European Journal of Human Genetics (2006) 14, 450–458

www.nature.com/ejhg

ARTICLE

Identification of probable genotyping errors by consideration of haplotypes

Tim Becker[1], Ruta Valentonyte[2,3], Peter JP Croucher[4,5], Konstantin Strauch[6], Stefan Schreiber[2,3], Jochen Hampe[2,3] and Michael Knapp*[,1]

[1]*Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany;* [2]*Institute for Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany;* [3]*University Hospital Schleswig-Holstein Campus Kiel, Schittenhelmstr. 12, Kiel, Germany;* [4]*Institute for Medical Informatics and Statistics, Christian-Alb rechts- University, Kiel, Germany;* [5]*University Hospital Schleswig-Holstein Campus Kiel, Brunswikerstr. 10, Kiel, Germany;* [6]*Institute for Medical Biometry and Epidemiology, Philipps-University Marburg, Marburg, Germany*

Undetected genotyping errors pose a problem in genetic epidemiological studies, as they may invalidate statistical analysis or reduce its power. Haplotype analysis requires an improved standard of the data, because a haplotype can be inferred correctly only if the genotypes of all its markers are correct. Here, we present a method that identifies probable genotyping errors in trio samples with the help of the estimated haplotype frequency distribution of the sample. If the likelihood of the most likely haplotype explanation depends strongly on just one genotype, in the sense that setting the genotype to be missing leads to a much more likely haplotype explanation, this genotype is considered as a potential genotyping error. We describe a method that systematically searches the whole data set for such potential errors. Based on the haplotype distribution of a real data set, we carry out a simulation study to estimate the sensitivity and specifity of the method. In addition, we apply our approach to the real data set itself. Potentially erroneous genotypes are re-determined via sequencing. The results of both the simulation study and of the application to the real data set show that a considerable proportion of true genotyping errors is detected and that the number of false-positive signals is acceptable. We conclude that it is indeed possible to identify probable genotyping errors by considering haplotypes. The method described here will be part of the next release of our FAMHAP software.
European Journal of Human Genetics (2006) **14,** 450–458. doi:10.1038/sj.ejhg.5201565; published online 25 January 2006

Keywords: genotype error; haplotype; frequency estimation

Introduction

Errors are inevitable when genotypes of many single-nucleotide polymorphisms (SNPs) in a large sample of individuals are produced by current high-throughput technologies. Various tools are available for identification of problematic genotypes. The applicability of these tools depends on the data structure of the sample. In population-based samples of unrelated individuals, the genotype distribution of each SNP can be tested for its deviation from the Hardy–Weinberg equilibrium (HWE). Such a deviation is indicative of problematic assays,[1] but does not identify individual genotyping errors. In cases of family-based data, genotyping errors can be detected by Mendelian inconsistencies (MIs); for example, when an allele in a child does not occur in either parent. However, only a fraction of the genotyping errors become apparent by such MIs. The detection rate depends on the marker allele frequency, the pedigree structure, and the error model.[2,3] For nuclear

***Correspondence: Dr M Knapp, Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Str. 25, D-53105 Bonn, Germany. Tel: +49 228 287 5810; Fax: +49 228 287 5854; E-mail: knapp@uni-bonn.de**
Received 7 July 2005; revised 6 October 2005; accepted 24 November 2005; published online 25 January 2006

families with a single affected child in which genotyping errors are assumed to occur independently with some fixed probability ε in each allele, Gordon *et al*[2] calculated that the detection rate is between 25 and 30% for diallelic markers such as SNPs. If there are at least two children in a family, genotyping errors can become visible as a recombination between tightly linked loci.[4–6]

The importance of the topic of genotyping errors stems from the observation that undetected errors may invalidate the statistical analysis. In the context of association analysis, non-differential genotyping errors decrease the power of case–control studies,[7] but inflate the type I error probability of the transmission disequilibrium test (TDT) for family-based association analysis.[8,9] It can be assumed that the effect of undetected genotype errors is exaggerated when haplotypes instead of single-marker loci are analyzed. Indeed, the integrity of a haplotype comprising m loci requires perfect allele ascertainment for each of the m loci. Therefore, genotyping error rates that are tolerable in the context of single-marker analysis can be unacceptable for analyses based on haplotypes. For example, Knapp and Becker[10] recently observed a dramatic inflation of the type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT) by Zhang *et al*[11] in the presence of genotyping error rates as small as 0.001.

As haplotype analysis requires a higher standard of data integrity, it would be valuable to develop methods for the identification of probable genotyping errors that exploit the kind of data typically available. Such a method is presented and evaluated in the present paper. From a sample of nuclear families that are genotyped for a set of tightly linked markers, maximum likelihood haplotype frequency estimates can be obtained.[12] Then, the likelihood of the most probable haplotype explanation can be calculated for each family. If it turns out that this likelihood critically depends on the genotype of a single marker, that is, a much more likely haplotype explanation would be enabled by replacing a single marker genotype, then this genotype is considered to represent a probable genotyping error. The details of our approach are described in the next section. We describe a simulation study, which is based on the haplotype distribution of a real data set. The simulation study is used to determine parameter configurations that lead to 'optimal' error detection. Finally, we apply our approach to the real data set itself in order to evaluate its performance. Genotypes that yield strong statistical evidence for errors are re-determined in a sequencing experiment in order to check the reliability of the error prediction.

