

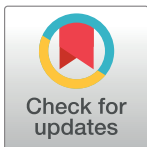
RESEARCH ARTICLE

Rules for resolving Mendelian inconsistencies in nuclear pedigrees typed for two-allele markers

Sajjad Ahmad Khan^{1,2*}, Sadaf Manzoor², Alamgir³, Amjad Ali², Dost Muhammad Khan¹, Umair Khalil¹

1 Department of Statistics, Abdul Wali Khan University Mardan, Khyber Pukhtunkhwa, Pakistan, **2** Department of Statistics, Islamia College, Peshawar, Khyber Pukhtunkhwa, Pakistan, **3** Department of Statistics, University of Peshawar, Khyber Pukhtunkhwa, Pakistan

* sajjadkhan@awkum.edu.pk, sajjadkahmad@gmail.com



Abstract

Gene-mapping studies, regularly, rely on examination for Mendelian transmission of marker alleles in a pedigree as a way of screening for genotyping errors and mutations. For analysis of family data sets, it is, usually, necessary to resolve or remove the genotyping errors prior to consideration. At the Center of Inherited Disease Research (CIDR), to deal with their large-scale data flow, they formalized their data cleaning approach in a set of rules based on PedCheck output. We scrutinize via carefully designed simulations that how well CIDR's data cleaning rules work in practice. We found that genotype errors in siblings are detected more often than in parents for less polymorphic SNPs and vice versa for more polymorphic SNPs. Through computer simulations, we conclude that some of the CIDR's rules work poorly in some circumstances, and we suggest a set of modified data cleaning rules that may work better than CIDR's rules.

OPEN ACCESS

Citation: Khan SA, Manzoor S, Alamgir, Ali A, Khan DM, Khalil U (2017) Rules for resolving Mendelian inconsistencies in nuclear pedigrees typed for two-allele markers. *PLoS ONE* 12(3): e0172807. doi:10.1371/journal.pone.0172807

Editor: Katriina Aalto-Setälä, University of Tampere, FINLAND

Received: December 1, 2016

Accepted: February 9, 2017

Published: March 2, 2017

Copyright: © 2017 Khan et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data set has been attached in the supporting information file named S1 File.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

A genotyping error arises when the observed genotype differs from the true underlying genotype [1, 2]. Even with the most modern techniques, the observed genotype does not always match the true underlying genotype, and this has been shown to occur at a rate of 0.5–7% for microsatellite markers [3]. Error rates are influenced by a number of factors, but are generally quoted between 0.25% and 1% for microsatellite genotyping [4]. Where family information is available, a proportion of genotyping errors can be detected as Mendelian inconsistencies, but this is more difficult for single nucleotide polymorphism (SNP) markers with only two alleles [5]. Several authors have shown that even a small error rate (i.e., 1–2%) can have a massive impact on linkage results [6–9].

In linkage and association analysis, investigators and researchers are fully aware of the consequences of genotyping errors at the marker loci [6, 10–13]. It is well known that misspecified marker allele frequencies, genotyping errors and Mendelian inconsistencies can lead to a

Table 1. CIDR's rules for removing Mendelian inconsistencies.

Situation in a Nuclear Family	Error Messages	Actions	Short Rule Name ^a
1 parent is inconsistent with 1 child	ERROR: Child 01 and Mother are inconsistent	Zero out the child's genotype	1P1C:C ₀
	OR ERROR: Child 01 and Father are inconsistent.		
1 parent is inconsistent with 2 + children	ERROR: Child 01 and Mother are inconsistent	Zero out the specific parent genotype	1P2+C:P ₀
	AND ERROR: Child 02 and Mother are inconsistent		
	OR ERROR: Child 01 and Father are inconsistent		
	AND ERROR: Child 02 and Father are inconsistent.		
2 parents are inconsistent with 1 child	ERROR: Child 01 is consistent with each parent separately, but not as a pair	Zero out the child's genotype	2P1C:C ₀
	OR ERROR: Child 01 and Mother are inconsistent		
	AND ERROR: Child 01 and Father are inconsistent.		
2 parents are inconsistent with 2 + children	ERROR: Child 01 is consistent with each parent separately, but not as a pair	Zero out the genotypes of the whole nuclear family	2P2+C:W ₀
	AND ERROR: Child 02 is consistent with each parent separately, but not as a pair.		
	OR ERROR: Child 01 and Father are inconsistent.		
	AND ERROR: Child 02 and Mother are inconsistent.		

