Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects

Daniel Kling [a,b,d,*], Jenny Welander [c], Andreas Tillmar [a], Øivind Skare [d,e], Thore Egeland [b], Gunilla Holmlund [a,c]

[a] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Artillerigatan 12, SE-587 58, Linköping, Sweden
[b] Department for Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Norway
[c] Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, SE-581 85, Linköping, Sweden
[d] Norwegian Institute of Public Health, P.O. Box 4040, Nydalen, NO-0403 Oslo, Norway
[e] Department of Public Health and Primary Health Care, University of Bergen, P.O. Box 7804, NO-5020 Bergen, Norway

ARTICLE INFO

Article history:
Received 21 March 2011
Received in revised form 12 June 2011
Accepted 5 July 2011

Keywords:
DNA microarray
Kinship analyses
Affymetrix
Distant relatedness
Single nucleotide polymorphism
Merlin
FEST

ABSTRACT

In the past decades, microarray technology has definitely put an edge to the field of genetic research. Our aim was to determine whether single nucleotide polymorphism (SNP) microarrays could be used as a tool in establishing genetic relationships where current molecular genetic methods are not sufficient. We used the Genechip, Affymetrix GenomeWide SNP Array 6.0, which detects more than 900,000 SNP markers dispersed throughout the human genome. The intention was to find a good selection of SNP markers that could be used for statistical evaluation of relatedness in a forensic setting. We conducted pairwise comparisons in the R-package FEST as well as pedigree comparisons in Merlin. Our methods were applied on two separate families, where relationships as distant as 3rd cousins were known. In addition, a question about a possible common ancestry between the two families was tested. Relationships as distant as 2nd cousins could be readily distinguished both from unrelated and other, genetically, closer relationships. This was achieved with a selection of 5774 markers, where each pair of markers was separated by a genetic distance of at least 0.5 cM (centiMorgan). When considering 3rd cousins, and more distant relationships, the number of markers needs to be extended, consequently decreasing the genetic distance between the markers. However, inclusion of a too large number of markers presents new challenges and our results imply that the use of too dense sets of markers always yields the highest probability for the genetically closest relationship hypothesis. Simulations confirm that this is most probably caused by the fact that the computational model assumes linkage equilibrium between markers, a problem that will be further evaluated. Our results do however suggest that SNP-data derived from microarrays are well suited for kinship determination provided linkage disequilibrium is properly accounted for.

1. Introduction

In the past decades, the use of DNA has revolutionized many fields of research. It still remains the most important tool to trace genetic relationships, both in forensic casework and in clinical research. In medical research it is often crucial to accurately establish the relationships between the individuals participating in a study. In genetic association studies, unknown kinship between cases and also between controls, or even between these two groups, may give rise to false associations [1,2]. Also in linkage analysis the results can be seriously biased as a consequence of unknown relationships between pedigree founders [3].

In forensic casework, DNA can be used in crime scene investigations to find or exonerate a perpetrator. In paternity testing, the conventional problem is to determine whether a man is the biological father of a child. DNA analyses can also provide evidence to determine a disputed relationship more distant than first generation relatives. In particular, immigration cases often present genetic relationships where current forensic genetic methods do not produce sufficient evidence [4]. In forensic genetics, the choice of markers is at present mostly limited to short tandem repeats (STRs); genetically due to their high variability and their ability to provide a high power of discrimination; technically due to their suitability for multiplex PCR analyses. However, one of the disadvantages when using STR-markers is their high mutation rate. In addition the multiplex assays are often limited to 16–20 markers [5,6]. The use of single nucleotide polymorphisms (SNPs) has recently received some attention in the establishing of genetic relatedness. In the forensic field, the SNPforID Consortium has established a set of SNPs which

* Corresponding author at: Norwegian Institute of Public Health, Familiegenetikk, Gaustadalléen 30, NO-0027 Oslo, Norway. Tel.: +47 210 77663.
E-mail address: Daniel.Kling@fhi.no (D. Kling).

1872-4973/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2011.07.007

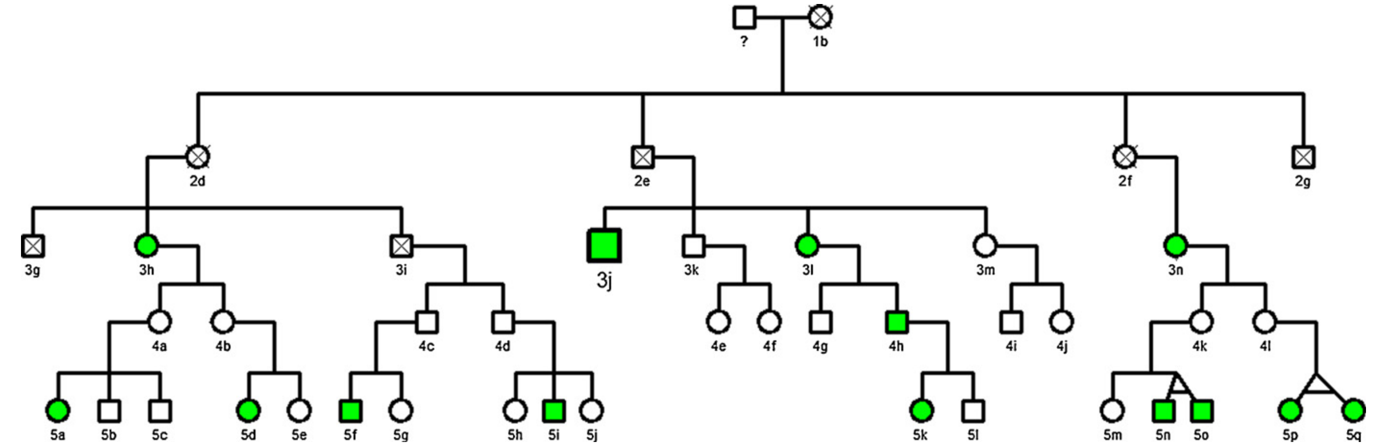

Fig. 1. Large Family. The pedigree describes a large family where relationships as distant as 3rd cousins were known. The question mark denotes an unknown paternal ancestor. Samples were drawn from the individuals marked with green.