Methods

Notation

Assume that a family consisting of two parents and a single child has been genotyped for a set B of markers. Let $G_B = (G_B^f, G_B^m, G_B^c)$ denote the unphased multi-locus genotypes of the family, with G_B^f, G_B^m, and G_B^c denoting the unphased multi-locus genotypes for father, mother, and child, respectively. Genotypes are allowed to be missing for some markers in some individuals, that is, it is not assumed that genotype information is complete. We use $r|s = G_B^\star$ to denote that the B-haplotypes r and s are compatible with the unphased multi-locus genotypes $G_B^\star$ of an individual. Then, an ordered quadruple (k,l,u,v) of B-haplotypes is called a haplotype explanation for G_B if $k|l = G_B^f$, $u|v = G_B^m$, and $k|u = G_B^c$. The set

$$C_{G_B} = \{(k,l,u,v) : k|l = G_B^f, u|v = G_B^m, k|u = G_B^c\}$$

is the set of possible haplotype explanations for G_B. (Note that this formulation implicitly assumes that no recombination occurs between marker loci.) With h_k denoting the frequency of haplotype k, it follows that

$$L_{G_B} = \max_{(k,l,u,v)\in C_{G_B}} h_k \cdot h_l \cdot h_u \cdot h_v \qquad (1)$$

is the likelihood of the most probable haplotype explanation for G_B.

Identification of potential genotyping errors

The basic idea of our approach for identification of genotyping errors is that the presence of a genotype error at a specific marker locus in a specific individual may become prominent through a sharp decrease of the likelihood of the most probable haplotype explanation for the family. Therefore, we modify the observed unphased multi-locus genotypes of the family by discarding (ie, setting to 'unknown') a single-marker genotype in an individual. Then, the likelihood of the most probable haplotype explanation for the modified data is calculated and compared to the corresponding likelihood of the observed data. A sharp increase of the likelihood points to a potential genotyping error.

More precisely, let $G_B^{l,j}$ denote the resulting unphased multi-locus genotypes of the family after the genotype at marker locus l in individual j ($j=1$: father; $j=2$: mother; $j=3$: child) has been discarded. As $C_{G_B} \subset C_{G_B^{l,j}}$, it follows that $L_{G_B} \le L_{G_B^{l,j}}$. If $L_{G_B^{l,j}}/L_{G_B} \ge c$, then we say that marker locus l in individual j provides a signal at threshold c for marker set B. As the frequency h_k of haplotype k is generally not known, haplotype frequencies have to be estimated. Provided that a sample of families is available, this estimation can be achieved by the program FAMHAP,[12] which estimates haplotype frequencies from samples consisting of general nuclear families. These estimates are then inserted into (1) to obtain L_{G_B} and $L_{G_B^{l,j}}$. It may happen that, owing to computer memory constraints, FAMHAP has to discard some families during the process of haplotype frequency estimation because they have too many haplotype explanations. For such families it is possible that no valide haplotype explanation can be found when its original genotypes are considered, but that

a valide haplotype explanation exists when a marker is set to be missing. In this situation we speak of a signal at threshold ∞.

Up to now, the set B of markers has been assumed to be fixed. In practice, genotypes of several hundreds of markers may have been obtained for the family sample. In this situation, it is neither technically feasible nor desirable to consider haplotypes consisting of all available loci. Thus, the question is how to select B for which $L_{G_B^{l,j}}/L_{G_B}$ has to be calculated. A natural way to deal with this problem is to divide the whole set of markers into haplotype blocks by any of the available algorithms[13] for this purpose. For examination of marker locus l, B may then be defined as all marker loci belonging to the block that contains l. After some experimentation with this approach, however, we focused on an alternative that more directly takes into account the following requirements: (i) for computational reasons, the number t of B-haplotypes with non-zero frequency should not be too large. Indeed, a naive algorithm for calculating the likelihood of the most probable haplotype explanation for G_B would have to check t^4 haplotype combinations for their compatibility with G_B. Although we employ an algorithm that generally results in a considerable reduction of this computational burden (details are given in the Appendix A), running time still presents a problem for large t; (ii) the number of loci in B should not be too small. Intuitively, it seems more impressive that the deletion of a single-marker genotype in an individual enables a much more likely haplotype explanation when haplotypes consisting of several loci are considered than when the haplotypes comprise only a very small number of loci; (iii) B should consist of neighboring loci.

In view of (i–iii), we consider the following strategy to identify individuals with potential genotyping errors at marker locus l, which is directed by two pre-defined constants maxhap and minloc. If the families have been genotyped for physically ordered marker loci $1, \ldots, m$, the set B_l of marker combinations consists of all subsets $B \subset \{1, \ldots, m\}$ satisfying (*i*) $t \leq$ maxhap for the number t of B-haplotypes with non-zero frequency; (*ii*) $|B| \geq$ minloc; and (*iii*) B is a connective subset of $\{1, \ldots, m\}$ that contains l. Then, the ratio $L_{G_B^{l,j}}/L_{G_B}$ is calculated for all $B \in B_l$ and we say that the strategy (c, maxhap, minloc) provides a signal for marker locus l in individual j at threshold c if

$$\max_{B \in B_l} \frac{L_{G_B^{l,j}}}{L_{G_B}} \geq c. \qquad (2)$$

Such a signal is a true signal when the marker genotype at locus l in individual j is really erroneous, otherwise it is a false signal. The sensitivity of the strategy (c, maxhap, minloc) for individuals of type j (ie, fathers, mothers, or children) is the fraction of genotype errors at any locus l in individuals of type j that results in a signal at this locus. The predictive value of a signal by strategy (c, maxhap, minloc) for individuals of type j is the fraction of all signals in individuals of type j that are true signals.

