European Journal of Human Genetics (2008) 16, 1225–1234

www.nature.com/ejhg

ARTICLE

The effect of pedigree structure on detection of deletions and other null alleles

Anna M Johansson*[,1,2] and Torbjörn Säll[1]

[1]*Department of Cell and Organism Biology, Lund University, Lund, Sweden;* [2]*Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden*

Deletions and other null alleles for genetic markers can be detected as a special case of non-Mendelian inheritance, ie when a parent and a child appear to be homozygous for different alleles. The probability to detect a deletion for a fixed overall number of investigated individuals was calculated for biallelic and multiallelic markers with varying allele frequencies. To determine the effect of increasing the number of parents and grandparents, the probability for this event was derived for a parent and one child, a trio, a trio with one grandparent and a trio with two grandparents. The results for biallelic markers show that for a fixed total number of individuals, a sample of trios with two grandparents is always more efficient than the other family types, despite a lower total number of founder chromosomes in the sample. For multiallelic markers the outcome varies. The effect of adding additional children to a nuclear family was also investigated. For nuclear families, the optimal number of children is two or three, depending on the allele frequencies. It is shown that adding children is more efficient than adding grandparents.
European Journal of Human Genetics (2008) **16,** 1225–1234; doi:10.1038/ejhg.2008.75; published online 16 April 2008

Keywords: deletion; null allele; pedigree; genetic marker

Introduction

Deletions are one of the many types of mutations that can affect a genome. The observed size of a deletion range from a single base pair up to an entire arm of a chromosome (see Lewis[1]). In recent years, there has been an increasing interest in investigating structural variants, including deletions.[2] One reason is because deletions may be the cause of some diseases. One such case is the occurrence of somatic deletions involved in cancer, which can be detected through 'loss of heterozygosity' in the tumour when compared to other tissues.[3] Another situation where deletions may be observed is when they occur as *de novo* deletions. These are notable when they subsequently cause disease in the offspring of the person within whom the deletion occurred in the germline.[4] Finally, deletions causing disease may be inherited. Such deletions may act as directly causing in one end of the spectrum, but may also act as risk alleles in complex diseases.[5,6]

However, just as in the case with other types of mutations deletions may have no phenotypic effect.[7,8] Alternatively, the effect is so weak that it is effectively undetectable. Such mutations may then appear as polymorphisms within populations. The presence of deletion polymorphisms in the human genome have recently been investigated in a number of large-scale projects.[5,9–12] These investigations demonstrate that a large number of deletion polymorphisms occur in humans. They also show that, although with few exceptions the deletion constitutes the rare allele, the frequency of the deletion may be relatively high. The presence and frequency of deletion polymorphisms is of interest for a number of reasons. In certain eukaryotic lineages such as in unicellular fungi, genome size is clearly under selective pressure.[13] A small genome size is assumed to have been of adaptive advantage due to

*Correspondence: Dr AM Johansson, Department of Cell and Organism Biology, Lund University, Sölvegatan 29, Lund SE-22362, Sweden.
Tel: +46 46 2227857; Fax: +46 46 147874;
E-mail: Anna.Johansson@cob.lu.se
Received 16 August 2007; revised 22 February 2008; accepted 11 March 2008; published online 16 April 2008

its ability to replicate faster. Such evolution can only occur in the presence of deletions as the basic mutational events. Thus, the genomic positions and allele distribution of deletions are of evolutionary importance.

In order to study deletions, it is necessary to accurately detect them. Small deletions of only a few bases can be detected by sequencing. However, deletions that cover the entire amplified product to be sequenced will not be detected since an individual that is heterozygous for the deletion will simply appear homozygous for the sequence. Such deletions could be detected using methods that measure the amount of DNA in specific chromosomal regions, like PCR-based, or hybridization-based methods (eg MLPA[14] and CGH arrays[15]). Another approach is to apply genetic methods that infer the presence of deletions from the pattern of marker segregation in families. This approach has been employed to detect both *de novo* and inherited deletions. For example, marker segregation was used to detect *de novo* deletions in the RB1 locus in patients with sporadic, bilateral retinoblastoma.[16] Inherited deletions were detected in autism kindreds using microsatellite markers[17] and in protein S deficiency using both microsatellites and SNPs.[18] The fact that analysis of segregation is one effective method to detect deletions has spurred theoretical investigations into the efficiency and power of these methods.[19,20] Amos *et al*[21] has modelled the probabilities of different configurations of a trio including *de novo* deletions, genotyping errors and also inherited deletions. In an earlier paper,[22] we described how deletions causing a dominant disease can be detected. Two recent papers present methods to infer copy number polymorphisms using quantitative measures of alleles and family information from nuclear families[23] and trios.[24] Kohler and Cutler[25] present a method to detect deletions using SNP data on trios. Their approach is to estimate the frequency of a deletion using Mendelian inconsistencies, departures from Hardy–Weinberg proportions and unusual patterns of missing data.

Here, we investigate a method to detect deletions without phenotypic effect by examining the segregation pattern of genetic markers. This approach was recently used in two large-scale projects.[5,10] The method is based on the fact that deletions act as null alleles with respect to marker loci in the deleted region. Null alleles with other causes than deletions are also of interest since undetected null alleles may cause errors in haplotype inference, parentage and population genetic analyses. The method is also relevant for searching for deletions involved in complex disease. Although a phenotypic effect is involved in these cases, the association between genotypes and phenotypes are usually sufficiently weak so that the methods of the present paper are more relevant than the methods presented in Johansson *et al*[22] Thus, in the case of a complex disease, we suggest the following: first, search for deletion independently of phenotype and then, if a deletion is found, investigate the co-segregation between the disease and the deletion.

We have derived the probability to detect the existence of a deletion as a function of family structure and allele frequencies both of the marker alleles and the deletion. These probabilities were used to compare the efficiency of detecting deletions in different pedigree structures where grandparents or children were added to a trio.