performs sufficiently well to be used in court cases and can be multiplexed in one PCR reaction [7–10]. SNPs possess several advantages, which make them favourable when establishing complex or distant relationships. For one, they have a very low mutation rate, approximately 10^{-8} [11]. In addition, they can be analyzed in short amplicons and are generally easy to multiplex. Furthermore, there is an abundance of SNP markers to choose from in the human genome; The most recent paper from the HapMap project shows a map of 3.1 million of SNPs in the genome and the expected total number are 9–10 millions [12]. However, single SNPs provide very little genetic information, since they mostly are biallelic. The shortage of information can, however, be counteracted by analysing a larger number of markers. SNPs can be massively typed on high-density microarrays, such as the Genechips produced by Affymetrix or the HumanMap chips provided by Illumina, and have been extensively used in medical genetics [13]. A great number of markers is crucial in cases of distant relationships. The use of the standard STR markers, as well as a small set of SNP markers and a set of VNTR (Variable Number of Tandem Repeat) markers will not be enough [14]. Although easy to accomplish, the use of a larger number of markers presents challenges for the computational model used to distinguish between alternative pedigree hypotheses.

Different algorithms can be used for the purpose of calculating likelihood for a given pedigree and genotype data. They all share certain characteristics and the choice of which one to use is mainly depending on the number of markers and the number of individuals, see Gao et al. for a review [15]. One such algorithm is the Elston–Stewart algorithm [16,17], which can be described as a peeling algorithm and peels in the direction of individuals. This means that the calculation is only linear in the number of individuals. In contrast, the Lander–Green algorithm allows for a linear increase in the number of calculations to the number of markers [18]. The algorithm is implemented in the software Merlin, the main software used in this study [19]. The drawback is that both algorithms grow exponentially in one direction. In other words, the Elston–Stewart algorithm is capable of handling large pedigrees, but little genotype data, perhaps 100 markers, while the Lander–Green algorithm can handle hundreds and thousands of markers but only approximately 25 individuals in each pedigree. Besides this, the most prominent challenge, for any model, is to take genetic linkage and linkage disequilibrium (LD) properly into account. Genetic linkage has been shown, in simulation studies, to provide conclusive information in cases of relatedness [20,21]. The Lander–Green algorithm is able to take linkage into account, but assumes linkage equilibrium (LE). Therefore measures were taken to avoid the influence of LD, mainly by setting a minimum distance between the chosen markers, but also by using different sets of markers; see Supplemental Fig. S2 for a more thorough description of the selection procedure [15,19]. In addition an evaluation of possible LD for each selection of markers was carried out in PLINK [22].

In this study, we wanted to investigate if data from thousands of SNP markers could be used to resolve distant relatedness issues. For this purpose we used DNA from individuals representing different relationships known a priori and selected SNP-data derived from microarrays. We also applied our findings on a case of genealogy with a presumable half 1st cousin relationship.

2. Materials and methods

2.1. Sample data

Nineteen blood samples were collected from two families, Figs. 1 and 2, each presenting a wide selection of a priori known relationships, e.g. parent–child, grandparent–grandchild relations, full siblings, 1st cousins, 2nd cousins and 3rd cousins and uncle–niece. These known relationships were used to ascertain the validity of the statistical calculations as well as to establish which relationships could actually be determined. Finally, data from all tested individuals were used to establish whether or not the two families were related two generations back. Allele frequencies from 60 unrelated Swedish individuals were used as a reference population.

2.2. Simulations

Data were also simulated to further investigate the impact of linkage disequilibrium for different marker densities. The simula-

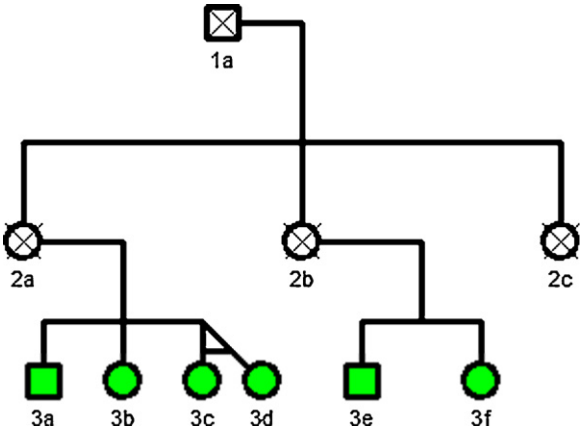

Fig. 2. Small Family. The pedigree describes a small family where relationships as distant as 1st cousins were known. Samples were drawn from the individuals marked with green.

tions were performed using FEST [20] where founder haplotypes with markers in linkage disequilibrium were drawn using the R package hapsim [23]. We used allele frequency information and LD data for chromosome 22 derived from HapMap for the CEPH (Utah residents with ancestry from northern and western Europe) population. To convert physical map distances (bp) to genetic map distances (cM), we used the Rutgers Combined Linkage-Physical Map of The Human Genome [24]. The following Bayesian approach was adopted; first, a true relation was drawn using a flat prior. Second, genotypes were simulated given the true relation: first the founder haplotypes assuming LD, then the genotypes of the descendants. Third, posterior probabilities were computed for each hypothesized family relation using Merlin. For each given marker density, these steps were repeated 5000 times, and then the posterior probabilities were averaged. Note that, if the likelihood computations were correct, the expected value of the posterior should equal the prior. This fact follows from $E[P[M = k|G]] = E[E[1_k(M)|G]] = E[1_k(M)] = P[M = k]$, where M is the family relation and G the genotype data. A bias in the averaged posterior probabilities, by not taking LD into account, would then be apparent as a deviation from the prior probabilities.