Instead of discarding the genotype at marker locus l in a single individual j, genotypes at marker locus l can be set to unknown in all members of the family. We denote the resulting unphased multi-locus genotypes of the family by $G_B^{l,0}$ and the likelihood of the most probable haplotype explanation by $L_{G_B^{l,0}}$. Then, a signal by strategy (c, maxhap, minloc) is defined analogously to (2). Note, however, that such a signal does not identify the genotype of a specific individual, but rather points to potential genotyping error(s) at locus l in the family. Therefore, a signal is a true signal when the marker genotype at locus l in at least one individual of the family is erroneous. The sensitivity of the strategy is the fraction of family/marker combinations with at least one genotype error resulting in a signal. The predictive value of a signal is the fraction of all signals which are true signals.

Real data set

The real data sample consisted of 659 family triads, that is, parents and a single child. Genotypes were obtained for 35 SNPs (listed in Supplementary Table S1) on chromosome 16 covering a region of 91 391 bases. In brief, genotypes of SNPs were obtained with the use of TaqMan™ technology (Applied Biosystems, Foster City, CA, USA). Probes and primers were either designed with the use of Primer Express software (Version 2.0.0, Applied Biosystems, Foster City, CA, USA) or obtained as ready-to-use assays from the Applied Biosystems store (www.appliedbiosystems.com). Fluorescence results were analyzed on an ABI Prism™-7900HT Sequence Detector System (384-well format) or ABI Prism™7700 Sequence Detector (96-well format, both from Applied Biosystems, Foster City, CA, USA). Genotype errors resulting in non-Mendelian segregation of markers in pedigrees were excluded from analysis. In 108 families, the total number of available genotypes for the family members was below 59. As the running time of the simulation study described below is negatively influenced by the presence of families with a large fraction of missing genotypes, these families were discarded. Thereafter, the data set consisted of 551 triads, 35 SNPs and 56 388 available genotypes. The percentage of available genotypes for an SNP varied between 94.7 and 99.5%.

Table 1 provides the frequency of the rarer allele for each SNP.

Simulation study

For the real data set, the frequencies of 35 marker haplotypes were estimated by the program FAMHAP.[12] FAMHAP identified 487 haplotypes with a frequency >0. The estimated frequency of 478 of these haplotypes was below 1%. The remaining nine haplotypes, which account for 90.5% of haplotypes in the sample, are listed in Table 2. This estimated haplotype distribution with 487 different

Table 1 Minor allele frequencies

Marker	*Frequency*	*Marker*	*Frequency*	*Marker*	*Frequency*	*Marker*	*Frequency*
SNP1	0.062	SNP10	0.214	SNP19	0.431	SNP28	0.134
SNP2	0.449	SNP11	0.431	SNP20	0.455	SNP29	0.212
SNP3	0.453	SNP12	0.433	SNP21	0.437	SNP30	0.209
SNP4	0.435	SNP13	0.022	SNP22	0.221	SNP31	0.126
SNP5	0.431	SNP14	0.443	SNP23	0.232	SNP32	0.118
SNP6	0.445	SNP15	0.217	SNP24	0.241	SNP33	0.126
SNP7	0.432	SNP16	0.021	SNP25	0.222	SNP34	0.022
SNP8	0.452	SNP17	0.022	SNP26	0.426	SNP35	0.030
SNP9	0.457	SNP18	0.437	SNP27	0.135		

Table 2 Haplotypes occurring in the real data set with a frequency of more than 1%

Haplotype	*Frequency (%)*
221112221112221221212212121111111112	31.9
212221112221211222121222111121111112	19.1
221112221112221221212212111112111112	11.1
212221112121212222121111212211222l2	10.1
212221112121212222121111211111211112	8.4
121112221112221221212212121111111112	5.9
212221112121212222121111212211111112	1.7
212112221112221221212212121111111112	1.2
221111212112111111222221111111111121	1.1

haplotypes was used to create a simulated replicate of the data as follows. Firstly, all parents in all families were randomly assigned two haplotypes according to this distribution. For determination of the alleles in the child being transmitted by one parent, it was assumed that recombinations between neighboring pairs of loci occur independently and that a physical distance of 1 Mb between two marker loci corresponds to a genetic distance of $\theta = 0.01$ between the two loci. Secondly, marker genotypes of individuals who were missing in the real data set were set to unknown. Thirdly, genotyping errors were introduced independently according to the stochastic error model, for which ε denotes the probability that, for each allele at each marker locus, the allele is changed. Note that the genotype of a heterozygous individual at a marker locus is altered only when exactly one of his or her alleles is switched. Such modified genotypes were recorded as a true typing error. In the fourth step, for families and marker loci with at least one true typing error, it was checked whether the typing error was evident as a Mendelian inconsistency. In this case, the genotypes of this marker were set to unknown in all individuals of the family. We simulated 100 replicated data sets in total, for which we determined the distribution of genotype errors and the sensitivity and predictive value of the proposed method under different constellations of the threshold c and the parameters maxhap and minloc.