Methods

Definitions and assumptions

In the following, we will consider a null allele at a genetic marker to be one that consistently fails to produce a detectable product or phenotype. We assume that dosage cannot be determined such that a heterozygote for a null allele is not distinguishable from a homozygote for the other allele. Further, we assume that there are no additional genotyping errors. We also assume that all null alleles present in the families are inherited, ie they are not due to *de novo* mutations within the pedigree. Null alleles can then be detected as a special case of apparent non-Mendelian inheritance; a parent and child will appear to be homozygous for different alleles (Figure 1). This event is denoted by C (for 'confirmed'). Note that the parent–offspring combination can be anywhere within the pedigree. In this paper, we do not consider the event of an individual with no detectable phenotype (missing value) as evidence of a homozygote for a null allele; such a result could also be due to bad quality DNA and technical mistakes during the genotyping process. In the calculations below we will consider one locus with the marker alleles A_1, A_2, …, A_m and a null allele A_0. The allele frequencies will be denoted by p_0 for the deletion/null allele and p_i,

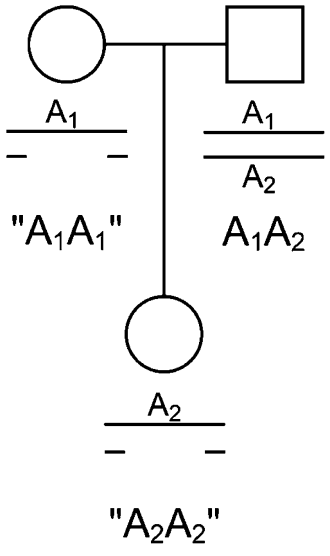

Figure 1 Illustration of actual genotypes and marker phenotypes and in a trio with a deletion. Citation marks indicate a misinterpreted genotype.

$i=1,\ldots, m$ are the frequencies of the marker alleles. For a biallelic marker, p_1 and p_2 denotes the frequencies of the two marker alleles. Some analyses are extended to two bialleleic loci (A and B), which are encompassed by the same deletion. The proportions of the five haplotypes will then be denoted by P_{11}, P_{12}, P_{21}, P_{22} and P_{00}, where P_{ij} is the proportion of the A_iB_j haplotype and P_{00} is the allele frequency of the deletion.

Calculations and comparisons

As mentioned above, the criterion for confirming a deletion is to observe a parent and a child who are homozygous for different alleles. Such a pattern will, along with other kinds of deviation from Mendelian inheritance, be readily detected by a program such as PedCheck.[26]

Expressions for the probability of C, $P(C)$, were derived for a number of different family structures as a function of allele frequencies for the deletion and the marker alleles (Appendix 1). It is trivial that $P(C)$ is higher in a larger family than in a smaller family, but larger families also require more individuals to be analysed. Thus, the relevant comparison is for a fixed total number of sampled individuals. We wanted to investigate which sampling design is most efficient, ie would give the highest probability to detect a null allele at a certain frequency in the population in at least one family. The probability to confirm a deletion in at least one family is calculated as $1-(1-P(C))^N$, where $P(C)$ is the probability to confirm a deletion in the family types respectively and N is the number of that family type.

Two series of comparisons were made. In the first comparison, the effect of adding parents and grandparents were investigated. In the second comparison, we investigated the effect of adding additional children to a trio, by using nuclear families with a varying number of children.

Comparison I: adding parents and grandparents

The probability to detect a deletion for a fixed overall number of investigated individuals was calculated for SNPs and multiallelic markers with varying allele frequencies. The family types are: one parent and one child (in this paper referred to as a duo), a trio, a trio with one grandparent (a tetra) and a trio with two grandparents (a pento). The family types can be seen in Figure 2. To investigate which family type that is most efficient, we made comparisons assuming a total sample size of 120, corresponding to 60 duos, 40 trios, 30 tetra and 24 pento. The results are illustrated for deletion frequencies of 0.05 and 0.01. In spite of the comparisons being made for a specific number of investigated individuals, it has full generality (Appendix 2).

Comparison II: adding children

The probability to detect a deletion for a fixed overall number of investigated individuals was calculated for SNPs.

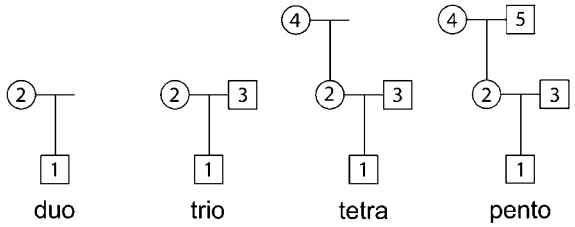

Figure 2 The family configurations used when studying the effect of adding parents and grandparents.

The probability to confirm a deletion was illustrated for a deletion frequency of 0.05 and a total of 120 individuals divided into 40 families with one child, 30 families with two children, 24 families with three children and 20 families with four children. The probability was also illustrated for a deletion frequency of 0.01 and a total of 420 individuals divided into 140 families with one child, 105 families with two children, 84 families with 3 children, 70 families with four children and 60 families with five children.

Results

Comparison I: adding parents and grandparents

The probabilities of confirming a deletion with allele frequency p_0 in the different pedigree structures using multiallelic markers are as follows:

Duo (one parent and one child). A deletion will be detected if the sampled parent has genotype A_iA_0 and the child inherits A_0 from this parent and A_j ($i\neq j$) from the unsampled parent. The probability for this event considering one allele is $2p_ip_00.5(1-p_i-p_0)$. The total probability summed over all alleles is $\sum_i 2p_ip_00.5(1-p_i-p_0)$. (Note: here and in all expressions below summation over i means for $i=1,\ldots, m$.) This can be simplified to

$$P(C_{\text{duo}}) = p_0 \sum_i p_i(1-p_i-p_0) \quad (1)$$

Trio (two parents and one child) give the following expression (Appendix 1 for derivation)

$$P(C_{\text{trio}}) = 2p_0 \sum_i p_i(1-p_i-p_0). \quad (2)$$

The probabilities to detect a null allele in a trio with one or two grandparents are as follows:

$$P(C_{\text{tetra}}) = 3p_0 \sum_i p_i(1-p_i-p_0) - 0.5p_0(1+p_0) \sum_i \sum_j p_ip_j(1-p_j-p_0) \quad (3)$$

for the tetra (trio with one grandparent) and

$$P(C_{\text{pento}}) = 4p_0 \sum_i p_i(1-p_i-p_0) - p_0(1+p_0) \sum_i \sum_j p_ip_j(1-p_j-p_0) \quad (4)$$

for the pento (trio with two grandparents). The double sums, $\sum_i \sum_j$, mean summation over $i=1,\ldots, m$ and $j=1,\ldots, m$ but skipping the cases when $i=j$.