^aLet P = parent, C = child, and W = whole family, then we name CIDR's rules as 1. 1P1C:C₀, 2. 1P2+C:P₀, 3. 2P1C:C₀, 4. 2P2+C:W₀, where C₀ = zero out the child's genotype, P₀ = zero out the specific parent genotype, and W₀ = zero out the genotypes of the whole nuclear family.

doi:10.1371/journal.pone.0172807.t001

systematic increase in false-positive rates. Power may be reduced, and parameter estimates may be biased and/or inconsistent [1–2, 7, 14–19].

For analysis of family data, it is, usually, necessary to resolve or remove the genotyping errors prior to analysis. There are no hard and fast rules about how to clean genotyping errors from pedigree data. Researchers, typically, clean their data using: PedCheck [20], MERLIN [21], MENDEL [22], SimWalk2 [23], or Sibmed [7].

At the Center of Inherited Disease Research (CIDR), to deal with their large-scale data flow; they formalized their data cleaning approach in a set of rules, which are presented in Table 1. They use PedCheck [20] to detect the inconsistencies, and then specific cleaning rules are triggered by certain combination of error messages from PedCheck. Here, we examine, via carefully designed simulations how well CIDR's data cleaning rules work in practice by answering the following three questions: i) How often are genotyping errors detected?; ii) How often are these rules applied?; iii) How often are these rules applied correctly?

Methods

We wish to evaluate how well the CIDR rules (see Table 1) work in terms of accurately removing Mendelian inconsistencies from the data set. The Center of Inherited Disease Research

(CIDR), routinely, uses PedCheck to detect Mendelian inconsistencies for each SNP. PedCheck [20] offers different error-checking levels: Level 0 checks for formatting errors in the pedigree structure data, Level 1 checks for Mendelian errors in nuclear families, Level 2 checks for all other Mendelian errors using the genotype elimination algorithm. CIDR runs Level 0 and 1 checks on a data set and removes Mendelian inconsistencies. After clearing up Level 0 and Level 1 errors, Level 2 checks are run to make sure that the data are free of Mendelian inconsistencies.

In our simulation study, we explore three questions:

1. How often are genotyping errors detected?
2. How often are these rules applied?
3. How often are these rules applied correctly?

We evaluate these questions via simulation study. In our simulation study, first we simulate error-free marker data for a single SNP (single nucleotide polymorphism) for 20,000 nuclear families (S1 File) with sibship sizes 2 to 6 [24] and SNP minor allele frequency (MAF) taking on the values 0.5, 0.4, 0.3, 0.2, 0.1 [25]. These simulations were done using the SIMULATE program [26]. Genotypes were simulated for all pedigree members. Secondly, we add in genotype errors using Mega2 [27]; this requires specification of the probability model for introducing errors, and the error rate. We introduce errors by picking a genotype at random with probability 0.01, and then changing the true genotype to one of the others with equal probability (for more see [28]). After introducing genotyping errors in data, we run Level 1 of PedCheck [20] to find the true underlying errors. Now we have two matched data sets: the original error free one, and the second one containing errors. After running PedCheck, we compute the percent of time genotype errors are detected in siblings, parents, or either. We also tabulate the percent of time each rule is applied. Finally, we compute how often each rule is applied correctly. We consider a rule to have been applied correctly if the genotype it zeroed out is a truly erroneous genotype.

Results

We present the results as Figures—for detailed counts, please see the supplemental tables. Fig 1 and S1 Table shows the percent of time genotyping errors are detected in siblings, parents, and in either. Using data set (S1 File), we observe that for sibship size 2 with MAF 0.3–0.5, error detection rates in parents and in siblings are very similar, while for sibship sizes 3 to 6, errors in parents are detected more often than errors in siblings. As the SNP becomes less polymorphic (i.e., MAF 0.1–0.2), errors in siblings are detected more often than errors in parents for all sibship sizes.