Table 3 Number[a] of signals ($\varepsilon = 0$, maxhap = 100, minloc = 2)

	Number of signals $Log_{10}>$				
	2	*2.5*	*3*	*3.5*	*4*
Any	445	146	38	12	5
Parents	202	48	7	1	0
Child	30	11	5	3	1

[a]Averages based on 100 replicated data sets

Sequencing

For investigation of the potentially erroneous genotypes, a re-sequencing experiment of all three individuals from the trios with flagged genotypes was performed. Forward and reverse primer pairs (Supplementary Table S1) for amplification of the target sequences were designed with Primer Express (Version 2.0.0, Applied Biosystems, Foster City, CA, USA). Ten nanograms of DNA were amplified with AmpliTaq® Gold DNA polymerase (Applied Biosystems; Darmstadt, Germany) and purified using SAP/Exo1 digestion. The products were sequenced using the BigDye™ chemistry (Aplera Inc., Foster City, CA) according to the manufacturer's protocol. Sequence detection was performed with an automated, 96-capillary fluorescence detection system ABI Prism®3700DNA analyser (Aplera Inc., Foster City, CA). In order to avoid possible mistakes introduced by polymerases, sequences of both DNA strands were analysed using the Sequencher program (Version 4.0.5 Gene Codes Corporation, Ann Arbor, MI, USA) and compared to ensure non-ambiguous genotypes. Finally, the genotypes, as determined by sequencing, were compared to the corresponding genotypes obtained using the TaqMan™ method.

Results

Results of simulation study

Firstly, we investigated the behavior of our approach in the absence of genotyping errors. In Table 3, average numbers

of (false positive) signals are listed for simulations without errors (ie, $\varepsilon = 0$). The row headed 'any' refers to the case where all genotypes of the family are set to zero at a given marker, the row headed 'parents' refers to the case where only one parent is set to be missing, and the row labeled 'child' applies to the case where the offspring genotype is set to be missing. As our data set consists of 551 trios which are typed at 35 markers, there are 19 285 trio-genotype constellations which may yield a signal under 'any' strategy, if we ignore the small portion of missing genotypes. Thus, 445 signals at threshold 100 correspond to a false-positive rate of 2.3%. For the 'child' strategy, the number of false-positive signals is much smaller; 30 signals at threshold 100 correspond to a false-positive rate of only 0.2%. This much smaller rate may be explained by the fact that only one family member is set to be 'missing'. Hence, the set of possible haplotype explanations will not grow as much as with the 'any' strategy. Moreover, both haplotypes of a child's haplotype explanation have to be compatible with the parents, that is, they have already been checked for compatibility twice. With a growing threshold, however, the number of false-positive signals becomes negligible for all strategies. We conclude that, in general, our routine produces an acceptable number of false signals in error-free data sets. Next, we investigated the performance of our routine in the presence of errors. Table 4 shows the distribution of the genotype errors in the simulated data set for an allele-wise error rate of $\varepsilon = 0.01$. On average, 26.8% of the genotyping errors (300 out of 1121) became visible as Mendelian inconsistencies, which confirms the calculations of Gordon *et al*[2] mentioned previously, and 644 out of the 822 undetected genotyping errors (78.3%) are errors in the parental genotypes. This increased proportion results from the fact that genotyping errors in offspring genotypes manifest themselves as MIs more readily.[3] Table 5 shows the sensitivity and predictive value of our approach at $\varepsilon = 0.01$. Note that the values in Table 5 refer only to the fraction of the genotyping errors that are not detectable as MIs. The first part of the table contains results for the situation in which all marker windows with two or more markers (minloc = 2), for which no more than 100 haplotypes have an estimated frequency different from zero (maxhap = 100), were considered. With the 'any' strategy, there are, on average, 831 signals at threshold 100, which approximately equals the number of genotyping errors that are consistent with Mendelian inheritance. Of course, not all of the signals are true, but 52% of the genotyping errors become visible and 50% of the signals actually point to a genotyping error. Together with the errors which become prominent as Mis, the sensitivity of error detection is thus about 64.8% and its predictive value is about 63.3%. These figures confirm that it is useful to search for genotyping errors using haplotypes. With higher thresholds, the predictive value, with respect to the errors that are not detected via MIs, grows markedly to 87%. Unfortunately, sensitivity drops rather rapidly with higher thresholds; at threshold 10 000, for instance, only 5% of the genotyping errors are detected. The 'child' case, has, in general, higher sensitivity than the 'parents' case, and except at lower thresholds also has a very good predictive value. The low predictive value at threshold c = 100 with the 'child' strategy seems to contradict the observations made from the simulations without genotyping errors. It is true that haplotypes in the children are checked for their integrity twice, as they occur both in parents and children, while the non-transmitted

Table 4 Distribution of genotype errors in simulated data sets ($\varepsilon = 0.01$)

	Mean[a]	*Std*	*Min*	*Max*
Total errors	1121	31.5	1058	1204
Mendelian inconsistent	300	17.3	254	340
Mendelian consistent	822	23.7	763	879
Any	804	22.9	744	860
Parents	644	21.8	585	697
Child	178	11.8	153	217

[a]Averages based on 100 replicated data sets.