For biallelic markers, such as SNPs, the expressions above reduce to

$$P(C_{\text{duo}}) = 2p_0p_1p_2 \quad (5)$$

$$P(C_{\text{trio}}) = 4p_0p_1p_2 \quad (6)$$

$$P(C_{\text{tetra}}) = 0.5p_0p_1p_2(11+p_0^2), \text{ and} \quad (7)$$

$$P(C_{\text{pento}}) = p_0p_1p_2(7+p_0^2). \quad (8)$$

For biallelic markers, it is possible to formally show which family type is most efficient. With a fixed total sample size of k, the probability to detect a deletion in at least one family is $1-(1-P(C))^N$ where $N=k/n$, and n is the number of individuals in a family. Comparing for example duos with trios then is to determine which of $1-(1-P(C_{\text{duo}}))^{k/2}$ and $1-(1-P(C_{\text{trio}}))^{k/3}$ that is largest. The actual comparisons can be seen in the Appendix 2.

The results for SNPs show that for a fixed total number of individuals, a sample of pentos is always more efficient than a sample of the other family types, despite a lower total number of founder chromosomes in the sample. We define founder chromosomes as chromosomes that are not inherited from an individual in the pedigree. The second most efficient family type is the tetra, while the trio is better than the duo. This is illustrated in Figure 3a and b where the probability to detect a deletion is plotted for deletion frequencies of 0.01 and 0.05, respectively. Thus, every addition of one ancestor increases the probability, even if the improvement is small, as when adding grandparents to a trio.

Some numerical examples of the results for multiallelic markers can be seen in Table 1. Rows 1–4 and 5–8 shows 2–5 equifrequent alleles and a deletion frequency of 0.01 and 0.05, respectively. The probabilities for detecting a deletion for multiallelic markers are higher than for biallelic markers. More alleles give higher probabilities. The probabilities for trio, tetra and pento are always very

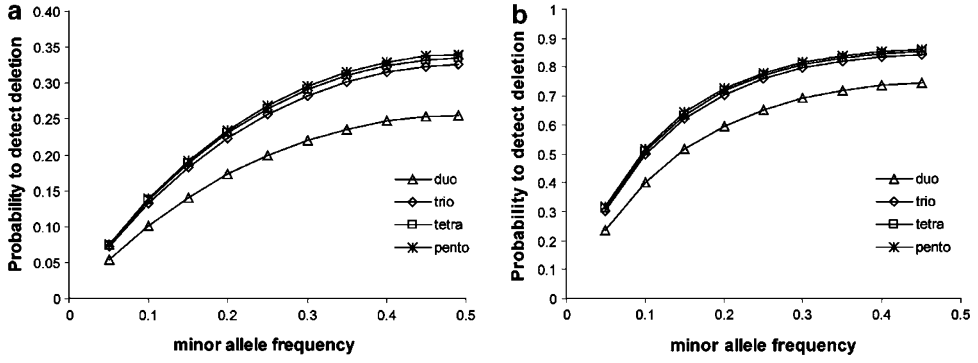

Figure 3 The probabilities to detect a deletion in at least one family as a function of the minor allele frequency for a total sample size of 120 (corresponding to 60 duo, 40 trio, 30 tetra, or 24 pento). (**a**) The frequency of the deletion is 0.01. (**b**) The frequency of the deletion is 0.05.

Table 1 Probabilities to detect deletions using multiallelic markers

Number of alleles	*Allele frequencies*	*Frequency of deletion*	*Duo*	*Trio*	*Tetra*	*Pento*
2	0.495×2	0.01	0.2553	0.3256	0.3344	**0.3398**
3	0.33×3	0.01	0.3252	0.4091	0.4098	**0.4105**
4	0.2475×4	0.01	0.3577	**0.4470**	0.4426	0.4403
5	0.198×5	0.01	0.3764	**0.4686**	0.4609	0.4566
2	0.475×2	0.05	0.7457	0.8423	0.8537	**0.8612**
3	0.3166, 0.3167, 0.3167	0.05	0.8400	0.9164	0.9187	**0.9209**
4	0.2375×4	0.05	0.8733	**0.9394**	0.9387	0.9391
5	0.19×5	0.05	0.8899	**0.9501**	0.9481	0.9476
4	0.1, 0.1, 0.1, 0.65	0.05	0.7447	0.8415	0.8451	**0.8482**
4	0.1, 0.2, 0.3, 0.35	0.05	0.8579	0.9290	0.9293	**0.9303**
3	0.1, 0.1, 0.75	0.05	0.6201	0.7277	0.7360	**0.7417**
10	0.095×10	0.05	0.9169	**0.9662**	0.9624	0.9605

Some examples for 60 duo, 40 trio, 30 tetra and 24 pento. The pedigree structure with the highest probability to detect a deletion in at least one family is marked in bold.

similar and substantially higher than for a duo. Which family configuration is best varies with the number of alleles and allele frequencies. A general trend where trios are best for many alleles can be seen, as well as that pentos are best when one allele is very common.

Comparison II: adding children

For a nuclear family with b children, a deletion can be detected when at least one parent and at least one child will appear to be homozygous for different alleles. The probability for this is (Appendix 1 for a derivation):

$$P(C_{\text{family}})=(1-0.5^b)4p_0\left(\sum_i p_i(1-p_i-p_0)+p_0\sum_i\sum_j p_ip_j\right)+(1-0.75^b)8p_0^2\sum_i p_i(1-p_i-p_0). \quad (9)$$

For a biallelic marker this reduces to

$$P(C_{\text{family}}) =4p_0p_1p_2[(1+p_0)(1-0.5^b)+2(1-p_0)(1-0.75^b)] \quad (10)$$

where b is the number of children. With an infinite number of children Equation (10) approaches the limit $4p_1p_2p_0(3-p_0)$. Note that the probability to confirm a null allele does not approach one when p_0 approaches one. The reason for this is that individuals with no detectable genotype are discarded from the analysis.

The effect of varying the number of children in the expression above was investigated numerically. Adding the first extra child to a trio increases the efficiency per investigated individual, adding the second does for some allele frequencies, but adding more children makes the family structure less efficient (Figure 4a and b). The optimal size of a nuclear family is thus one with two or three children. The probability to detect a least one deletion is very similar for two and three children, with a higher probability for two children for uneven allele frequencies and higher probability for three children with more even allele frequencies (Table 2).