For each genotype error that is detected, a CIDR rule, as defined in Table 1, is triggered (Fig 2 and S2 Table). Rules 1P1C:C₀ and 2P1C:C₀ are inversely proportional in application. As the MAF decreases from 0.5 to 0.1, the frequency of application of rule 1P1C:C₀ decreases from ~58% to ~18%, while that of rule 2P1C:C₀ increases from ~18% to ~60%. Rule 2P2+C:W₀ is only applied 1–12% of the time. For MAF 0.1–0.2, rule 2P2+C:W₀ is applied 1–6% of the time, and for MAF 0.3–0.5, rule 2P2+C:W₀ is applied 6–12% of the time. Similarly, overall, rule 1P2+C:P₀ is applied from 12% to 31% of the time. It is applied most frequently when the MAF is 0.1 and the sibship size is 2.

Fig 3 and S3 Table show how often each rule is applied correctly. Since rule 2P2+C:W₀ zeros out the whole pedigree if it contains one or more genotyping error, it is always “correctly” applied. So we excluded rule 2P2+C:W₀ from Fig 3 and S3 Table.

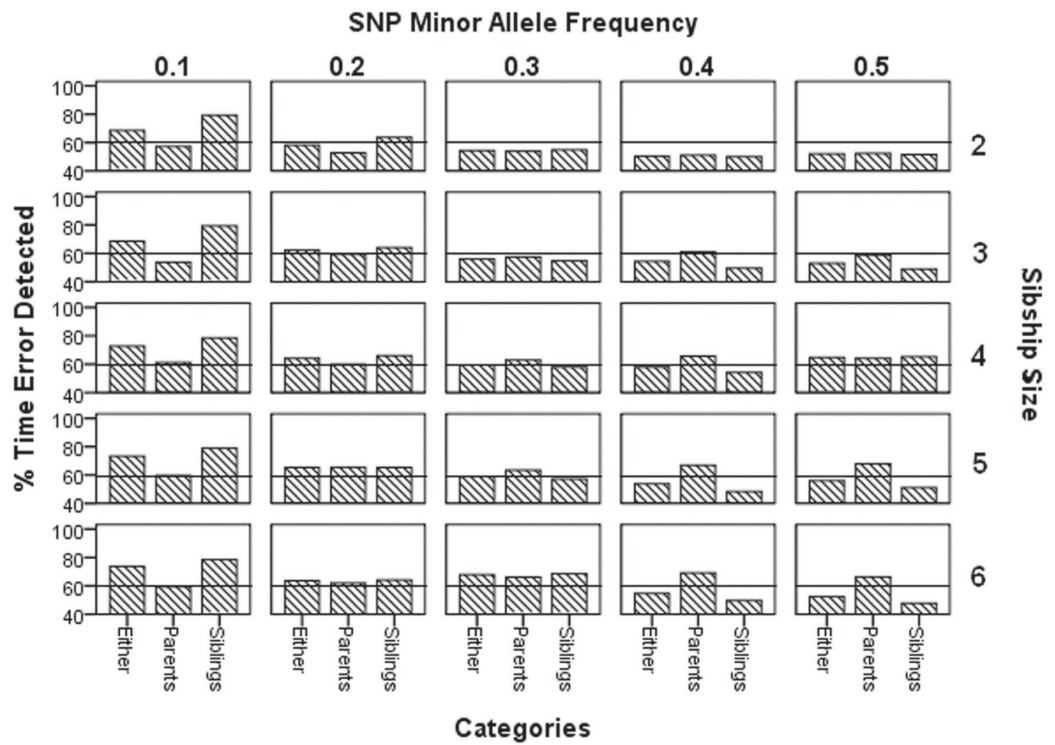


Fig 1. Percent of time a genotyping error is detected.