Table 5 Sensitivity and predictive value ($\varepsilon = 0.01$, maxhap = 100)

		Number of signals[a]					*Sensitivity*					*Predictive value*				
					$log_{10}>$				$log_{10}>$					$log_{10}>$		
Minloc		*2*	*2.5*	*3*	*3.5*	*4*	*2*	*2.5*	*3*	*3.5*	*4*	*2*	*2.5*	*3*	*3.5*	*4*
2	Any	831	496	257	146	49	0.52	0.29	0.23	0.16	0.05	0.50	0.47	0.73	0.88	0.87
2	Parents	308	49	33	26	21	0.32	0.05	0.03	0.03	0.03	0.66	0.59	0.68	0.78	0.85
2	Child	440	379	202	109	18	0.87	0.83	0.78	0.53	0.09	0.35	0.39	0.68	0.86	0.91
5	Any	564	300	212	134	43	0.49	0.26	0.22	0.15	0.05	0.69	0.70	0.84	0.92	0.92
5	Parents	281	43	30	25	20	0.30	0.04	0.03	0.03	0.03	0.69	0.66	0.74	0.82	0.88
5	Child	273	230	174	105	17	0.87	0.83	0.77	0.52	0.09	0.57	0.64	0.79	0.88	0.92

[a]Averages based on 100 replicated data sets.

haplotypes of the parents are checked for integrity only once[10] and that as a consequence, the chance of detect genotyping errors in the parents by MIs is reduced. On the other hand, if setting a child's genotype to 'missing' leads to additional haplotype explanations, it is likely that they produce a strong signal as all four haplotypes of the former most likely haplotype explanation may change. This phenomenon may explain the lower predictive value at threshold c = 100. However, as the 'child' strategy shows both high sensitivity and predictive value at threshold c = 1000, the performance at threshold c = 100 is not important. In the second part of Table 5, sensitivity and predictive value are computed for the same permutation replicates, but now only signals that stem from haplotypes which extend over at least five markers (minloc = 5) are considered. Independent of the considered thresholds and the individuals who are set to be missing, the predictive value improves markedly when compared to minloc = 2. This improvement is compromised only by a small reduction of the sensitivity. The data confirms our initial conjecture that signals that are found using haplotypes consisting of many markers may be more meaningful than those that are formed by few markers. Hence, the usage of a higher minloc is recommendable. We also considered the influence of the maxhap parameter on the performance of the method. It turns out that both sensitivity and predictive value are slightly reduced when lower values of maxhap are chosen (data not shown). Therefore, it is sensible to choose maxhap as high as running time requirements allow. Furthermore, we simulated data under the high error rate of $\varepsilon = 0.05$. In this situation, the portion of errors that become visible as MIs remained as before. While the predictive value was slightly better than under the more realistic (lower) error rate, sensitivity was greatly reduced (24% at threshold c = 100, 0% at threshold c = 10 000). This phenomenon can be explained as follows. The haplotype frequencies which are used to compute the signals are estimated from the data set and are influenced by the overall error rate of the data. If the genotyping quality of a data set is very low, then the true haplotype distribution is not represented adequately and it becomes more difficult to find potential genotyping errors with the help of the estimated haplotype frequencies.

Table 6 Real data: number of signals (maxhap = 100, minloc = 2)

	Number of signals $log_{10}>$				
	2	*2.5*	*3*	*3.5*	*4*
Any	466	227	111	63	41
Parents	294	102	35	20	11
Child	94	60	49	40	29

Application to real data set

Table 6 lists the number of signals for the real data set at five different thresholds with maxhap = 100 and minloc = 2. The values are higher than those observed in the simulations without errors (Table 3), but lower than those for the simulations with an error rate of $\varepsilon = 0.01$ (Table 5). We conclude that the rate of genotyping errors in the real data set is lower than $\varepsilon = 0.01$, but that the data set is not free from errors. Therefore, we decided to redetermine those genotypes that gave strong signals. All 41 signals that were obtained with the 'any' strategy at threshold 10 000 can be found in Table 7. The genotypes of the respective markers were re-determined for all members of the particular family using the described sequencing technique. The new genotypes were assumed to be the correct ones. The sequencing failed to generate six individual genotypes. The results of the sequencing experiment confirmed many of the predictions made by our method. There are five instances in which the opposite homozygote to that genotyped was predicted with this being confirmed to be the true genotype from the sequencing data (father of family 1284 and both parents of family 830 at two different loci). In total, 23 out of the 41 signals (about 56%) turned out to correspond to actual genotyping errors and, except for three instances (ID 1269-SNP 3, ID 522-SNP 26 and ID 1327-SNP 10), the true genotypes were those suggested by our routine. The predictive value for the real data set is somewhat below the values determined in the simulation study. However, in a further 13 instances, either the families' genotypes are completely correct, but the families actually have a genotyping error at another marker which also shows a signal, or the sequencing failed to produce results for some family members. Family 1083, for instance, shows an equally high signal (4.89e + 04) both for markers 2 and 3 when markers 1-4 are considered. The sequencing results demonstrate that marker 3 is correctly genotyped, but that marker 2 actually has a genotyping error. In this way, the signal for marker 3 is induced by a genotyping error in its vicinity. Five out of the 41 signals only are false positives in a narrow sense. Even in these five cases, it is possible that there are genotyping errors at other markers that did not produce signals above the threshold of 10 000 and which, therefore, were not checked in the sequencing experiment. This is indeed not unlikely, as the false-positive signals are all among those signals that did not reach the threshold of 100 000. In summary, we believe that the real data set proves that the prediction of genotyping errors works satisfactorily.