2.3. Microarray analysis

DNA was extracted as described by Lindblom and Holmlund [25]. The DNA concentration was quantified with Nanodrop (Thermo Scientific, Wilmington, DE, USA) and adjusted to 50 ng/μl prior to the microarray assay. Samples were analyzed on the Affymetrix GenomeWide SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) according to the manufacturer's protocol.

2.4. Selection of markers

The raw data was analyzed in the software Genotyping Console version 4.0 (GTC), supplied by Affymetrix. From the original 900,000 markers, different selections of autosomal SNP markers were made. The selection criteria included minor allele frequencies (MAF), minimum distances between two neighbouring markers as well as Hardy Weinberg p-value for each marker (see Supplemental Fig. S2 for a graphical explanation of the selection procedure). In addition to the previously mentioned criteria, a subsequent evaluation of the LD between selected markers was carried out in PLINK. Two different approaches to evaluate the presence of LD were used. First computation of pairwise r^2 values between each SNP and the 100 most proximally located SNPs. For each selection of markers, the fractions of pairwise r^2 values above a "limit" (limit = 0.1; 0.2; 0.3; 0.5; 0.8) were calculated. Second, we searched for the presence of haploblocks that can be defined as a cluster of closely located SNPs in strong LD [26]. The number of haploblocks was estimated from the "haplotype block estimation" option in PLINK [22]. This estimation uses the algorithm published by Gabriel et al. [27].

A more complex selection procedure could possibly account for information content, as described by Krawczak et al. [28]. This paper describes a formula which can be used to address the issue in paternity cases. However, we consider more general pedigrees using linked markers and therefore these measures of informativity cannot be used.

2.5. Statistical calculations

Likelihoods for the hypothetical pedigree structures were obtained from the software Merlin [19]. In addition the R-package FEST, which provides a front-end user interface to Merlin, was used to perform simple pairwise comparisons between individuals [20]. FEST lets the user include certain predefined hypotheses in the analysis. There are three different simple types of pairwise relationships: (1) S–n–m – the sharing of two common ancestors n and m generations back, (2) HS–n–m, the sharing of one common ancestor n and m generations back. When $n = m$, we abbreviate to S–n and HS–n. Finally (3) PC–n denotes a parent–child relationship spaced by n generations. FEST was used due to its relative ease with which it allows the user to calculate the likelihoods for a large number of alternative hypotheses. In addition FEST provides an in-built thinning procedure for genotype data. However, FEST has some constraints. Firstly, pedigree structures with inbreed loops and marriage loops are impossible to specify in terms of simple pairwise relationships. Secondly, inclusion of genotypes from more than two individuals in each analysis is impossible, which might be necessary in distant relatedness cases.

The likelihoods, obtained from Merlin and FEST, were converted to posterior probabilities according to a Bayesian approach using flat priors. An in-house software (freely available from the corresponding author), was used to perform extensive tests in Merlin. In this study three different minor allele frequencies were tested; 0.2, 0.3 and 0.4. For each minor allele frequency, 10 separate analyses were performed based on different minimum distances between selected markers. The minimum distance was evenly spaced between 0.05 and 2 cM, yielding approximately 49,000 and 1800 markers respectively. The numbers vary slightly depending on which minor allele frequency was chosen. In addition, for each minimum distance and MAF, three separate selections, not including the same SNPs, were made in order to minimize the possible influence of linkage disequilibrium.

2.6. Genotyping errors

Genotyping errors may have an impact on the calculations [29,30]. A study was undertaken to establish the degree of genotyping errors. One control sample was typed eleven consecutive times and approximately 4000 markers, approximately 0.4% of the original 900,000, were excluded from all analyses due to overrepresentation of inconclusive results. This is an *ad-hoc* solution that requires further development for future applications, possibly by inferring an error frequency and implementing this into the statistical model. One example of a model accounting for genotyping errors is provided by Epstein et al. [31].

3. Results

3.1. Pairwise comparisons with known relationships using FEST

Using different sets of markers, pairwise relationships were shown to yield high posterior probabilities for relationships as distant as 2nd cousins (Tables 1 and 2). In Table 1, the calculated posterior probabilities are shown based on a selection of 5774 markers for six known relationships. The first row contains the true relationships; S-1 denotes full siblings, S-2 full cousins, S-3 full 2nd cousins, S-4 full 3rd cousins, while HS-1 denotes half siblings, HS-2 half cousins and PC-2 a grandfather–grandchild relationship. Table 2 shows the results where instead a selection of 12,453 markers was used to calculate likelihoods (the number of comparisons included to calculate the averaged posterior for each true relationship depends on the available data, see Supplemental Table S1). When calculating the posterior probability for a 3rd cousin relationship, see S-4 Tables 1 and 2, the highest probability achieved was 0.9991 in favour of the true hypothesis, with a selection of 12,453 markers. Although sufficient to establish the 3rd cousin relationship, comparing two unrelated individuals only yielded 0.64 in favour of the

Table 1
Posterior probabilities for each tested relationship, based on a selection of 5774 SNP markers (markers separated by at least 0.5 cM). A Bayesian approach with flat priors has been used to calculate posterior probabilities.