doi:10.1371/journal.pone.0172807.g001

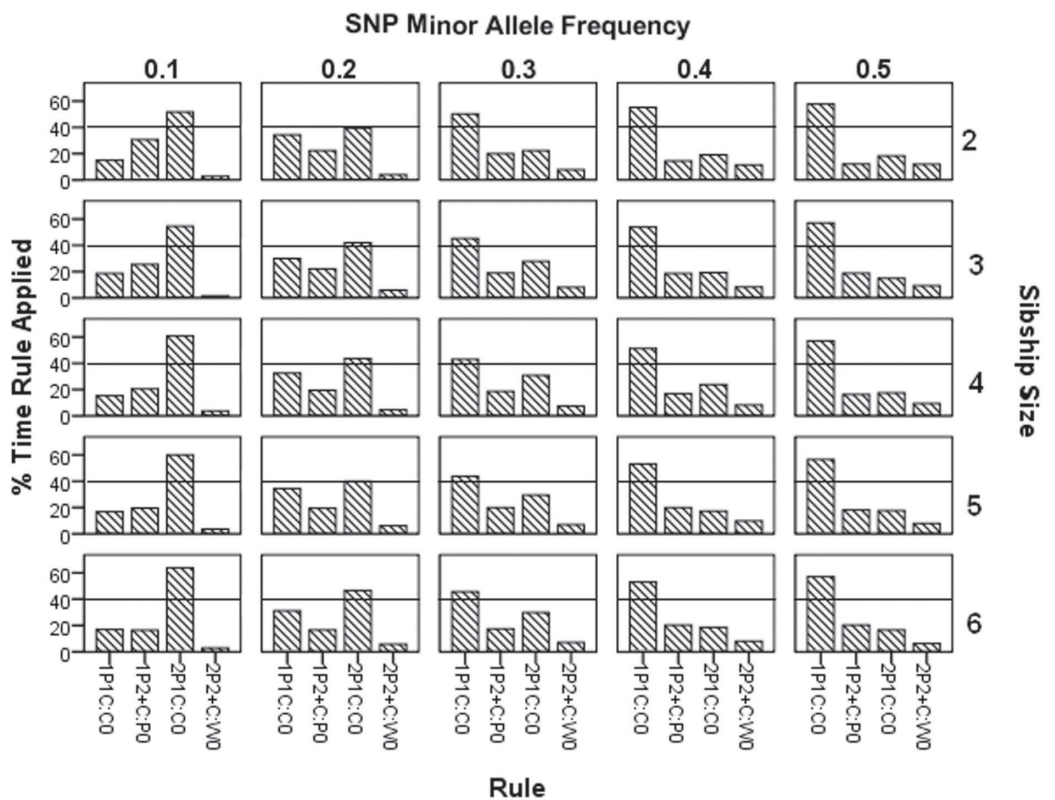


Fig 2. Percent of time each rule is applied.

doi:10.1371/journal.pone.0172807.g002

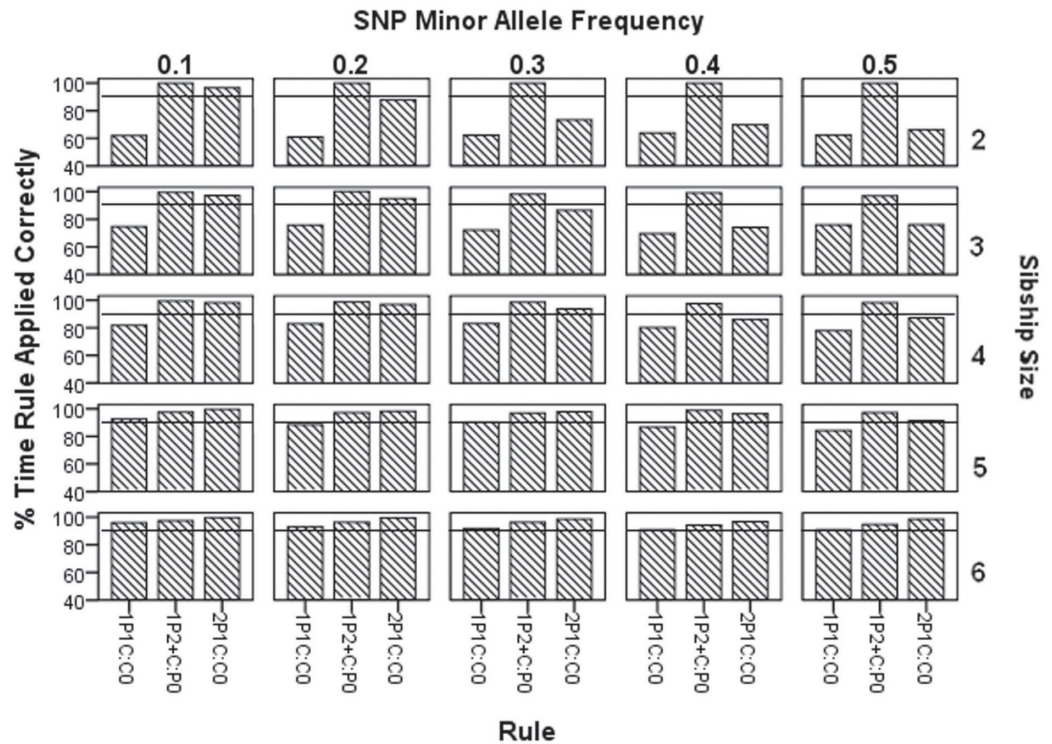


Fig 3. Percent of time each rule is applied correctly.