Discussion

We have developed a systematic search routine for potential genotyping errors, which relies on estimated haplotype frequencies and have implemented this routine

Table 7 Signals at threshold 10 000 and sequencing results (real data set)

Family ID	*SNP*	*Window*	*Ratio*[a]	*Taqman*[b]	*Predicted*[c]	*Sequencing*[d]	*Hit*[e]
427	7	7–10	1.15e+04	12 22 12	12 12 12	12 22 12	No
427	11	8–11	1.13e+04	12 11 12	12 12 12	12 11 12	No
508	3	3–21	1.60e+04	12 11 12	12 11 11	12 11 12	No
522	26	23–30	1.09e+04	12 12 11	12 12 12	12 11 11	Yes
646	18	18–25	6.77e+06	12 22 12	12 12 12	12 12 12	Yes
666	22	18–24	7.08e+04	22 12 12	12 22 12	22 12 12	No[f]
666	23	18–26	6.26e+07	12 22 12	12 12 12	12 12 12	Yes
707	14	13–19	1.41e+04	12 12 22	12 12 12	12 12 22	No[f]
707	18	16–21	5.30e+04	12 12 12	12 12 11	12 12 11	Yes
707	20	16–24	3.31e+05	12 12 11	12 12 12	12 12 11	No[f]
707	23	21–29	6.89e+04	12 11 12	12 11 11	12 11 11	Yes
756	15	15–29	1.42e+07	12 11 12	12 12 12	12 12 12	Yes
814	29	15–30	3.80e+04	12 11 11	12 11 12	12 11 11	No
821	22	15–35	∞	22 12 12	22 12 22	22 12 22	Yes
821	23	4–23	4.23e+04	22 12 22	22 12 12	22 12 22	No[f]
821	30	22–35	∞	11 12 11	11 12 12	11 12 11	No[f]
830	15	15–22	1.20e+08	22 11 12	11 22 12	22 11 12	No[f]
830	22	10–22	4.96e+08	22 11 12	11 22 12	11 22 12	Yes
830	25	25–29	3.46e+06	11 22 12	22 11 12	22 11 00	Yes
1083	2	1–4	4.89e+04	12 12 12	12 12 11	12 00 11	Yes
1083	3	1–4	4.89e+04	12 12 22	12 12 12	12 12 22	No[f]
1083	23	15–29	8.14e+04	11 12 11	11 12 12	11 12 12	Yes
1109	4	4–18	1.38e+05	12 11 12	12 11 11	12 11 11	Yes
1109	5	4–5	1.89e+04	12 11 11	12 11 12	12 11 11	No[f]
1109	23	18–24	1.16e+05	12 11 12	12 11 11	12 11 11	Yes
1147	18	18–21	9.72e+04	11 12 12	11 12 11	11 12 11	Yes
1147	19	18–19	6.28e+04	22 12 22	22 12 12	22 12 22	No[f]
1269	2	1–3	3.97e+04	12 12 12	12 12 11	12 00 12	No[f]
1269	3	3–4	4.82e+05	12 12 22	12 22 12	00 12 12	Yes
1269	4	4–17	4.09e+06	12 22 12	12 12 12	12 21 12	Yes
1269	18	16–25	2.77e+06	12 22 12	12 12 12	12 12 12	Yes
1269	22	16–25	7.36e+06	12 22 12	12 12 12	12 22 12	No[f]
1284	15	10–24	1.22e+04	11 12 00	22 12 22	22 12 22	Yes
1327	10	10–30	1.25e+07	12 11 12	12 12 12	00 12 11	Yes
1370	4	4–23	5.21e+05	12 12 11	12 12 12	12 12 12	Yes
1370	5	4–5	1.89e+04	12 12 12	12 12 11	12 12 12	No[f]
11127	30	12–32	7.71e+04	12 11 12	12 12 12	12 11 12	No
11128	31	31–35	5.57e+04	12 12 22	12 12 12	00 12 22	No[f]
11128	33	27–33	1.06e+05	12 12 12	12 12 22	12 12 22	Yes
11227	18	16–30	6.09e+05	12 22 12	12 12 12	12 12 12	Yes
11759	3	3–22	1.99e+05	22 12 12	12 12 12	12 12 12	Yes