Adding grandparents or children: which is better?

To answer this question, we compared the probabilities to detect a deletion for a biallelic marker for nuclear families with two or three children with a tetra and a pento, respectively. The probability for a biallelic marker in a single family is higher for a nuclear family with two children than for a tetra. The probability is thus also higher given a fixed sample size since the number of individuals in each family is the same. Comparing a nuclear family with three children with a pento in the same way gives the result that a nuclear family with three children is more efficient for detecting deletions than a pento (Table 2). The conclusion is thus that it is more efficient to add children than grandparents.

Multiple loci

So far only single marker analysis has been considered. Since a deductive method rather than inferential is under investigation here, genuine multipoint analysis does not exist as such. The information from several loci that do not confirm the existence of null alleles cannot be combined to confirmation in the sense above. If two adjacent markers confirm the presence of a null allele simultaneously, then it can be concluded that a deletion encompass both markers. If more than one of the investigated markers are situated within the deleted region, it is a higher probability that at least one marker will confirm a null allele. For this reason, two aspects of the simultaneous analysis of two markers have been considered, the probability that at least one marker will confirm a null allele, $P(C, \geq 1)$, and the probability that both will confirm null-alleles and thereby a deletion, $P(C, 2)$. This is done for two biallelic markers in the duo and trio family structure, respectively.

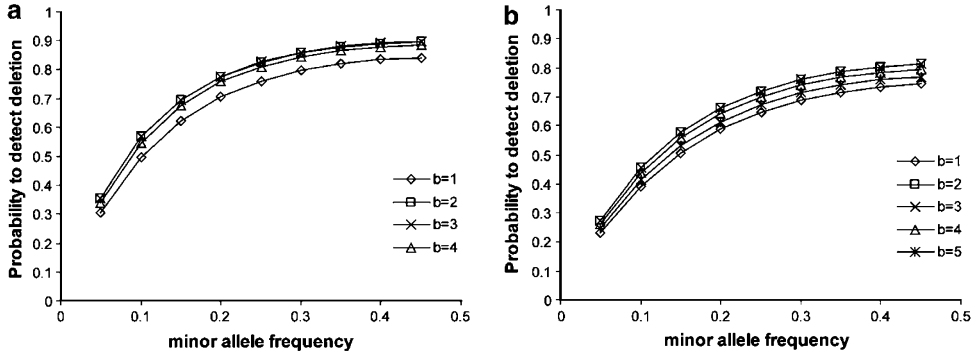

Figure 4 (**a**) The probabilities to detect a deletion in at least one family as a function of the minor allele frequency for a total sample size of 120 (corresponding to 40 families with one child, 30 families with two children, 24 families with three children, or 20 families with four children) when the frequency of the deletion is 0.05. (**b**) The probabilities to detect a deletion in at least one family as a function of the minor allele frequency for a total sample size of 420 (corresponding to 140 families with one child, 105 families with two children, 84 families with three children, 70 families with four children, or 60 families with five children) when the frequency of the deletion is 0.01.

Table 2 Probabilities to detect deletions using biallelic markers

		Pedigree structure					
		Number of children in nuclear family					
Frequency of deletion	*Minor allele frequency*	*1*	*2*	*3*	*4*	*Tetra*	*Pento*
0.01	0.05	0.072508	**0.087639**	0.087619	0.083032	0.074713	0.076044
0.01	0.1	0.132946	0.159636	**0.159651**	0.151650	0.136889	0.139283
0.01	0.15	0.182995	0.218485	**0.218564**	0.208032	0.188301	0.191538
0.01	0.2	0.224000	0.266152	**0.266306**	0.253906	0.230367	0.234268
0.01	0.25	0.257031	0.304174	**0.304401**	0.290638	0.264214	0.268629
0.01	0.3	0.282928	0.333740	**0.334033**	0.319290	0.290725	0.295530
0.01	0.35	0.302334	0.355751	**0.356097**	0.340674	0.310576	0.315665
0.01	0.4	0.315724	0.370863	**0.371249**	0.355383	0.324264	0.329544
0.01	0.45	0.323418	0.379520	**0.379929**	0.363818	0.332127	0.337515
0.1	0.05	0.496336	**0.565629**	0.564611	0.543496	0.508468	0.516105
0.1	0.1	0.727721	0.795945	**0.796058**	0.777439	0.741064	0.749709
0.1	0.15	0.841461	0.895639	**0.896395**	0.883136	0.853094	0.860753
0.1	0.2	0.900258	0.941528	**0.942475**	0.933365	0.909830	0.916182
0.1	0.25	0.932007	0.963907	**0.964842**	0.958422	0.939879	0.945117
0.1	0.3	0.949659	0.975334	**0.976197**	0.971416	0.956304	0.960724
0.1	0.35	0.959443	0.981263	**0.982056**	0.978223	0.965283	0.969162
0.1	0.4	0.964395	0.984131	**0.984878**	0.981535	0.969784	0.973360
0.1	0.45	0.965911	0.984989	**0.985721**	0.982529	0.971155	0.974633

The total sample size is 120. The pedigree structure with the highest probability to detect a deletion in at least one family is marked in bold.

For a duo, the probability of detection of a null allele at one locus or both is

$$P(C_{\text{duo}}, \geq 1) = 2P_{00}(P_{11}P_{22} + P_{12}P_{21} + P_{11}P_{12} + P_{11}P_{21} + P_{22}P_{12} + P_{22}P_{21}) \quad (11)$$

And the probability of detection at both loci is

$$P(C_{\text{duo}}, 2) = 2P_{00}(P_{11}P_{22} + P_{12}P_{21}). \quad (12)$$

For a trio, the corresponding probabilities are

$$P(C_{\text{trio}}, \geq 1) = 4P_{00}(P_{11}P_{22} + P_{12}P_{21} + P_{11}P_{12} + P_{11}P_{21} + P_{22}P_{12} + P_{22}P_{21}) \quad (13)$$

and

$$P(C_{\text{trio}}, 2) = 4P_{00}(P_{11}P_{22} + P_{12}P_{21}). \quad (14)$$

The derivation of Equations (11–14) is in Appendix 1. The ratio between $P(C, \geq 1)$ and $P(C, 2)$ can take any value between 0 and 1. One obvious result is that, if linkage disequilibrium (LD) is complete in the sense that either $P_{11} = P_{22} = 0$ or $P_{12} = P_{21} = 0$, then $P(C, \geq 1) = P(C, 2)$, ie detection always occurs simultaneously for both markers. If on the other hand LD is zero and allele frequencies are equal, $P(C, 2)/P(C, \geq 1) = 1/3$. On the other hand, under these conditions, $P(C, \geq 1)$ takes its maximum value. Since closely situated markers often are in LD and markers are selected for high minor allele frequencies, the probability of simultaneous detection will be relatively high.