doi:10.1371/journal.pone.0172807.g003

For rule 1P1C:C₀, we observe in Fig 3 that for sibship size 2, the action taken is 61–64% correct. When we increase the sibship size from 2 to 3, then the percentage of the correctness of rule 1P1C:C₀ also increases to 70–76%. Similarly in the same fashion, the correctness of rule 1P1C:C₀ is 78–83%, 84–92% and 91–96% for sibship sizes 4, 5 and 6 respectively. In addition to noting this systematic increase in correctness of rule 1P1C:C₀ with increase in sibship size, it is also important to note that there are not any drastic changes in the correctness of rule 1P1C:C₀ with change in MAF within a given sibship size.

Rule 1P2+C:P₀ is correctly applied 100% of the time for sibships of size 2, regardless of the MAF. As the sibship size increases, the frequency of correct application only declines very slightly: for sibship size 6, rule 1P2+C:P₀ is applied correctly 94–97% of the time.

For rule 2P1C:C₀, Fig 3 shows that for sibship size 2 with MAF 0.5, action taken for rule 2P1C:C₀ (see Table 1) is 66% correct. For the same sibship size, the percentage of correctness of action taken for rule 2P1C:C₀ is 70%, 73%, 88% and 97% respectively for MAF 0.4, 0.3, 0.2 and 0.1. In similar fashion when sibship size increases, the percentage of the correctness of action taken for rule 2P1C:C₀ also increases, whereas the percentage of correctness of action are 66%, 76%, 87%, 91% and 99% respectively for sibship size 2, 3, 4, 5 and 6 with MAF 0.5. The action taken for rule 2P1C:C₀ is highly correct for higher sibship sizes and also for low MAF when the sibships are smaller.

Conclusion and discussion

In this study, we simulated data for nuclear families to examine the behavior of the CIDR data cleaning rules (Table 1). These rules are determined by the results of running PedCheck [19] to detect Mendelian inconsistencies. We examined how often a given error is detected, how often the rules are applied, and how often each rule is correctly applied.

Fig 1 shows how often errors are detected, and indicates that usually more errors are detected in parents than in siblings when MAF is 0.3–0.5. Similarly, when MAF is smaller (0.1–0.2), more errors are detected in siblings than parents. Overall true errors are detected at the rate of 51–74%. Douglas et al. [29] derived at the rate of 30–48% for 2 alleles model and 51–74% for 4 alleles model.

Fig 2 shows how often each rule is applied, and indicates that rule 1P1C:C₀ is applied most frequently in nuclear families for MAF 0.3–0.5. Similarly, for SNP allele frequencies 0.1–0.2, rule 2P1C:C₀ is applied most frequently, while rule 1P2+C:P₀ is applied moderately (15–20%). Rule 2P2+C:W₀ is always the least frequently applied rule (1–12%) across all MAF values and sibship sizes.

Fig 3 shows that how often actions taken by the CIDR rules shown in Table 1 are correct. Note that rule 2P2+C:W₀ is excluded from Fig 3 because the concept of ‘correctness’ is not applicable to it if correctness means ‘the rule correctly zeroed out *only* the erroneous genotype’. Rule 2P2+C:W₀ zeros out all the genotypes for the entire family. So while it does zero out the erroneous genotype, it also zeros out several correct genotypes.

Rule 1P2+C:P₀ is almost always applied correctly (94–100%) as it is always correctly applied when there is one detectable true error in the nuclear family, and the underlying true error is in the parents (Fig 3). Alternatively, if there is only one true error in the pedigree in a single child, then it will not trigger rule 1P2+C:P₀ because it will not cause one parent to be inconsistent with 2 or more children. When rule 1P2+C:P₀ is applied less than 100% correctly, this is due to more than one true (and detectable) error occurring within a given family, which is a rare event in the smaller sibship sizes.

Rule 1P1C:C₀ performs most incorrectly on smaller sibships (Fig 3) and is applied most frequently when it is extremely wrong (Fig 2), and is consistently more frequently wrong than the other rules. Rule 2P1C:C₀ becomes more correctly applied as the MAF becomes smaller. CIDR’s rules are more often correctly applied as the sibship sizes get larger and MAF becomes smaller.