True relationship	S-1	S-2	S-3	S-4	PC-2	Unrelated 1	Unrelated 2
S-1	>**0.99999**	<0.00001	–	–	<0.00001	–	–
HS-1	<0.00001	<0.00001	–	–	–	<0.00001	–
S-2	<0.00001	**0.993**	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	–	0.007	–	–	–	<0.00001	–
S-3	–	<0.00001	**0.9999**	–	–	0.0035	–
PC-2	–	–	–	–	**0.9999**	–	–
S-4	–	–	–	**0.81**	–	–	0.19
Unrelated	<0.00001	<0.00001	<0.00001	0.19	<0.00001	**0.997**	**0.81**

The true relationships in the first row and corresponding probability in bold. A hyphen in a specific row means exclusion of the relationship as an alternative hypothesis. S-1 means full siblings, HS-1 half-siblings, S-2 full 1st cousins, HS-2 half 1st cousins, S-3 full 2nd cousins, PC-2 grandparent–grandchild relation and S-4 means full 3rd cousins. The same 5774 markers have been used in all comparisons. Due to the varying availability of pairwise *true* relationships (Supplemental Table S1), the number of examples included for each relationship varies; For S-1 five comparisons, S-2 ten comparisons, S-3 four comparisons, S-4 ten comparisons, PC-2 nine comparisons, Unrelated ten comparisons.

unrelated relationship, see Table 2 and the comparison denoted *Unrelated 2*. Inclusion of a large number of markers revealed to always favour the genetically closest relationship, also when unrelated was the true relationship. The threshold value, when this phenomenon starts to occur depends on which relationship is tested. As a rule of thumb, when testing relationships closer than 2nd cousins, more than 20,000 SNP markers should not be included to obtain reliable results. See Fig. 3(a)–(c) which describe an approximate threshold for three different relationships, S-2, S-3 and unrelated.

3.2. Pairwise comparisons with "unknown" relationships using FEST

The question whether the two families in Figs. 1 and 2 were related to each other was first examined with FEST. Pairwise comparisons between the individuals in the third generation, i.e. 3a/3b/3c/3d/3e/3f and 3h/3j/3l/3n, Figs. 1 and 2, were performed. The following hypotheses were included, *unrelated*, *half 1st cousins* (HS-2) and *full 1st cousins* (S-2). Table 3 shows an extraction of the results with various selections of markers. All comparisons yielded high probabilities for the two families to be unrelated.

3.3. Comparisons with "unknown" relationships using Merlin

We tested alternative hypotheses for the unknown relationship between the two families in Figs. 1 and 2, including data from all typed individuals in the third generation. All tests, independent of marker selections, revealed high posterior probability for the unrelated hypothesis (Table 4). The hypotheses tested assumed, however, that the individuals in the family in Fig. 1 were full-cousins. Separate tests also confirmed this relationship (see Supplemental Table S2 and Fig. S1).

3.4. Evaluation of linkage disequilibrium using PLINK

For each selection of markers we performed pairwise LD evaluations in PLINK. We tested for LD between markers separated by less than 100 SNPs, which roughly means comparing markers located less than 50 Mb apart, in a selection of 5774 markers. Of course this distance depends not on the number of markers but on the minimum distance chosen between two selected markers, e.g. choosing markers separated by 0.1 cM yields a distance of roughly 10 Mb. Table 5 describes the results for a selection of marker sets. Evidently, selecting markers located 0.05 cM apart, roughly 29,200 markers, yields a higher percentage of r^2 values above 0.5, while in a selection of 5800 markers, the number is considerably lower. Also, r^2 values above 0.3 are comparatively rare in the latter selection. Furthermore, Fig. 4 describes the relation between number of markers and the number of haploblocks. According to the estimation the dependence is approximately exponential, meaning that choosing more markers will yield an exponential increase in the number of haploblocks, i.e. markers in tight LD.

3.5. Simulations using FEST

Table 6 summarizes the simulation results, based on genotype data from chromosome 22, where we consider the hypotheses full cousins (S-2), half cousins (HS-2) and unrelated. We see that by reducing the distance between markers, the averaged posterior probability is shifted progressively towards full cousins, the genetically closest relationship. These results are in concordance with our experience for real data (see also Supplemental Table S3 where the same simulations have been conducted without accounting for LD).

Table 2
Posterior probabilities for each tested relationships, based on 12,453 SNP markers (markers separated by at least 0.25 cM). A Bayesian approach with flat priors has been used to calculate posterior probabilities.

True relationship	S-1	S-2	S-3	S-4	PC-2	Unrelated 1	Unrelated 2
S-1	>**0.99999**	<0.00001	–	–	<0.00001	–	–
HS-1	<0.00001	0.0002	–	–	–	<0.00001	–
S-2	<0.00001	**0.9998**	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	–	<0.00001	–	–	–	<0.00001	–
S-3	–	<0.00001	>**0.99999**	–	–	0.0017	–
PC-2	–	–	–	–	**0.99998**	–	–
S-4	–	–	–	**0.9991**	–	–	0.36
Unrelated	<0.00001	<0.00001	<0.00001	0.0009	<0.00001	**0.9983**	**0.64**

The true relationships in the first row and corresponding probability in bold. A hyphen in a specific row means exclusion of the relationship as an alternative hypothesis. S-1 means full siblings, HS-1 half-siblings, S-2 full 1st cousins, HS-2 half 1st cousins, S-3 full 2nd cousins, PC-2 grandparent–grandchild relation and S-4 means full 3rd cousins. The same 5774 markers have been used in all comparisons. Due to the varying availability of pairwise *true* relationships (Supplemental Table S1), the number of examples included for each relationship varies; For S-1 five comparisons, S-2 ten comparisons, S-3 four comparisons, S-4 ten comparisons, PC-2 nine comparisons, Unrelated ten comparisons.

Please cite this article in press as: D. Kling, et al., DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects, Forensic Sci. Int. Genet. (2011), doi:10.1016/j.fsigen.2011.07.007

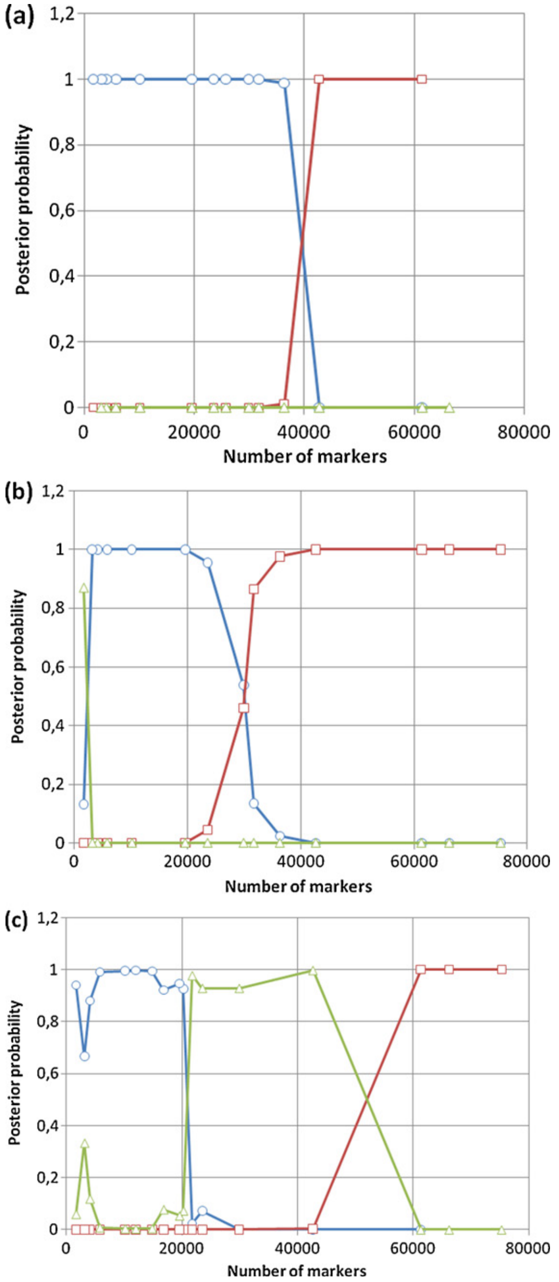

Fig. 3. (a)-(c). Graphs displaying the posterior probability for each hypothesis against the number of markers. (a) True relationship full 1st cousins (blue line) versus alternative hypotheses of full siblings (red line) and unrelated (green line). (b) True relationship full 2nd cousins (blue line) versus alternative hypotheses of full 1st cousins (red line) and unrelated (green line). (c) True relationship Unrelated (blue line) versus alternative hypotheses of full 2nd cousins (green line) and full 1st cousins (red line). The upper threshold value, when the true relationship no longer receives the highest posterior probabilities seems to be, for full 1st cousins: ~35,000 markers (markers separated by 0.05 cM), for full 2nd cousins: ~20,000 markers (markers separated by 0.1 cM), and for unrelated ~20,000 markers (markers separated by 0.1 cM).

4. Discussion

DNA has proven to be the most important tool to evaluate genetic relationships, both in forensic casework [32–34] and in medical research [1–3]. During the last decade mtDNA and gonosomal (X, Y) markers have been used to establish relatedness when lineages of maternal or paternal inheritance can be followed [35,36]. However, as soon as a line of inheritance is broken the genetic analyst loses track. Using thousands of autosomal SNP markers we showed that distant relationships could be established where the above-mentioned methods did not prevail. Although too early to draw any definite guidelines or conclusions, we believe the methods proposed in this study can be applied whenever complex family relations need to be resolved, as in for example genealogy studies.

The robustness of the tests was shown by using different sets of markers. The marker selection was based on a set of criteria that each chosen SNP-marker had to fulfil. Different minor allele frequencies did not appear to influence the results notably, though the issue was not extensively investigated. The distance between the markers did, however, show more impact on the results; especially when the number of markers exceeded 20,000, which is approximately equal to a distance of 0.1 cM between each pair of markers. It was apparent that a too dense selection of markers rendered the genetically closest relationship as the most probable, see Fig. 3(a)–(c). This phenomenon is, most likely, a consequence of linkage disequilibrium, which is also evident in Table 5 where more dense selections of markers yield a greater percentage of high r^2 values, but also, interestingly, an exponential increase in the number of haploblocks, see Fig. 4. Our simulations also further corroborates these results, see Table 6, where there obviously is a shift towards the closest relationship as more markers are included in the simulations. One of the reasons to why the results favour the closest relationship can possibly be explained by the “random” sharing of uncommon alleles. According to this admittedly speculative conjecture, a dense selection of markers amplifies the effect, as the uncommon alleles can possibly be in LD with other closely located uncommon alleles.

The Lander–Green algorithm, used to calculate the likelihoods, assumes the markers to be in LE and the likelihood computation collapses using many markers that are in LD. One reason to why the calculations fail is the large difference between the observed and the expected haplotype frequencies when dense sets of markers are used. Moreover, unrelated individuals will share certain haplotypes, as mentioned previously, due to a common ancestry, although further back, and they will appear related, i.e. false positives will arise [37,38]. In 2008 Kurbasic and Hossjer presented an extension to the Lander–Green algorithm in order to account for linkage disequilibrium [39]. They combined the Markov chain for inheritance vectors (i.e. Lander–Green) with another Lth order Markov chain that models LD structure. In this extension, the Markov chain contains information about the genotypes of the pedigree founders of L consecutive located loci. Kurbasic and Hossjer applied their method on a smaller simulation study (L = 1) and pointed out that the method is very computationally intensive unless the pedigrees are small and L is small. This limitation was also shown when the algorithm was implemented with a small number of forensically relevant STR markers [40]. Using a combination of kinship coefficient and IBS statistics, Manichaikul et al. recently presented a software, KING, which allows pairwise comparisons to be conducted on large sample material [41]. The authors claim the problem with LD is circumvented based on *large sample theory*. The KING software calculates a kinship coefficient, i.e. a rough estimate of an abstract family relationship, and not a forensically relevant probability value for a given pedigree hypothesis. We used the software on our material and the performance is comparable with our methods, for relationships closer than 3rd cousin. Using KING, 3rd cousin relationships could not be readily resolved. In addition, KING does not provide an answer to our main problem, determining the most likely pedigree.

