daas, Bruno Morolli, Andre G Uitterlinden en Jan Vijg: Linkage Analysis by two dimensional DNA typing. Am J Hum Genet 1993; 53:1289-1297.

Martin A. van der Meulen, Meine J.P. van der Meulen and Gerard J. te Meerman: Recurrence risk for germinal mosaicism revisited. J Med Genet 1995; 32:102-104

Corien C Verschuuren-Bemelmans, Ewout RP Brunt, Margaret Burton, Rob GJ Mensink, **Martin A van der Meulen**, Nico H Smit, Irene Stolte Dijkstra, Charles HCM Buys, Hans Scheffer: Refinement by linkage analysis in two large families of the candidate region of the third locus (SCA3) for autosomal dominant cerebellar ataxia type 1. Hum Genet 1995; 96:691-694

G.J. te Meerman, **M.A. van der Meulen** and L.A. Sandkuijl: Perspectives of identity by descent (IBD) mapping in fouder populations, Clin Exp Allergy 1995; 25(suppl 2):97-102.

Hendrik G. de Vries, **Martin A. van der Meulen**, Rima Rozen, Dickie J.J. Halley, Hans Scheffer, Leo P. ten Kate, Charles H.C.M. Buys and Gerard J. te Meerman: Haplotype identity between individuals who share a CFTR mutation allele Identical by descent: demonstration of the usefulness of the haplotype sharing concept for gene mapping in real populations, Human Genetics 1996; 98:304-309

Martin A. van der Meulen and Gerard J. te Meerman: Association and haplotype sharing due to Identity by Descent, with an application to genetic mapping. In: Genetic mapping of disease genes, edited by JH Edwards, IH Pawlowitzki and E Thompson, Academic Press, to appear april 1997.

Martin A Van der Meulen, Gerard J Te Meerman and Lodewijk A Sandkuijl: Calculation of Recurrence Risk in case of possible mosaicism, submitted.

A Dorum, P Moller, EJ Kamsteeg, H Scheffer, M Burton, KR Heimdal, LO Mehle, E Hovig, CG Trop, AH Van der Hout, **MA Van der Meulen**, CHCM Buys and GJ Te Meerman: Haplotype analysis, a strategy for identifying prevalent mutations, demonstrating a Norwegian BRCA1 founder mutation, submitted.