Effect of missing data

The calculations above are derived under the assumption that no data is missing. The effect of missing data varies among the different family types. If it is assumed that γ is the probability of missing one genotype in one individual the probability of a complete duo is $(1-\gamma)^2 \approx 1-2\gamma$, if γ is small. Thus, the probability of an incomplete duo is 2γ. In either case, whether it is the parent or the child that is missing, a duo will be lost leaving just a single individual. In the case of a trio, the probability of loss is 3γ in analogy. However, if either of the parents is lost, a duo will remain. The probability of a completely lost family is only γ. For a tetra, the probability of a remaining trio is γ, a three generation trio is γ (ie grandparent, parent and child), and a duo will remain with the probability 2γ. For a pento, a tetra will remain with the probability 3γ, a trio with the probability γ and a duo with the probability γ. Thus, when individuals are lost in the larger family types, the remaining individuals will form an informative family, a fact that emphasizes the relative efficiency of larger families.

Discussion

Every method to detect deletions has advantages and disadvantages. Because of this, physical and segregational methods have been combined to support each other in the search for deletion. For example, this was the strategy in the investigation by Conrad *et al*[5] and McCarroll *et al*.[10] One distinct advantage of using markers in families is that it directly allows for deduction of marker haplotypes outside the deletion, on the chromosome carrying the deletion. If more than one of the investigated individuals is found to carry a deletion within a certain region, the haplotype for the adjacent markers can be used to investigate whether the deletion has a single origin, or if

the same region has been deleted several times. Moreover, if a deletion is found to have a single origin, the extended haplotype can be used to estimate the age of the deletion.

An important aspect is that in many cases, the marker data already is available. These data might have been generated for purposes such as family-based association, linkage mapping or characterization of the pattern of LD in a region. If a set of markers has been typed in families, these data should be further analysed for possible deletions or other null alleles. Merely discarding data that does not fit Mendelian inheritance might lead to loss of important information. As an illustration, Amos[27] showed that the pattern of missing values can carry important information. Amos[27] found that missing values for microsatellites had the same value for reconstructing a population tree as the data itself. This implied that in the dataset he analysed, a large proportion of the missing values were due to null allele homozygotes. When a single marker indicate a null allele, it is not possible to distinguish between a deletion and other causes. This may be considered a problem if the goal is to identify deletions. However, this information is vital if inference methods are used to detect null alleles and estimate their frequencies.

The probability to detect a null allele for a single marker using segregation in pedigrees is higher than the probability to get departure from Hardy–Weinberg equilibrium using the same number of unrelated individuals (AM Johansson unpublished data). Therefore, family data are valuable for detecting deletions.

A rather strong assumption underlying the analysis is that of perfect genotyping. This is of course not the case in a real investigation. The limitation imposed by this assumption varies with the conditions. If the marker method is combined with a physical method as in McCarroll *et al*,[10] it is a very limited problem. A deletion is then confirmed when the two methods both indicate a deletion. If several adjacent markers indicate a deletion, its existence can be inferred with great certainty without independent confirmation. The most critical situation is then when a single marker in a single generational step indicates a null-allele. Then, genotyping errors, deletions and other null alleles cannot be distinguished. This problem can be accounted for in a number of ways. One is to introduce probabilities for genotyping errors. Such an approach is presented for trios by Amos *et al*[21] A problem with this method is that it relies on estimates of genotyping errors. An alternative way would be to re-type interesting SNPs both with the same set of primers and with newly designed primers. If the pattern is repeated with the original primers, a genotyping error can be excluded whereas differentiation between deletions and other null-alleles is not possible. If newly designed primers also indicate a null-allele it is with great certainty a deletion.

An important problem when planning a study is whether it is more efficient to increase the size of the pedigrees by adding grandparents or to sample more children. We showed that addition of children to the pedigree is more efficient. This result is both surprising and encouraging, because it is much easier to collect information from additional children rather than a grandparent, since the grandparents may no longer be living. However, for other purposes, it might be better to have a design with grandparents since they would provide additional founder chromosomes and also make inferences about haplotypes more accurate.

The result that larger families are more efficient in spite of the overall smaller number of founder-chromosomes might seem surprising. However, it is only when a null-allele is transmitted from a typed parent to a typed child that detection is possible. Thus, the relevant comparison between experimental designs is to count the total number of transmitted chromosomes that can be observed. In a duo, only one transmitted chromosome can be observed per family. If the total number of investigated individuals is n_T, then a total of $0.5n_T$ transmitted chromosomes can be observed. If trios are used, $n_T/3$ families can be studied but they represent two transmitted chromosomes each. Thus $(n_T/3)2 \approx 0.67n_T$ transmitted chromosomes are observed in total. In a tetra, there are three transmissions but only two of them definitely represent founder chromosomes. The third is a founder chromosome with the probability 0.5 and an already observed chromosome with the probability 0.5. A retransmission is of course less informative than the transmission of a founder chromosome but is still useful since it is possible to miss a deletion in the first transmission and detect it in the second. The relative value of a retransmission depends on the allele frequencies of the loci involved but if we arbitrarily assign a value of 0.5, the expected number of transmissions within a tetra is $2 + 0.5 \times 1 + 0.5 \times 0.5 = 2.75$. Under this assumption, the expected total number of observable transmitted chromosomes using tetras will be $(n_T/4)2.75 \approx 0.69n_T$. Using the same assumption, a pento represents 3.5-transmitted chromosomes and as a consequence the total number will be $(n_T/5)3.5 = 0.70n_T$. These calculations give an explanation of both why larger families tend to be more efficient and also why the largest jump in efficiency occurs between duos and trios. If the same calculation is made for a nuclear family with two children, the expected number will be $2 + 0.25 \times 2 + 0.5 \times 1.5 + 0.25 \times 1 = 3.5$, yielding a total of $(n_T/4)3.5 \approx 0.88n_T$ chromosomes, which is compatible with the observations above that adding children is more efficient than adding parents.