As for Merlin/FEST, true relationships as distant as 3rd cousins could be distinguished with satisfactory posterior probabilities, using 12,453 markers. Unfortunately, inclusion of more distant relationships, e.g. 3rd cousins, as an alternative hypothesis when comparing two truly unrelated individuals, yields unsatisfactory probabilities, such as only a 64% posterior probability in favour of

Please cite this article in press as: D. Kling, et al., DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects, Forensic Sci. Int. Genet. (2011), doi:10.1016/j.fsigen.2011.07.007

Table 3
Posterior probabilities for the hypothesis of relationship between the two families, see Figs. 1 and 2, based on analyses using FEST.

Number of markers	19,518	12,453	10,144	5774	4074	3151
Comparison 1						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.024	8.5e−6	0.0002	0.0001	7e−5	0.00086
Unrelated	0.975	0.99999	0.9998	0.9999	0.9999	0.999
Comparison 2						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.0007	6e−6	5.1e−5	0.003	0.00085	0.006
Unrelated	0.999	0.9999	0.9999	0.997	0.999	0.994

Posterior probabilities for the included hypotheses. S-2 means full 1st cousins, HS-2 means half 1st cousins, see text for further details. Each value represents a posterior probability for a given selection of markers, see column header. Comparison 1 and 2, represents two separate tests to whether the two families in Figs. 1 and 2 are related.

Table 4
Posterior probabilities for the hypothesis of relationship between the two families, see Figs. 1 and 2, based on analyses using Merlin. A Bayesian approach with flat priors has been used.

Number of markers	19,518	12,453	10,144	5774	4074	3151
Comparison 1						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.024	8.5e−6	0.0002	0.0001	7e−5	0.00086
Unrelated	0.975	0.99999	0.9998	0.9999	0.9999	0.999
Comparison 2						
S-2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
HS-2	0.0007	6e−6	5.1e−5	0.003	0.00085	0.006
Unrelated	0.999	0.9999	0.9999	0.997	0.999	0.994

The question was whether the two families were sharing a common paternal ancestor two generations back. Data was included from all individuals in the third generations of the two families. For each minor allele frequency, two different distances between two neighbouring markers has been tested, see column headings.

Table 5
Evaluation of linkage disequilibrium. The table describes the proportion of pairwise comparisons with a r^2-value above each limit. In addition the number of haploblocks in each selection has been calculated using PLINK. Limitcm stands for the minimum genetic distance between two markers in the selection.

	Proportion of pairwise-SNP with r^2 higher than r^2 limit					
	limitcm0.5	limitcm0.25	limitcm0.15	limitcm0.1	limitcm0.075	limitcm0.05
Number of markers	5865	10,227	14,869	19,420	23,263	29,277
Number of pair-wise comparisons	471,704	903,536	1,363,046	1,813,620	2,194,050	2,789,420
r^2 limit						
0.1	0.0168	0.0176	0.0568	0.0418	0.0225	0.0615
0.2	0.0020	0.0022	0.0102	0.0085	0.0053	0.0178
0.3	0.0015	0.0014	0.0073	0.0060	0.0038	0.0132
0.5	0.0014	0.0012	0.0057	0.0042	0.0026	0.0090
0.8	0.0014	0.0010	0.0044	0.0026	0.0014	0.0047
Number of haploblocks[a]	4	80	355	824	1482	2896

[a] Estimated in PLINK.

Table 6
Averaged posterior probabilities for simulated relationships. The table describes averaged posterior probabilities (with standard deviations in parentheses) from 5000 simulations of genotype data on chromosome 22. Markers are assumed to be in LD and are evenly spaced over the chromosome. Prior probabilities are equal to 1/3.

Number of markers on chr 22	Distance (cM) between markers	Number of markers if extended to all chromosomes	HS-2	S-2	Unrelated
1		45[a]	0.3333 (0.0000)	0.3335 (0.0002)	0.3332 (0.0002)
10	8.778	453	0.3333 (0.0000)	0.3339 (0.0005)	0.3327 (0.0005)
100	0.798	4535	0.3334 (0.0001)	0.3345 (0.0012)	0.3321 (0.0012)
200	0.397	9070	0.3332 (0.0001)	0.3355 (0.0013)	0.3312 (0.0013)
500	0.158	22,675	0.3337 (0.0002)	0.3445 (0.0015)	0.3218 (0.0015)
1000	0.079	45,349	0.3335 (0.0002)	0.3592 (0.0015)	0.3073 (0.0016)
1500	0.053	68,024	0.3338 (0.0003)	0.3927 (0.0015)	0.2735 (0.0016)

[a] Due to the variation in genetic length of different chromosomes, the number is not 22.

the true hypothesis, i.e. *unrelated*, see Table 2. This value is certainly not convincing in forensic genetics, nor should it be in medical genetic research. We applied our findings, based on tests using data from known relationships, on two families concerning a common paternal ancestor two generations back. The results from Merlin and FEST were unambiguous and showed that the two families did not share a common ancestor and thus, according to our findings, are unrelated.

A 2nd cousin relationship appears to be the limitation to what can be determined with current methods, or by any means presently available. It is debatable what the term unrelated really stands for [42]. The genetic material is quickly diluted as each

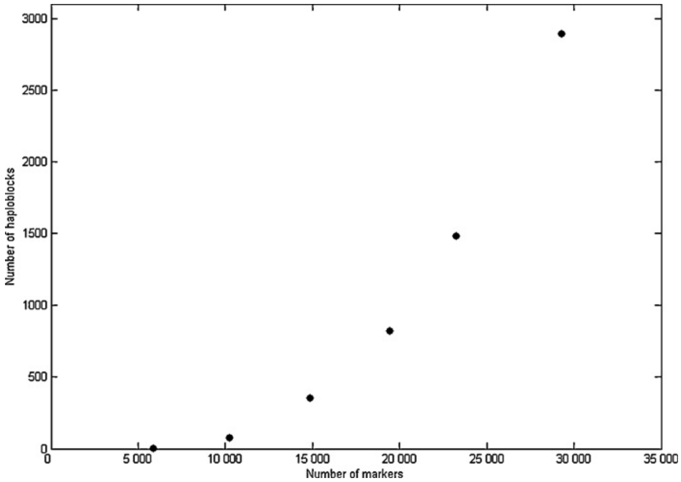

Fig. 4. Graph displaying the number of haploblocks (y-axis) versus the number of markers (x-axis). The number of haploblocks for each selection of markers has been calculated using PLINK. The graph displays an approximate exponential relationship between the number of haploblocks and the number of markers. The dot at 5865 markers corresponds to a distance of at least 0.5 cM between two markers, while the dot at 29,277 markers corresponds to a distance of at least 0.05 cM, see also Table 5.

generation passes. Perhaps the average background relatedness, shared by all individuals of the same ethnicity, lies not very far from the 3rd cousin relationship. Indeed the latest release from the HapMap project demonstrates that two unrelated individuals in the CEU population share in average 0.34% of their alleles through identity by descent (IBD) [12]. This is in fact approximately equal to the expected sharing of alleles (IBD) between two 3rd cousins. To investigate this further, more families need to be analyzed, where relationships such as half siblings, half cousins and half 2nd cousins are known. Simulation studies can be performed but they are more complicated since they raise the issue of how to model and account for linkage disequilibrium. For example, relationships can be simulated based on true haplotypes, where the issue of how to model LD in simulations is irrelevant, but haplotypes are complicated and computer demanding to infer. PHASE and IMPUTE, as well as similar available software, offer the advantage of inferring haplotypes from genotyping data, without any family or pedigree information [43,44]. We simulated relationships where instead the founder haplotypes were created using an approximate LD map from the HapMap project and the results agreed with our previous findings.

Regarding the statistical calculation, we suggest creating a new model, or modifying an existing one, which accounts for linkage disequilibrium. LD might be turned into an advantage if a proper model is developed. Moreover, other algorithms should be considered, i.e. other than Lander–Green, which is used in Merlin. Indeed, algorithms that can handle large and complex pedigrees with a large number of markers should be evaluated. For large and complex pedigrees, with thousands of markers, approximate approaches, such as Monte-Carlo Markov chain (MCMC), utilized in the software MORGAN for example, might be a good candidate. [15,45]. The existence of block-like structures, with clusters of tightly linked SNPs may also prove useful [26,46]. Merlin provides the possibility to calculate likelihoods based on specified cluster information [47]. Although theoretically promising the current implementation of the method in Merlin was, in our study, unable to handle more extended pedigrees with an average amount of clusters, i.e. 3rd cousins and 5000 clusters.

There are in addition alternative methods for the determination of the most probable relationship between individuals. One such approach is utilizing identity by state (IBS). This approach may not be optimal from a statistical point of view, but can nevertheless be useful to illustrate distant relationships [48,49].

In conclusion, genotype data from high-density SNP arrays have proved to be useful in the investigation of distant genetic relationships. In this study we solved a real case of half 1st cousinship using different selections of SNP markers. Relationships as distant as 2nd cousins could also be unambiguously resolved. However, 3rd cousins and more distant relationships revealed hard to distinguish from unrelated. Nevertheless, this task should not be insurmountable using a good computer algorithm and enough reference material to work with. Parameters such as genotyping errors and LD should be more thoroughly investigated as well as IBS approaches. Our conclusions regarding the relation between the two families (Figs. 1 and 2) are primarily based on a small number of established relationships (Tables 1 and 2) and thus further simulations and families are needed to verify our results. Even so, we are confident that our methods can be used to solve other cases of disputed distant family relationships.

Acknowledgements

Thanks to Prof. Peter Söderkvist, Department of Clinical and Experimental Medicine, Linköping University, for kindly providing the population reference material and to Master Student Iftikhar Ahmad, Department of Mathematical Science, Chalmers University, for the work on genotyping errors. We would also like to thank the two families, without whom this study would not have been possible.

Disclaimer: The study was funded by the National Board of Forensic Medicine, Sweden, Grant No. 4155225; Center of Forensic Science, Linköping University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2011.07.007.

References

[1] D.L. Newman, M. Abney, M.S. McPeek, et al., The importance of genealogy in determining genetic associations with complex traits, Am. J. Hum. Genet. 69 (2001) 1146–1148.

[2] B.F. Voight, J.K. Pritchard, Confounding from cryptic relatedness in case-control association studies, PLoS Genet. 1 (2005) e32.

[3] A.L. Leutenegger, E. Genin, E.A. Thompson, et al., Impact of parental relationships in maximum lod score affected sib-pair method, Genet. Epidemiol. 23 (2002) 413–425.

[4] A.O. Karlsson, G. Holmlund, T. Egeland, et al., DNA-testing for immigration cases: the risk of erroneous conclusions, Forensic Sci. Int. Genet. 172 (2007) 144–149.

[5] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435–445.

[6] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2006) 253–265.

[7] B. Budowle, A. van Daal, Forensically relevant SNP classes, Biotechniques 44 (603–608) (2008) 610.

[8] J.C. Glaubitz, O.E. Rhodes, J.A. Dewoody, Prospects for inferring pairwise relationships with single nucleotide polymorphisms, Mol. Ecol. 12 (2003) 1039–1047.

[9] J.J. Sanchez, C. Phillips, C. Borsting, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.

[10] C. Borsting, J.J. Sanchez, H.E. Hansen, et al., Performance of the SNPforID 52 SNP-plex assay in paternity testing, Forensic Sci. Int. Genet. 2 (2008) 292–300.

[11] D.E. Reich, S.F. Schaffner, M.J. Daly, et al., Human genome sequence variation and the influence of gene history, mutation and recombination, Nat. Genet. 32 (2002) 135–142.

[12] K.A. Frazer, D.G. Ballinger, D.R. Cox, et al., A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–861.

[13] S.F. Grant, H. Hakonarson, Microarray technology and applications in the arena of genome-wide association, Clin. Chem. 54 (2008) 1116–1124.

[14] M. Nothnagel, J. Schmidtke, M. Krawczak, Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci, Int. J. Legal Med. 124 (2010) 205–215.

Please cite this article in press as: D. Kling, et al., DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects, Forensic Sci. Int. Genet. (2011), doi:10.1016/j.fsigen.2011.07.007

[15] G. Gao, D.B. Allison, I. Hoeschele, Haplotyping methods for pedigrees, Hum. Hered. 67 (2009) 248–266.

[16] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523–542.

[17] T. Egeland, P.F. Mostad, B. Mevag, et al., Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. Genet. 110 (2000) 47–59.

[18] E.S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 2363–2367.

[19] G.R. Abecasis, S.S. Cherny, W.O. Cookson, et al., Merlin—rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.

[20] O. Skare, N. Sheehan, T. Egeland, Identification of distant family relationships, Bioinformatics 25 (2009) 2376–2382.

[21] T. Egeland, N. Sheehan, On identification problems requiring linked autosomal markers, Forensic Sci. Int. Genet. 2 (2008) 219–225.

[22] S. Purcell, B. Neale, K. Todd-Brown, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.

[23] G. Montana, HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients, Bioinformatics 21 (2005) 4309–4311.

[24] Rutgers Combined Linkage-Physical Map of The Human Genome, http://compgen.rutgers.edu/RutgersMap/DownloadMap.aspx (accessed 18.05.11).

[25] B. Lindblom, G. Holmlund, Rapid DNA purification for restriction fragment length polymorphism analysis, Gene Anal. Tech. 5 (1988) 97–101.

[26] J. Ge, B. Budowle, J.V. Planz, et al., Haplotype block: a new type of forensic DNA markers, Int. J. Legal Med. 124 (2010) 353–361.

[27] S.B. Gabriel, S.F. Schaffner, H. Nguyen, et al., The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.

[28] M. Krawczak, Informativity assessment for biallelic single nucleotide polymorphisms, Electrophoresis 20 (1999) 1676–1681.

[29] F. Pompanon, A. Bonin, E. Bellemain, et al., Genotyping errors: causes, consequences and solutions, Nat. Rev. Genet. 6 (2005) 847–859.

[30] E. Sobel, J.C. Papp, K. Lange, Detection and integration of genotyping errors in statistical genetics, Am. J. Hum. Genet. 70 (2002) 496–508.

[31] M.P. Epstein, W.L. Duren, M. Boehnke, Improved inference of relationship for pairs of individuals, Am. J. Hum. Genet. 67 (2000) 1219–1231.

[32] D.W. Gjertson, C.H. Brenner, M.P. Baur, et al., ISFG: recommendations on biostatistics in paternity testing, Forensic Sci. Int. Genet. 1 (2007) 223–231.

[33] M. Tracey, Short tandem repeat-based identification of individuals and parents, Croat. Med. J. 42 (2001) 233–238.

[34] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

[35] J. Ge, A. Eisenberg, J. Yan, et al., Pedigree likelihood ratio for lineage markers, Int. J. Legal Med. 125 (2011) 519–525.

[36] R. Szibor, X-chromosomal markers: past, present and future, Forensic Sci. Int. Genet. 1 (2007) 93–99.

[37] Q. Huang, S. Shete, C.I. Amos, Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis, Am. J. Hum. Genet. 75 (2004) 1106–1112.

[38] J.M. Keith, A. McRae, D. Duffy, et al., Calculation of IBD probabilities with dense SNP or sequence data, Genet. Epidemiol. 32 (2008) 513–519.

[39] A. Kurbasic, O. Hossjer, A general method for linkage disequilibrium correction for multipoint linkage and association, Genet. Epidemiol. 32 (2008) 647–657.

[40] A.O. Tillmar, T. Egeland, B. Lindblom, et al., Using X-chromosomal markers in relationship testing: calculation of likelihood ratios taking both linkage and linkage disequilibrium into account, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.11.004.

[41] A. Manichaikul, J.C. Mychaleckyj, S.S. Rich, et al., Robust relationship inference in genome-wide association studies, Bioinformatics 26 (2010) 2867–2873.

[42] B.S. Weir, A.D. Anderson, A.B. Hepler, Genetic relatedness analysis: modern data and new challenges, Nat. Rev. Genet. 7 (2006) 771–780.

[43] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978–989.

[44] J. Marchini, B. Howie, S. Myers, et al., A new multipoint method for genome-wide association studies by imputation of genotypes, Nat. Genet. 39 (2007) 906–913.

[45] A.W. George, E.M. Wijsman, E.A. Thompson, MCMC multilocus lod scores: application of a new approach, Hum. Hered. 59 (2005) 98–108.

[46] K. Zhang, P. Calabrese, M. Nordborg, et al., Haplotype block structure and its applications to association studies: power and study designs, Am. J. Hum. Genet. 71 (2002) 1386–1394.

[47] G.R. Abecasis, J.E. Wigginton, Handling marker–marker linkage disequilibrium: pedigree analysis with clustered markers, Am. J. Hum. Genet. 77 (2005) 754–767.

[48] H. Miyazawa, M. Kato, T. Awata, et al., Homozygosity haplotype allows a genome-wide search for the autosomal segments shared among patients, Am. J. Hum. Genet. 80 (2007) 1090–1102.

[49] E.D. Roberson, J. Pevsner, Visualization of shared genomic regions and meiotic recombination in high-density SNP data, PLoS One 4 (2009) e6711.