If we focus on where the rules are correctly applied (i.e. greater than 95% of the time), then we might come up with the following alternative rules, which should be better than CIDR’s rules:

1. Delete Rule 1P1C:C₀, and instead zero out all the genotypes of the whole nuclear family when rule 1P1C:C₀’s triggers apply.
2. Keep Rule 1P2+C:P₀
3. Apply Rule 2P1C:C₀ as a function of the MAF and sibship size; otherwise zero out all the genotypes of the whole family instead. Rule 2P1C:C₀ is only to be applied to sibship of size 2 if the MAF ≤ 0.1 ; to size 3 if the MAF ≤ 0.2 ; to size 4 if the MAF ≤ 0.3 ; to size 5 if the MAF ≤ 0.4 ; and to size 6 for all values of the MAF.
4. Keep Rule 2P2+C:W₀

We can also use another alternative approach which might be better than the CIDR’s rules—we may use Pedcheck’s Level 4 checking, and then zero out the genotype of any person whose alternative genotypes have at least one odds ratio of 1.0. Bedzioch et al. [30] examines genotyping errors using Level 4 of PEDCHECK for 4 data sets and conclude that Level 4 checking worked quite well (even when PEDCHECK did not indicate the most probable genotyping error in a few cases).

Supporting information

S1 File. Data set.

(PDF)

S1 Table. Percent of time genotype errors detected. The denominator is the number of actual errors present and the numerator is the number of errors detected by PedCheck.

(DOCX)

S2 Table. Percent of time each rule is applied.

(DOCX)

S3 Table. Percent of time each rule is correctly applied. The denominator is the number of times each rule is applied and the numerator is the number of times each rule is applied correctly.

(DOCX)

Author Contributions

Conceptualization: SAK.

Data curation: SAK SM.

Formal analysis: SAK A.

Funding acquisition: SAK.

Investigation: SAK AA DMK.

Methodology: SAK UK.

Project administration: SAK SM.

Resources: SAK SM A.

Software: AA DMK.

Supervision: A UK.

Validation: A SAK.

Visualization: SAK.

Writing – original draft: SAK.

Writing – review & editing: SM A AA.

References

1. Neale BM, Ferreira M. AR, Medland SE and Posthuma D. *Statistical Genetics: Gene Mapping Through Linkage and Association*. Taylor and Francis Group, New York, NY 10016. 2008.
2. Ziegler A and König IR. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. 2006.
3. Brzustowicz LM, Merette C, Xie X, Townsend L, Gilliam TC and Ott J. Molecular and statistical approaches to the detection and correction of errors in genotype databases. *Am. J. Hum. Genet.* 1993; 53:1137–1145. PMID: [8213837](#)
4. Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW and Foote SJ. Identification and analysis of error types in high-throughput genotyping. *Am. J. Hum. Genet.* 2000; 67:727–736 doi: [10.1086/303048](#) PMID: [10924406](#)