[a]Likelihood ratio of the new most likely haplotype explanation and the most likely haplotype explanation of the real data.
[b]Original Taqman genotypes. Alleles are coded '1' and '2'. Genotypes are reported in the order 'father's genotype', 'mother's genotype', 'offspring genotype'.
[c]Genotypes as predicted by the most likely haplotype explanation found with our routine.
[d]Genotypes obtained via resequencing.
[e]This column indicates if the signal corresponds to a true genotyping error.
[f]Cases in which either parts of the sequencing failed or where the original genotyping was correct, but where there is an error at a nearby marker in the family.

in our FAMHAP software. Both simulation studies and application to a real data set confirm the usefulness of our approach. Together with those genotyping errors that are readily visible as MIs, the method achieved an average sensitivity of 63% and a predictive value of 79.7% for signals at threshold $c = 100$ with maxhap = 100 and minloc = 5 in the simulation study. When applied to the real data set, only five out of 41 signals at threshold $c = 10\,000$ were false positives in a strict sense. Our method works properly as long as the genotyping error rate is low enough to allow for an accurate estimation of the true haplotype frequency distribution of the data. We did not investigate the impact of deviations from Hardy–Weinberg equilibrium or of more complex error models on the performance of our method because of running time restrictions. The error routine took about 30 min to screen the real data set with maxhap = 50, but with maxhap = 100 it increased to more than 2 h. If one is willing to assume that our simple error model holds, the true error rate can be approximated via the number of observed Mendelian inconsistencies. As shown by Gordon *et al*[2] and confirmed by our simulation study, between 25 and 30% of the genotyping errors are detectable as MIs, independent of the actual error rate. A general strategy to optimize the search

for genotyping errors in a sample of trios can be outlined as follows. Genotyping should be carried out blind to the family relationships of the individuals. For each marker, the number of observed MIs is used to estimate the overall genotyping error rate of the marker in the sample. As an alternative, the method proposed by Morris and Kaplan[14] can be used to allow for more complex error models and to determine their parameters. Next, the haplotype frequency distribution of the sample is determined with the EM-algorithm. Then, a simulation study as described in this paper should be carried out, based on the estimated haplotype frequencies, and, in addition to our simulation study, based on the estimated error model for each marker. From the simulation study, the optimal threshold for signal detection can be determined. Finally, these thresholds can be applied to the real data set for detecting possible genotyping errors. How should the potential genotyping errors be dealt with? The strictest way is to exclude all families from the analysis, which show MIs or signals at the pre-specified threshold for at least one marker. A less stringent approach is to treat the potentially erroneous genotypes as missing and to carry out the analysis on the modified data set. Finally, one could even impute those genotypes that produce signals according to the new most likely haplotype explanation and analyze the imputed data. Note that none of the strategies guarantees that subsequent association tests are valid. As mentioned before, ignoring genotypes which become visible as MIs lead to an inflated type I error rate of the TDT. With our method, it is possible to detect further potential genotyping errors than by looking at Mendelian inconsistencies only, but still all genotyping errors will not be found. Therefore, there is good reason to believe that subsequent analysis is improved in terms of its validity, but the possibility that the inflation of the type I error rate even grows cannot be excluded. In order to clarify this issue, we plan to carry out a simulation study that evaluates the validity and power of different test strategies in the presence of genotyping errors, dependent on the strategy used to deal with observed genotyping errors. Until such results are available, in practice, a sensible strategy would be to try all of these possibilities to deal with potential genotyping errors and to compare the results of the subsequent association analysis. If the results do not depend on the chosen strategy, one can have good confidence in the analysis, but if the results do depend on the chosen treatment of potential genotyping errors, it is recommendable to re-genotype conspicuous markers or individuals.

Acknowledgements

Our work was supported by grant Kn 378/1 (Project D1 of FOR 423) from the Deutsche Forschungsgemeinschaft. The authors wish to thank T.Wesse, A.Dietsch and H.Hinz (Kiel) for expert technical help.

Electronic database information

URLs for data presented herein are as follows: FAMHAP: Haplotype Frequency Estimation, http://www.uni-bonn.de/~umt70e/becker.html

References

1 Hosking L, Lumsden S, Lewis K *et al*: Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet* 2004; **12**: 395–399.
2 Gordon D, Heath SC, Ott J: True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Hered* 1999; **49**: 65–70.
3 Douglas JA, Skol AD, Boehnke M: Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet* 2002; **70**: 487–495.
4 Ehm MG, Kimmel M, Cottingham Jr RW: Error detection for genetic data using likelihood methods. *Am J Hum Genet* 1996; **58**: 225–234.
5 Douglas JA, Boehnke M, Lange K: A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet* 2000; **66**: 1287–1297.
6 Mukhopadhyay N, Buxbaum SG, Weeks DE: Comparative study of multipoint methods for genotype error detection. *Hum Hered* 2004; **58**: 175–189.
7 Mote VL, Anderson RL: An investigation of the effect of misclassification on the properties of χ^2 tests in the analysis of categorical data. *Biometrika* 1965; **52**: 95–109.
8 Gordon D, Heath SC, Liu X, Ott J: A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 2001; **69**: 371–380.
9 Mitchell AA, Cutler DJ, Chakravarti A: Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* 2003; **72**: 598–610.
10 Knapp M, Becker T: Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *Am J Hum Genet* 2004; **74**: 589–591 (Letter).
11 Zhang S, Sha Q, Chen HS, Dong J, Jiang R: Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet* 2003; **73**: 566–579.
12 Becker T, Knapp M: Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol* 2004; **27**: 21–32.
13 Anderson EC, Novembre J: Finding haplotype block boundaries by using the minimum-description-length principle. *Am J Hum Genet* 2003; **73**: 336–354.
14 Morris RW, Kaplan N: Testing for association with a case-parent design in the presence of genotyping errors. *Genet Epidemiol* 2004; **26**: 142–152.