Acknowledgements

We thank Professor Carsten Wiuf, Alison Anastasio and two anonymous reviewers for valuable comments. This study was funded by the Nilsson-Ehle foundation. R code implementing Equations (1-4) can be obtained from the authors upon request.

References

1 Lewis R: *Human Genetics: Concepts and Applications*, 7th edn. New York: McGraw-Hill, 2007.
2 Sharp AJ, Cheng Z, Eichler EE: Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 2006; **7**: 407–442.
3 Dutt A, Beroukhim R: Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol* 2007; **19**: 43–49.
4 Sebat J, Lakshmi B, Malhotra D *et al*: Strong association of *de novo* copy number mutations with autism. *Science* 2007; **316**: 445–449.
5 Conrad DF, Andrews TD, Carter NP, Hurles ME, Prichard JK: A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006; **38**: 75–81.
6 Estivill X, Armengol L: Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 2007; **3**: e190.
7 Gilles F, Goy A, Remache Y, Manova K, Zelenetz AD: Cloning and characterization of a Golgin-related gene from the large-scale polymorphism linked to the PML gene. *Genomics* 2000; **70**: 364–374.
8 Silverstein S, Lerer I, Buiting K, Abeliovich D: The 28-kb deletion spanning D15S63 is a polymorphic variant in the Ashkenazi Jewish population. *Am J Hum Genet* 2001; **68**: 261–263.
9 Hinds DA, Kloek AP, Jen M, Chen X, Frazier KA: Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 2006; **38**: 82–85.
10 McCarroll SA, Hadnott TN, Perry GH *et al*: Common deletion polymorphisms in the human genome. *Nat Genet* 2006; **38**: 86–92.
11 Redon R, Ishikawa S, Fitch KR *et al*: Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
12 Wong KK, deLeeuw RJ, Dosanjh NS *et al*: A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 2007; **80**: 91–104.
13 Gerstein AC, Chun HJE, Grant A, Otto SP: Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet* 2006; **2**: 1396–1401.
14 Schouten JP, McElgunn CJ, Waaijer R *et al*: Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acid Res* 2002; **30**: e57.
15 Pinkel D, Segraves R, Sudar D *et al*: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 1998; **20**: 207–211.
16 Alonso J, García-Miguel P, Abelairas J, Mendiola M, Peastaña Á: A microsatellite fluorescent method for linkage analysis in familial retinoblastoma and deletion detection at the RB1 locus in retinoblastoma and osteosarcoma. *Diagn Mol Pathol* 2001; **10**: 9–14.
17 Yu CE, Dawson G, Munson J *et al*: Presence of large deletions in kindreds with autism. *Am J Hum Genet* 2002; **71**: 100–115.
18 Johansson AM, Hillarp A, Säll T *et al*: Large deletions of the PROS1 gene in a large fraction of mutation-negative patients with protein S deficiency. *Thromb Haemost* 2005; **94**: 951–957.
19 Daiger SP, Chakravarti A: Deletion mapping of polymorphic loci by apparent parental exclusion. *Am J Med Genet* 1983; **14**: 43–48.
20 Wilkie AOM: Detection of cryptic abnormalities in unexplained retardation: a general strategy using hypervariable subtelomeric DNA polymorphisms. *Am J Hum Genet* 1993; **53**: 688–701.
21 Amos CI, Shete S, Chen J, Yu RK: Positional identification of microdeletions with genetic markers. *Hum Hered* 2003; **56**: 107–118.
22 Johansson AM, Halldén C, Säll T: Detecting deletions in families affected by a dominant disease by use of marker data. *Hum Hered* 2005; **60**: 26–35.
23 Kosta K, Sabroe I, Göke J *et al*: A bayesian approach to copy-number polymorphism analysis in nuclear pedigrees. *Am J Hum Genet* 2007; **81**: 808–812.
24 Wang K, Li M, Hadley D *et al*: PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**: 1665–1674.
25 Kohler JR, Cutler DJ: Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. *Am J Hum Genet* 2007; **81**: 684–699.
26 O'Connell JR, Weeks DE: PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998; **63**: 259–266.
27 Amos W: The hidden value of missing genotypes. *Mol Biol Evol* 2006; **23**: 1995–1996.

Appendix 1

Derivation of $P(C)$ for the different family types:

In the following C_{ij} will be used as the event that a null allele is confirmed by transmission of a deletion from individual i to individual j.

Duo:

In a duo only one transmission occurs, from the parent to the child, ie C_{21}, where the numbers refer to the numbering of individuals in Figure 2. The derivation of the probability for this is shown in the main text, where it is shown that

$$P(C_{\text{duo}}) = P(C_{21}) = p_0 \sum_i p_i(1 - p_i - p_0).$$

Trio:

In a trio confirmation can occur through two paths, C_{21} and C_{31}. Thus,

$$P(C_{trio}) = P(C_{21} \cup C_{31}) = P(C_{21}) + P(C_{31}) - P(C_{21} \cap C_{31}).$$

Here, $P(C_{31}) = P(C_{21}) = p_0 \sum p_i(1 - p_i - p_0)$ (see above) and $P(C_{21} \cap C_{31}) = 0$. The latter since homozygosity for a deletion is scored as missing data.

Accordingly, $P(C_{\text{trio}}) = 2p_0 \sum_i p_i(1 - p_i - p_0)$.

Tetra:

In a tetra, confirmation can occur through three paths, C_{42}, C_{21} and C_{31}. Thus,

$$\begin{aligned} P(C_{\text{tetra}}) =& P(C_{21} \cup C_{31} \cup C_{42}) = P(C_{21}) + P(C_{31}) + P(C_{42}) \\ & - P(C_{21} \cap C_{31}) - P(C_{21} \cap C_{42}) - P(C_{31} \cap C_{42}) \\ & + P(C_{21} \cap C_{31} \cap C_{42}). \end{aligned}$$

Here, $P(C_{21}) = P(C_{31}) = P(C_{42}) = p_0 \sum p_i(1 - p_i - p_0)$, and $P(C_{21} \cap C_{31}) = P(C_{21} \cup C_{31} \cup C_{42}) = 0$.

$P(C_{21} \cap C_{42})$ can be found in the following way. (In the following '*i*' will refer to individual number i in the family.) '4' has to be heterozygous for the deletion and a specific allele A_i, which occurs with the probability $2p_0p_i$. Then, '4' has to segregate the deletion to '2' which occurs with probability 0.5. In addition '2' has to receive an allele A_j, $i \neq j$, from the unscored parent which occurs with probability p_j. Finally '1' has to receive the deletion from '2' and an allele from '3' which is different from A_j. This occurs with probability $0.5(1-p_j-p_0)$. The combined probability given A_i in '4' and A_j in '2' is $0.5p_0p_ip_j(1-p_j-p_0)$ which has to be summed over all combination of alleles to get the full probability, ie $0.5p_0 \sum_i \sum_j p_ip_j(1 - p_j - p_0)$.

$P(C_{31} \cap C_{42})$ can be found in the following way. The '4' has to be heterozygous for the deletion and a specific allele A_i, which occurs with the probability $2p_0p_i$. Then, the '4' has to segregate the deletion to '2' which occurs with probability 0.5, and the '2' has to receive an allele A_j, $i \neq j$, from the unscored parent which occurs with probability p_j. In the next generation, '2' has to segregate A_j to '1', which occurs with probability 0.5. Finally, '3' has to be heterozygous for the deletion, and an allele which is different from A_j and segregates the deletion to '1'. The probability for this is $2p_0(1-p_j-p_0)0.5$. Accordingly, $P(C_{31} \cap C_{42}) = 0.5 p_0^2 \sum_i \sum_j p_i p_j (1 - p_j - p_0)$.

Finally, the probability for a tetra is achieved from combining the terms, ie

$$P(C_{\text{tetra}}) = 3p_0 \sum_i p_i(1 - p_i - p_0) - 0.5p_0(1 + p_0) \sum_i \sum_j p_i p_j (1 - p_j - p_0).$$

Pento:

In a pento, confirmation can occur through four paths, C_{21}, C_{31}, C_{42} and C_{52}. Thus,

$$\begin{aligned} P(C_{\text{pento}}) =& P(C_{21} \cup C_{31} \cup C_{42} \cup C_{52}) \\ =& P(C_{21}) + P(C_{31}) + P(C_{42}) + P(C_{52}) \\ & - P(C_{21} \cap C_{31}) - P(C_{21} \cap C_{42}) - P(C_{21} \cap C_{52}) \\ & - P(C_{31} \cap C_{42}) - P(C_{31} \cap C_{52}) - P(C_{42} \cap C_{52}). \\ & + P(C_{21} \cap C_{31} \cap C_{42}) + P(C_{21} \cap C_{31} \cap C_{52}) \\ & + P(C_{21} \cap C_{42} \cap C_{52}) + P(C_{31} \cap C_{42} \cap C_{52}) \\ & - P(C_{21} \cap C_{31} \cap C_{42} \cap C_{52}) \end{aligned}$$

Here,

$$P(C_{52}) = P(C_{42}) = P(C_{31}) = P(C_{21}) = p_0 \sum_i p_i(1 - p_i - p_0),$$

and

$$\begin{aligned} P(C_{21} \cap C_{31}) =& P(C_{42} \cap C_{52}) = P(C_{21} \cap C_{31} \cap C_{42}) \\ =& P(C_{21} \cap C_{31} \cap C_{52}) \\ =& P(C_{21} \cap C_{42} \cap C_{52}) = P(C_{31} \cap C_{42} \cap C_{52}) \\ =& P(C_{21} \cap C_{31} \cap C_{42} \cap C_{52}) = 0. \end{aligned}$$

$P(C_{21} \cap C_{52}) = P(C_{21} \cap C_{42})$ which is equal to $P(C_{21} \cap C_{42})$ in a tetra, ie $0.5p_0 \sum_i \sum_j p_i p_j (1 - p_j - p_0)$.

$P(C_{31} \cap C_{52}) = P(C_{31} \cap C_{42})$ which is equal to $P(C_{31} \cap C_{42})$ in a tetra, ie $0.5p_0^2 \sum_i \sum_j p_i p_j (1 - p_j - p_0)$.

Finally, the probability for a pento is achieved from combining the terms, ie

$$P(C_{\text{pento}}) = 4p_0 \sum_i p_i(1 - p_i - p_0) - p_0(1 + p_0) \sum_i \sum_j p_i p_j (1 - p_j - p_0).$$

Nuclear family:

For a nuclear family with b children, a deletion can be detected when at least one parent and at least one child will appear to be homozygous for different alleles. This happens if one or both parents are heterozygous for the deletion, and a child inherits this deletion and then inherits another allele than the parent with the deletion has from the parent it does not inherit the deletion from. This happens if

(1) one parent has genotype A_iA_0 and the other parent has genotype A_jA_k, where $i \neq j$ and $i \neq k$ but where $j = k$ is not excluded. At least one child out of b should be A_jA_0 or A_kA_0. The probability for this event is $2[2p_i2p_0(1-p_i-p_0)(1-0.5^b)]$,
(2) one parent has genotype A_iA_0 and the other parent has genotype A_iA_j where $i \neq j$ and at least one child out of b having genotype A_jA_0. The probability for this event is $2[2p_0p_i2p_0(1-p_i-p_0)(1-0.75^b)]$,
(3) one parent has genotype A_iA_0 and the other parent has genotype A_jA_0 where $i \neq j$ and at least one out of b children has genotype A_iA_0 or A_jA_0. The probability for this event is $2[2p_0p_i2p_0p_j(1-p_i-p_0)(1-0.5^b)]$.

The probability to detect a null allele in a nuclear family is thus the sum of the probabilities from 1, 2 and 3 which can be simplified to

$$P(C_{\text{family}}) = (1 - 0.5^b)4p_0 \left(\sum_i p_i(1 - p_i - p_0) + p_0 \sum_i \sum_j p_i p_j \right) + (1 - 0.75^b)8p_0^2 \sum_i p_i(1 - p_i - p_0).$$

For a biallelic marker this reduces to

$$4p_0p_1p_2[(1 + p_0)(1 - 0.5^b) + 2(1 - p_0)(1 - 0.75^b)]$$

As a control, we can set b to 1 and use only two alleles except the null allele, which gives

$$\begin{aligned} & 4p_0p_1p_2[(1 + p)0.5 + 2(1 - p)0.25] \\ & = 4p_0p_1p_2(0.5 + 0.5p_0 + 0.5 - 0.5p_0) = 4p_0p_1p_2. \end{aligned}$$

This is Equation (6) for the trio in the main text.

Analysis of two markers:

Below it is assumed that a deletion encompass two marker loci, A and B, both being bialleleic.

Duo:

Simultaneous confirmation of null alleles at two adjacent loci, ie a deletion, occurs if, (1) the parent is heterozygous for the deletion, (2) the parent segregates the deletion to the child and (3) the gamete from the unscored parent carries the complementing haplotype to the haplotype of the other chromosome of the parent. There are two pairs of complementing haplotypes, ie A_1B_1 *vs* A_2B_2 and A_1B_2 *vs* A_2B_1. For example, one case of simultaneous detection at two loci occurs if the parent is A_1B_1/del, it segregates *del* and the gamete from

the unscored parent carries A_2B_2, which occurs with the probability $(2P_{00}P_{11})P_{22}\frac{1}{2}$. The total probability will then be:

$$(2P_{00}P_{11})P_{22}\tfrac{1}{2} + (2P_{00}P_{22})P_{11}\tfrac{1}{2} + (2P_{00}P_{12})P_{21}\tfrac{1}{2} + (2P_{00}P_{21})P_{12}\tfrac{1}{2} = 2P_{00}(P_{11}P_{22} + P_{12}P_{21}).$$

The probability for a confirmation at one or two loci can be found using the relation, $P(A \cup B) = \mathrm{P}(A) + P(B) - P(A \cap B)$, where A means detection at marker locus A and B at marker locus B. $P(A)$ and $P(B)$ are found using Equation (5) in the main text where $p_0 = P_{00}$, $p_1 = P_{11} + P_{12}$ and $p_2 = P_{21} + P_{22}$ for A and $p_0 = P_{00}$, $p_1 = P_{11} + P_{21}$ and $p_2 = P_{12} + P_{22}$ for B. The total probability will then be

$$2P_{00}((P_{11} + P_{12})(P_{21} + P_{22})) + 2P_{00}((P_{11} + P_{21})(P_{12} + P_{22})) - 2P_{00}(P_{11}P_{22} + P_{12}P_{21}) = 2P_{00}(P_{11}P_{22} + P_{12}P_{21} + P_{11}P_{12} + P_{11}P_{21} + P_{22}P_{12} + P_{22}P_{21}).$$

Trio:

Just as in the case of a single locus the probability of the trio has to be twice the probability of a duo, ie the corresponding probabilities are

$$4P_{00}(P_{11}P_{22} + P_{12}P_{21})$$

and

$$4P_{00}(P_{11}P_{22} + P_{12}P_{21} + P_{11}P_{12} + P_{11}P_{21} + P_{22}P_{12} + P_{22}P_{21}).$$

Appendix 2

Below, we formally compare the efficiency of the different family types, given a fixed overall number, n_T, of tested individuals. Only cases where complete families of each type are tested will be considered. Consider a comparison of family type I, where a single family includes n_I individuals and type J including n_J individuals each. Then n_T/n_I families of type I and n_T/n_J families of type J will be compared with respect to the probability of detecting at least one case of the deletion. If $Q(\text{type-}X) = (1 - P(C_{\text{type}X}))$ is the probability of a single family of not detecting the deletion, the comparison can be made through the ratio

$$Q(\text{type}I)^{\frac{n_T}{n_I}} / Q(\text{type}J)^{\frac{n_T}{n_J}}.$$

If the ratio is less than 1, then the ratio raised to any number, ie n_T, will also be less than one and equally, if the ratio is larger than one it will continue to be so when raised to n_T. As a consequence, one can choose $n_T/n_I = n_J$ and $n_T/n_J = n_I$, retaining full generality. Accordingly, $P(C_{\text{type}I}, n_J) - P(C_{\text{type}J}, n_I) = [1 - (1 - P(C_{\text{type}I})^{n_J}] - [1 - (1 - P(C_{\text{type}J})^{n_I}]$ has been calculated for each comparison of family types below.

Comparison of duo and trio

Using the method described above,

$$P(C_{\text{trio}}, 2) - P(C_{\text{duo}}, 3) = [1 - (1 - 4p_0p_1p_2)^2] - [1 - (1 - 2p_0p_1p_2)^3] = [8p_0p_1p_2 - 16p_0^2p_1^2p_2^2] - [6p_0p_1p_2 - 12p_0^2p_1^2p_2^2 + 8p_0^3p_1^3p_2^3] = 2p_0p_1p_2 - 4p_0^2p_1^2p_2^2 - 8p_0^3p_1^3p_2^3$$

Since $p_0 + p_1 + p_2 = 1$, the maximum value of $p_1p_2p_0$ is 1 of 27. In other words, the terms where p_1, p_2 and p_0 are raised to two needs a factor that is 27 times the factor multiplied to $p_1p_2p_0$ in order to influence the result. Hence, the difference above is positive and accordingly the trio design is more efficient than duo.

Comparison of trio and tetra

Using the method described above,

$$P(C_{\text{tetra}}, 3) - P(C_{\text{trio}}, 4) = [1 - (1 - 0.5p_0p_1p_2(11 + p_0^2))^3] - [1 - (1 - 4p_0p_1p_2)^4] = 0.5p_0p_1p_2 + 1.5p_0^3p_1p_2 + 5.25p_0^2p_1^2p_2^2 - 16.5p_0^4p_1^2p_2^2 - 89.625p_0^3p_1^3p_2^3 + g(p_0p_1p_2),$$

where $g(p_0p_1p_2)$ represents terms of higher order. Using the logic above, it is clear that this difference is positive and hence the tetra design is more efficient than the trio design.

Comparison of tetra and pento

Using the method described above,

$$P(C_{\text{pento}}, 4) - P(C_{\text{tetra}}, 5) = [1 - (1 - p_0p_1p_2(7 + p_0^2))^4] - [1 - (1 - 0.5p_0p_1p_2(11 + p_0^2)^5] = 0.5p_0p_1p_2 + 1.5p_0^3p_1p_2 + 8.5p_0^2p_1^2p_2^2 - 29p_0^4p_1^2p_2^2 - 291.75p_0^3p_1^3p_2^3 + g(p_0p_1p_2),$$

where $g(p_0p_1p_2)$ represents terms of higher order. Using the logic above, it is clear that this difference is positive and hence the pento design is more efficient than the tetra design.