5. Gordon D, Heath SC and Ott J. True pedigree errors more frequent than apparent errors for single nucleotide polymorphism. *Human Heredity*. 1999; 49:65–70 PMID: [10077724](#)
6. Buetow KH. Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am. J. Hum. Genet.* 1991; 49:985–994 PMID: [1928104](#)
7. Douglas JA, Boehnke M and Lange K. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am. J. Hum. Genet.* 2000; 66:1287–1297 doi: [10.1086/302861](#) PMID: [10739757](#)
8. Abecasis GR, Cherny SS, Cardon LR. The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet.* 2001; 9:130–134 doi: [10.1038/sj.ejhg.5200594](#) PMID: [11313746](#)
9. Hess M, Rhydderch J, LeClair L, Buckley R, Kawase M, Hauser L. Estimation of genotyping error rate from repeat genotyping, unintentional recaptures and known parent-offspring comparisons in 16 micro-satellite loci for brown rockfish (*Sebastes auriculatus*). *Molecular Ecology Resources*. 2012; 12(6): 1114–1123. doi: [10.1111/1755-0998.12002](#) PMID: [22958648](#)
10. Lathrop GM, Hooper AB, Huntsman JW, Ward RH. Evaluating pedigree data: I. The estimation of pedigree error in the presence of marker mistyping. *Am. J. Hum. Genet.* 1983; 35:241–262 PMID: [6573130](#)
11. Terwilliger JD, Weeks DE and Ott J. Laboratory errors in the reading of marker alleles cause massive reductions in LOD score and lead to gross overestimation of the recombination fraction. *Am. J. Hum. Genet.* 1990; Suppl. 47:A201
12. Goring HHH, Terwilliger JD. Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am. J. Hum. Genet.* 2000; 66:1107–1118 doi: [10.1086/302798](#) PMID: [10712221](#)
13. Kang SJ, Gordon D and Finch SJ. What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology* 2004; 26:132–141 doi: [10.1002/gepi.10301](#) PMID: [14748013](#)
14. Seaman SR and Holmans P. Effect of genotyping error on type I error rate of affected sib pair studies with genotyped parents. *Hum. Hered.* 2005; 59:157–164 doi: [10.1159/000085939](#) PMID: [15925894](#)
15. Hao K, Li C, Rosenow C and Wong WH. Estimation of genotype error rate using samples with pedigree information—an application on the GeneChip Mapping 10K array. *Genomics* 2004; 84:623–630 doi: [10.1016/j.ygeno.2004.05.003](#) PMID: [15475239](#)
16. Cherny SS, Abecasis GR, Cookson WOC, Sham PC and Cardon LR. The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. *Genet. Epidemiol.* 2001; 21 (Suppl 1):S117–S122
17. Goring HHH, Terwilliger JD. Linkage analysis in the presence of errors III: Marker loci and their map as nuisance parameters. *Am. J. Hum. Genet.* 2000; 66:1298–1309 doi: [10.1086/302846](#) PMID: [10731467](#)
18. Terwilliger JD and Ott J. *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, 1994
19. Ott J. Strategies for characterizing highly polymorphic markers in human gene mapping. *Am. J. hum. Genet.* 1992; 51:283–290 PMID: [1642229](#)
20. O'Connell JR and Weeks DE. PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* 1998; 63:259–266 doi: [10.1086/301904](#) PMID: [9634505](#)
21. Abecasis GR, Cherny SS, Cookson WO and Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* 2000; 30:97–101.
22. Lange K, Weeks DE and Boehnke M. Programs for pedigree analysis: MENDEL, FISHER and dGENE. *Genet. Epidemiol.* 1988; 5:471–472. doi: [10.1002/gepi.1370050611](#) PMID: [3061869](#)
23. Sobel E, Papp J and Lange K. Detection of genotyping errors. *Am. J. Hum. Genet.* 2002; 70:496–508 doi: [10.1086/338920](#) PMID: [11791215](#)
24. Beekman M., Blanché H., Perola M., Hervonen A., Bezrukov V., Sikora E., . . . On behalf of the GEHA consortium, C. Genome-wide linkage analysis for human longevity: Genetics of Healthy Ageing Study. *Aging Cell*. 2013; 12(2), 184–193. <http://doi.org/10.1111/accel.12039> doi: [10.1111/accel.12039](#) PMID: [23286790](#)
25. Zhu Z, Yuan D, Luo D, Lu X, Huang S. Enrichment of Minor Alleles of Common SNPs and Improved Risk Prediction for Parkinson's Disease. *PLoS ONE*. 2015; 10(7): e0133421. doi: [10.1371/journal.pone.0133421](#) PMID: [26207627](#)
26. Terwilliger JD, Speer M and Ott J. Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet. Epidemiol.* 1993; 10:217–224
27. Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE. Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics*. 2005; 21:2556–2557. doi: [10.1093/bioinformatics/bti364](#) PMID: [15746282](#)

28. Mukhopadhyay N, Buxbaum SG and Weeks DE. Comparative study of multi-point methods for genotype error detection. *Human Heredity* 2004; 58:175–189 doi: [10.1159/000083545](https://doi.org/10.1159/000083545) PMID: [15812175](https://pubmed.ncbi.nlm.nih.gov/15812175/)
29. Douglas JA, Skol AD and 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 doi: [10.1086/338919](https://doi.org/10.1086/338919) PMID: [11791214](https://pubmed.ncbi.nlm.nih.gov/11791214/)
30. Bedzioch MD, DeFrance HB and Jarvik G. An examination of the genotyping error detection function of SIMWALK2. *BMC Genetics*. 2003; 4; Supp1, pp:S40