Appendix A

Here, we describe the algorithm used to identify the most likely haplotype explanation for a trio. Consider a trio sample, which is genotyped at m tightly linked markers and let $H=\{1, 2, \ldots, t\}$ be the set of haplotypes with an estimated frequency greater than zero. We consider a fixed trio and its most likely haplotype explanation (k, l, u, v). For one of the m markers, we set the genotype for one individual of the trio to be missing. We aim to determine the most likely haplotype explanation (k_0, l_0, u_0, v_0) that is compatible with the modified multi-marker genotype. A naive algorithm to find the most probable haplotype explanation would check all t^4

haplotype combinations by four nested loops. Each loop would represent one of the four parental haplotypes. We have made three improvements to this naive algorithm. First of all, the order of the nested loops can be optimized. The first loop runs over the father's putatively transmitted haplotype and the second loop runs over the mother's putatively transmitted haplotype. In this way, it is possible to check for compatibility with the child's genotype after only two nested loops. If the child's genotype is not compatible, it is not necessary to compute the two remaining inner loops. Secondly, it is not necessary that the loops run over all haplotypes of H. For instance, if the father is homozygous at marker i for allele 1 and if $h \in H$ is a haplotype with allele 2 at position i, it is clear that there is no haplotype explanation for the trio with h as one of the father's haplotypes. Similarly, if the child is homozygous for allele 1 at position i, it is clear that neither the father's transmitted haplotype nor the mother's transmitted haplotype can be h. With these simple rules, it is possible to define a set $FT \subseteq H$ which includes all haplotypes that could possibly be transmitted from the father to the child. In the same way, we define a set FNT of haplotypes which may plausibly represent the non-transmitted haplotypes of the father and analogous sets MT, MNT for the mother. Now, it is sufficient that the outer loop of the algorithm runs over FT, that the first inner loop runs over MT, that the second inner loop runs over FNT and the most inner loop runs over MNT. Thirdly, many loops can be stopped prematurely, if the haplotypes are ordered according to their frequency estimate, beginning with the most frequent haplotype. As we are looking only for the most likely haplotype explanation for the modified genotype data, we need not run over all its possible haplotype explanations. Let $ft_1 \in FT$ be the most frequent haplotype of FT and let fnt_1, mt_1, mnt_1 be defined analogously. Suppose that we are, for instance, in the third of the nested loops, and that we are considering the jth most frequent haplotype $fnt_j \in FNT$. Now, if the likelihood of the haplotype explanation $(ft_1, fnt_j, mt_1, mnt_1)$ is smaller than the likelihood of the most likely haplotype explanation found so far, it is not necessary to finish the third loop and its inner loop, because the likelihood can only be smaller than the likelihood of $(ft_1, fnt_j, mt_1, mnt_1)$. Instead, it is possible to continue with the next value of the second loop. This is particularly helpful for families with a very low typing ratio.

We give an example for our algorithm. Suppose that three SNPs were genotyped and that haplotype $111 := h_1$ has an estimated frequency of $f_{111} = 0.5$, haplotype $112 := h_2$ has an estimated frequency of $f_{112} = 0.3$, haplotype $212 := h_3$ and haplotype $221 := h_4$ have an estimated frequency of $f_{212} = f_{221} = 0.05$, and that the remaining haplotypes all have a frequency of 0.025. We consider a trio for which the second marker has been set to missing. Let $G_f = ((1, 2), (0, 0), (1, 2))$ be the father's modified multi-marker genotype, let $G_m = ((1, 2), (0, 0), (1, 2))$ be the mother's modified multi-marker genotype and let $G_c = ((1, 1), (0, 0), (1, 2))$ be the child's modified multi-marker genotype. A naive algorithm would have to check $8^4 = 4096$ haplotype configurations (ft, fnt, mt, mnt) for compatibility with the multi-locus genotypes. The sets FT and MT contain only haplotypes that have allele 1 at the first marker, the sets FNT and MNT contain only haplotypes that have allele 2 at the first marker. Thus, the number of combinations that have to be checked can be reduced to $4^4 = 256$. When haplotypes are ordered by their frequencies, we have $FT = MT = \{h_1, h_2, 121, 122\}$ and $FNT = MNT = \{h_3, h_4, 211, 222\}$. The outer loop will start with $ft = h_1$. As $mt = h_1$ will not be compatible with the child after this choice of ft, the first value of the second loop that does not lead to a premature stop will be $mt = h_2$. The first value of the third loop is $fnt = h_3$ and then $mnt = h_3$ is not possible and it follows that $mnt = h_4$. Thus, the haplotype explanation (h_1, h_3, h_2, h_4) has been found, and it is easily seen that it is one of two haplotype explanations with maximal likelihood, the other being (h_2, h_4, h_1, h_3), which differs only with respect to the parental origin of the haplotypes.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg).