

MIT Open Access Articles

*My sister's keeper?: genomic research
and the identifiability of siblings*

The MIT Faculty has made this article openly available. **Please share**
how this access benefits you. Your story matters.

Citation: Cassa, Christopher et al. "My sister's keeper?: genomic research and the identifiability of siblings." BMC Medical Genomics 1.1 (2008): 32.

As Published: <http://dx.doi.org/10.1186/1755-8794-1-32>

Publisher: BioMed Central Ltd.

Persistent URL: <http://hdl.handle.net/1721.1/52440>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution



Research article

Open Access

My sister's keeper?: genomic research and the identifiability of siblings

Christopher A Cassa*^{1,2}, Brian Schmidt², Isaac S Kohane^{1,3} and Kenneth D Mandl^{1,3}

Address: ¹Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA, ²Clinical Decision Making Group, CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA and ³Harvard Medical School, Boston, MA, USA

Email: Christopher A Cassa* - cassa@mit.edu; Brian Schmidt - bschmidt@alum.mit.edu; Isaac S Kohane - isaac_kohane@harvard.edu; Kenneth D Mandl - kenneth_mandl@harvard.edu

* Corresponding author

Published: 25 July 2008

Received: 26 November 2007

BMC Medical Genomics 2008, 1:32 doi:10.1186/1755-8794-1-32

Accepted: 25 July 2008

This article is available from: <http://www.biomedcentral.com/1755-8794/1/32>

© 2008 Cassa et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Genomic sequencing of SNPs is increasingly prevalent, though the amount of familial information these data contain has not been quantified.

Methods: We provide a framework for measuring the risk to siblings of a patient's SNP genotype disclosure, and demonstrate that sibling SNP genotypes can be inferred with substantial accuracy.

Results: Extending this inference technique, we determine that a very low number of matches at commonly varying SNPs is sufficient to confirm sib-ship, demonstrating that published sequence data can reliably be used to derive sibling identities. Using HapMap trio data, at SNPs where one child is homozygous major, with a minor allele frequency ≤ 0.20 , ($N = 452684$, 65.1%) we achieve 91.9% inference accuracy for sibling genotypes.

Conclusion: These findings demonstrate that substantial discrimination and privacy risks arise from use of inferred familial genomic data.

Background

Genomic data are increasingly integrated into clinical environments, stored in genealogical and medical records[1,2] and shared with the broader research community[3,4] without full appreciation of the extent to which these commodity level measurements may disclose the health risks or even identity of family members. While siblings, on average, share half of their contiguous chromosomal segments, well over half of a sibling's allelic values can be inferred using only population-specific allele frequency data and the genotypes of another sib. The informed consent process for research and clinical genomic data transmission must therefore include rigor-

ous treatment of accurately quantified disclosure risks for all who will be impacted by such activity.

It is remarkably easy to positively identify a person with fewer than 40 independent, commonly varying SNPs, using a physical sample or a copy of those values[5]. As DNA sequences cannot be revoked or changed once they are released, any disclosure of such data poses a life-long privacy risk. Unlike conventional fingerprints, which provide little direct information about patients or relatives, SNP genotypes may encode phenotypic characteristics, which can link sequences to people[6]. Despite these privacy issues[7,8], use of genetic sequencing is increasing in

both forensics[9] and clinical medicine. The recent genetic fingerprinting provision in the renewal of the federal Violence Against Women Act[10], alone, may result in one million new sequenced individuals each year, markedly increasing the number of available links between identities and genotypes. This genetic fingerprinting has an impact on people beyond those directly sequenced—genetic testing partially reveals genotypes of siblings and other family members.

At each locus in a child's genome, each parent transmits only one of his or her two chromosomes. If we have the genotype of one child, and would like to use that information to help infer the genotype of a sibling, we consider both the known parental genotypes (for the alleles they have transmitted to their first sibling,) and also consider those chromosomes they have but have not transmitted. We assume that the unknown parental alleles are drawn from a reference population, such as one of the HapMap populations. Now, considering the genotype of the inferred sibling (2nd child), with probability 0.25, the sibling will receive the same 2 chromosomes transmitted to the first child, in which case they will have the same genotype. With probability 0.25, the inferred sibling will receive both previously untransmitted chromosomes, in which case the sibling will have the same genotype distribution as the reference population. If only one of the same chromosomes is transmitted, then one chromosome will be the same and the other will be drawn from the population.

Methods

To quantify the risk of SNP disclosure to relatives, we demonstrate a model for inferring sibling genotypes using proband SNP data and population-specific allele frequency databases, such as the HapMap[10,11]. We also evaluate the probability that two people, in a selected pool of individuals, are siblings given a match at an independent subset of SNPs, and show that this number can be made remarkably low with appropriate SNP selection.

Enhanced ability to infer sibling genotypes

First, consider the case where one sibling's genotype is known to be 'AA', and the goal is to determine the probability that a second sibling's genotype will also be 'AA' at that locus. Because there is additional knowledge—the familial relationship between the two sibs—the prior probability of the second sib carrying a specific genotype at a selected SNP will be altered under the new constraint. A conditional probability expression that sums over the nine possible parental genotypic combinations (for example, maternal genotype 'Aa' with paternal genotype 'AA') at a single SNP, each denoted as i can be used:

$$p(\text{Sib}_2\text{AA} \mid \text{Sib}_1\text{AA}) = \sum_{i=1}^9 p(\text{Sib}_2\text{AA} \mid \text{parental comb. } i) p(\text{parental comb. } i \mid \text{Sib}_1\text{AA})$$

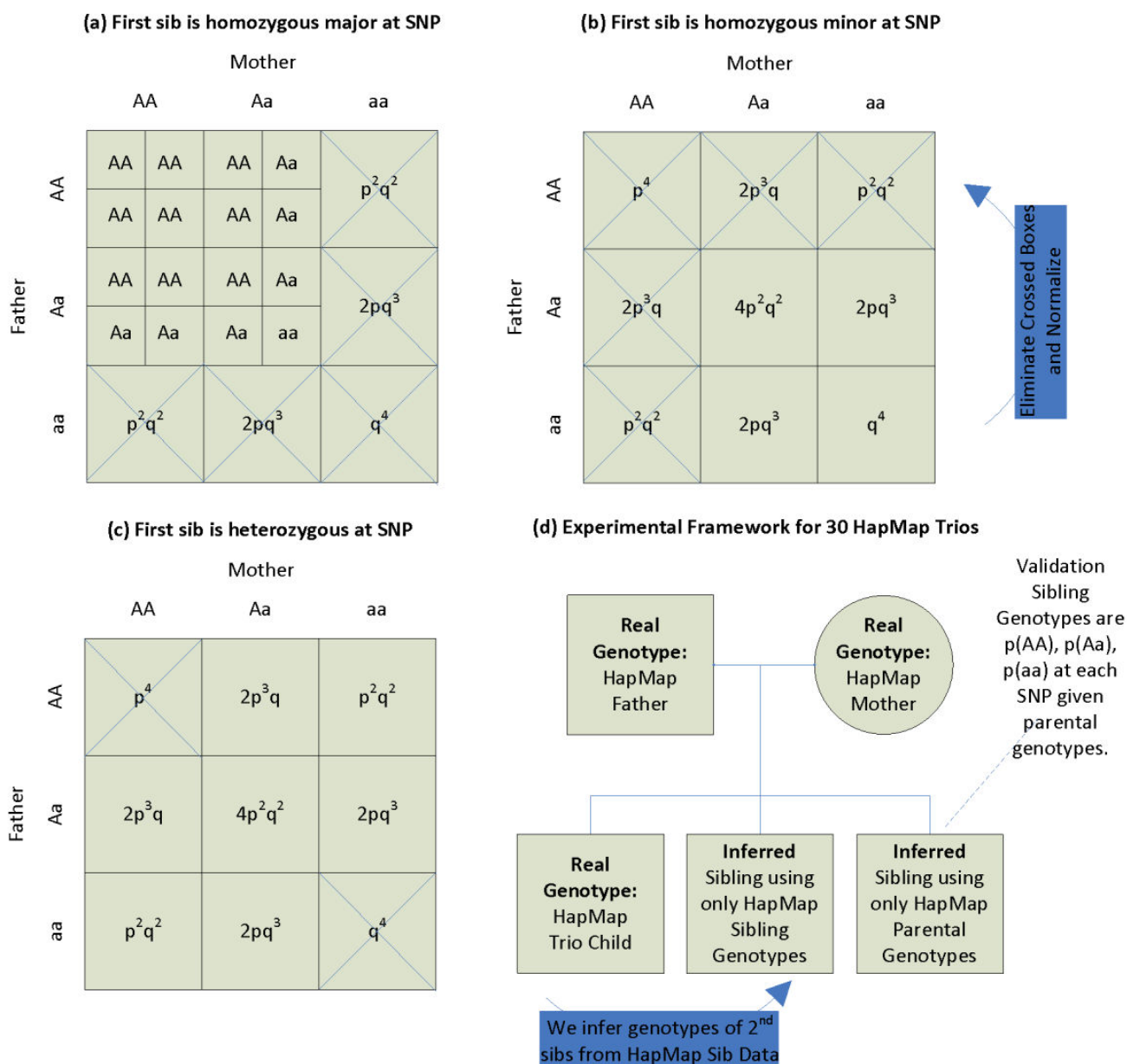
$$= \sum_{i=1}^9 \frac{p(\text{Sib}_2\text{AA} \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} p(\text{parental comb. } i \mid \text{Sib}_1\text{AA})$$

where Sib_1AA and Sib_2AA refer to Sib_1 and Sib_2 genotypes 'AA' at a selected SNP, respectively.

With unknown parental genotypes, we would calculate $p(\text{Sib}_2\text{AA})$ considering all nine possible parental genotype combinations, but knowledge that Sib_1 has genotype 'AA' allows exclusion of any parental combinations where either parent has genotype 'aa', as that would require the transmission of at least one copy of the 'a' allele to Sib_1 , if non-paternity and new mutations are excluded. HapMap SNP population frequencies, p and q , for each selected SNP, can be used to calculate the probabilities of each parental combination, i . Once these values have been calculated, the genotype of the first sibling eliminates possible parental genotypic candidates (Figs. 1A–C), and the remaining probabilities are normalized.

Measuring the information content of Sibling genotype data

When calculating the probability of a specific Sib_2 genotype given a known Sib_1 genotype, it is possible to directly measure the benefit of the proband genotype information in improving Sib_2 inferences. This involves measuring the difference between the prior Hardy-Weinberg probability for the genotype, given only population frequencies, and the posterior probability, as calculated by the conditional expression above. To measure the information content provided by the first sibling's genotype, we propose the use of a likelihood ratio test statistic, comparing models where two individuals are known to be siblings versus two individuals that are known to be unrelated. There are a total of nine possible likelihood ratios, $\Lambda_{\text{Ind}_1, \text{Ind}_2 \text{ genotypes'}}$ for each of the possible individual genotypic combinations, such as $\text{Ind}_1 \text{AA}$:

**Figure 1**

(a-c) Refining mechanism for homozygous major SNPs: when the first sibling is homozygous major (a), homozygous minor (b), or heterozygous (c) at a given SNP, this constrains the possible parental genotypes; in the first case, five of nine parental genotypic combinations can be eliminated (crossed boxes). Using HapMap CEPH SNP population frequencies, p and q , the probability frequencies are populated for the remaining squares, and normalized. The probability that subsequent sibs will be homozygous major, heterozygous, or homozygous minor can then be calculated using the probabilities that parents would contribute specific allelic values. (d) For each of 30 HapMap CEPH trios, the Sib₁ genotype and the SNP population frequencies are used (without the parent genotypes) to infer $p('AA')$, $p('Aa')$, and $p('aa')$ for subsequent siblings. Those probabilities are then validated against those that would be expected given only the parental genotypes at each SNP.

$$\begin{aligned}
\Lambda_{Ind_1, Ind_2 \text{ genotypes}} &= \frac{p(Ind_2 \text{ genotype} | Ind_1 \text{ genotype} \cap \text{siblings})}{p(Ind_2 \text{ genotype} | Ind_1 \text{ genotype} \cap \text{unrelated})} \\
&= \frac{p(Sib_2 \text{ genotype} | Sib_1 \text{ genotype} \cap \text{siblings})}{\left(\frac{p(Ind_2 \text{ genotype} \cap p(Ind_1 \text{ genotype} \cap \text{unrelated}))}{p(Ind_1 \text{ genotype} \cap \text{unrelated})} \right)} \\
&= \frac{\sum_{i=1}^9 p(Sib_2 \text{ genotype} | \text{parental comb. } i) p(\text{parental comb. } i | Sib_1 \text{ genotype})}{\left(\frac{p(Ind_2 \text{ genotype} \cap p(Ind_1 \text{ genotype} \cap \text{unrelated}))}{p(Ind_1 \text{ genotype} \cap \text{unrelated})} \right)} \\
&= \frac{\sum_{i=1}^9 \frac{p(Sib_2 \text{ genotype} \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} p(\text{parental comb. } i | Sib_1 \text{ genotype})}{\left(\frac{p(Ind_2 \text{ genotype}) \cdot p(Ind_1 \text{ genotype}) \cdot (1 - \frac{1}{N})}{p(Ind_1 \text{ genotype}) \cdot (1 - \frac{1}{N})} \right)} \\
&\equiv \frac{\sum_{i=1}^9 \frac{p(Sib_2 \text{ genotype} \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} p(\text{parental comb. } i | Sib_1 \text{ genotype})}{p(Ind_2 \text{ genotype})}
\end{aligned}$$

The denominator becomes $p(Ind_2 \text{ genotype})$, which is either p^2 , $2pq$, or q^2 . This is intuitive; when considering two unrelated individuals, the probability that the 2nd has a specific genotype can only be identified using the population frequencies for that genotype. The numerator is the posterior probability expression derived in Table 1, also in terms of p and q . The log of this odds ratio can then be used as a statistic for measuring relatedness, depending only on the SNP allele frequency and the Sib₁ genotype (Fig. 2).

The allele frequency, p , that maximizes this statistic can then be found numerically for each $\Lambda_{Ind_1, Ind_2 \text{ genotypes}}$ expression, to identify which allele frequencies and conditions are most informative for genotypic inferences. These results are below in Table 2.

Confirming sib-ship with two non-matching sets of SNP genotypes

The above inference technique can be extended to confirm sib-ship in two non-matching samples of SNP

sequence data. Given a set of matches at M independent loci from a pool of N individuals, an expanded form of Bayes Theorem can be used to calculate $p(\text{sibs} | \text{match at } M \text{ loci})$ directly:

$$\begin{aligned}
p(\text{sibs} | \text{match at } M \text{ loci}) &= \frac{p(\text{match at } M \text{ loci} | \text{sibs}) p(\text{sibs})}{p(\text{match at } M \text{ loci} | \text{sibs}) p(\text{sibs}) + p(\text{match at } M \text{ loci} | \text{sibs}) p(! \text{ sibs})} \\
&= \frac{[p(\text{both AA} | \text{sibs}) + p(\text{both Aa} | \text{sibs}) + p(\text{both aa} | \text{sibs})]^M \left(\frac{1}{N}\right)}{[p(\text{both AA} | \text{sibs}) + p(\text{both Aa} | \text{sibs}) + p(\text{both aa} | \text{sibs})]^M \left(\frac{1}{N}\right) + p(\text{match} | ! \text{sibs})^M \left(1 - \frac{1}{N}\right)}
\end{aligned}$$

$p(\text{match} | ! \text{sibs})$ can be calculated for each SNP using the population frequency; it is the probability that two unrelated individuals in the population would share the same genotype, 'AA', 'Aa', or 'aa'. The expression $p(\text{match} | ! \text{sibs})$ is effectively the same as $p(\text{match})$ as long as the sample pool, N , is large enough, as the probability of sib-ship is very low in a large pool. For three different pool sizes, ($N = 100,000; 10,000,000; 6,000,000,000$), we have created a sib-ship probability surface that varies with the number of matched SNPs and MAF of those SNPs (Fig. 3) and published supporting values for these probabilities in Table 3. For SNPs that commonly vary in the population, a small number of genotypic matches are required to confirm sib-ship.

Modeling a series of SNP inferences using a binomial distribution

A binomial distribution can be used to represent a series of sibling genotypic inferences, such as the probability of correct inferences at 50 SNP loci, if each inference meets specific criteria. Independent inferences can be treated as a random variable with probability p of success, as long as independent SNPs are selected, with the same MAF and Sib₁ genotype.

$$p(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $p(k, n, p)$ refers to the probability that k correct inferences were made out of n attempted inferences when the probability of success for each inference attempt is p .

Table 1: Sib₂ inference error reduction when Sib₁ genotype is known.

Sib ₂	Sib ₁	Prior Prob.	Posterior Prob.	Error Reduction
AA	AA	p^2	$p^2 + pq + 1/4q^2$	$ p^2 - [p^2 + pq + 1/4q^2] $
Aa	AA	$2pq$	$pq + 1/2q^2$	$ 2pq - [pq + 1/2q^2] $
aa	AA	q^2	$1/4q^2$	$ q^2 - [1/4q^2] $
AA	Aa	p^2	$1/2p^2 + 1/4pq$	$ p^2 - [1/2p^2 + 1/4pq] $
Aa	Aa	$2pq$	$1/2p^2 + (2/3) \cdot pq + 1/2q^2$	$ 2pq - [1/2p^2 + (2/3) \cdot pq + 1/2q^2] $
aa	Aa	q^2	$1/4pq + 1/2q^2$	$ q^2 - [1/4pq + 1/2q^2] $
AA	Aa	p^2	$1/4p^2$	$ p^2 - [1/4p^2] $
Aa	Aa	$2pq$	$1/2p^2 + pq$	$ 2pq - [1/2p^2 + pq] $
aa	Aa	q^2	$1/4p^2 + pq + q^2$	$ q^2 - [1/4p^2 + pq + q^2] $

The error reduction depends only on the allele frequencies, and at all frequencies, the error is reduced, improving the quality of genotypic inference.

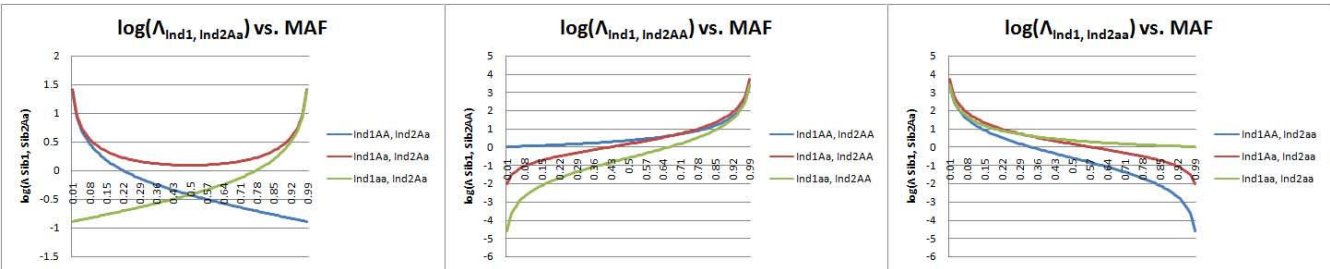


Figure 2
Log likelihood ratio test statistic for sibling inferences: for each Sib_1 genotype, the log likelihood ratio for each possible Sib_2 inference is shown versus MAF. These charts describe how informative the Sib_1 genotype is when inferring each Sib_2 genotype.

This measure will enable those who attempt to infer SNP genotypes to calculate the probability of matching at a subset of independent SNPs.

The cumulative binomial measures the probability of reaching up to k successes in n trials with probability p of success at each attempt:

$$F(k; n, p) = P(X \leq k) = \sum_{j=0}^k \binom{n}{j} p^j (1-p)^{n-j}$$

If n guesses are considered (i.e. n SNPs are genotyped and used for sib inference), $F(k, n, p)$ is the probability that at least k of those will be correct.

Results
Validation of SNP genotype inference using HapMap trio data

We then empirically infer sibling genotypic sequences from HapMap trio child genotypes using the above technique. At 700,000 SNP loci on chromosomes 2, 4, and 7, in each of 30 HapMap CEPH trios, the trio sibling, Sib_1 ,

known genotypes are combined with the CEPH and global HapMap SNP allele frequencies to produce genotypic inferences of a hypothetical sib, Sib_2 , at these loci. The inference method produces three genotypic probabilities for Sib_2 (or subsequent siblings): $p(Sib_2AA|Sib_1 genotype)$, $p(Sib_2Aa|Sib_1 genotype)$, and $p(Sib_2aa|Sib_1 genotype)$ for each SNP, which we call the SNP probability vector.

The ability to correctly infer a sibling genotype from a trio child genotype can be validated by comparing whether the best estimated genotype, using only the sibling genotype and population frequencies, matches the best estimated genotype using the parental genotypic data (Fig. 1D). We do this by comparing the plural, largest, value in the SNP probability vector, with the plural value in the SNP probability vector that would be expected given the parental genotypes and Mendelian Inheritance. The fraction of correct inferences for SNPs where the Sib_1 is homozygous major or heterozygous versus MAF are graphed in Figs. 4A–B, respectively. There were insufficient SNPs where the trio child was homozygous minor, so they have been excluded from this analysis. The appendix contains details about the HapMap population used as well as the distance and scoring metric used.

Table 2: Finding the MAF that maximizes the log likelihood ratio test statistic for each Sib_2 genotypic inference type.

Sib_2	Sib_1	Maximizing MAF	$\text{Log}(\Lambda_{\text{ind1, ind2 genotypes}})$
AA	AA	0.01	3.407
Aa	AA	0.01	3.699
aa	AA	0.01	3.389
AA	Aa	0.99	1.396
Aa	Aa	0.01, 0.99	1.407
aa	Aa	0.01	1.396
AA	aa	0.99	3.389
Aa	aa	0.99	3.699
aa	aa	0.99	3.407

The *maximizing MAF* is the allele population frequency at which the most information will be derived about the Sib_2 genotype from Sib_1 under that Sib genotypic combination. Note: There are two equally maximizing MAF values for $\text{Log}(\Lambda_{Sib1Aa, Sib2Aa})$, 0.01 and 0.99, both resulting in a value of 1.407.

For inferences at SNPs where the trio child, Sib_1 , was homozygous major, with $MAF < 0.05$ ($N = 300512, 43.2\%$), we are able to correctly infer the genotype of other siblings, e.g. Sib_2 , with 98.5% accuracy when using population-specific allele frequency data. At SNPs with $MAF < 0.20$ ($N = 452684, 65.1\%$) we achieve 91.9% average accuracy. For SNPs where the first sibling is heterozygous, with $MAF > 0.20$ ($N = 125796, 18.1\%$), it is possible to infer the correct genotype of the second sibling with 57.7% average accuracy. Without Sib_1 genotypes, all inferences for homozygous major SNPs with $MAF \geq 0.33$ and heterozygous SNPs with $MAF \leq 0.33$ would be incorrect when validated against plural parental values. At these allele frequencies, as well as others, use of Sib_1 genotypes markedly improves Sib_2 inferences.

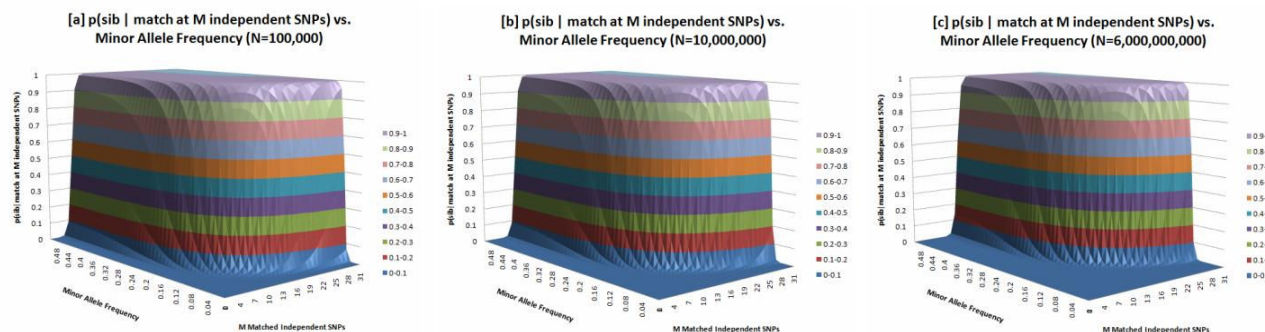


Figure 3

Sib-ship identifiability surfaces: these surfaces describe the probability of sib-ship as a function of M , the number of partially matched independent SNPs (between two individuals) and MAF. We show this across three sample size pools— $N =$ (a) 100,000; (b) 10,000,000; (c) 6,000,000,000 people. At high MAFs even very large increases in the potential sample pool size will not prevent sib-ship confirmation with relatively few matched SNPs. For example, if loci with $MAF = 0.25$ are selected, the number of matched SNPs to confirm sib-ship with $p = 0.999$ is 50 with a candidate pool of 100,000 and increases to only 80, in a group of 6 billion.

Deriving propensity to disease from sibling SNP data

Additionally, sibling SNP data can be used to quantify an individual's disease propensity through genotypic inference, without that individual's actual sequence data. For example, the likelihood ratio test statistic above may also be used to describe relative risk, using a multiplicative model.

$$\begin{aligned} \Gamma_{Sib_2 genotype | Sib_1 genotype} &= \frac{\text{probability with sibling knowledge}}{\text{probability without sibling knowledge}} \\ &= \frac{p(Sib_2 genotype | Sib_1 genotype)}{p(Sib_2 genotype)} \\ &= \frac{\sum_{i=1}^9 \frac{p(Sib_2 genotype \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} p(\text{parental comb. } i | Sib_1 genotype)}{p(Sib_2 genotype)} \end{aligned}$$

For example, the relative risk of $Sib_2 Aa$, carrying one copy of the disease allele 'a', is provided by information from the $Sib_1 aa$ genotype:

$$\begin{aligned} \Gamma_{Aa | Sib_1 aa} &= \frac{p(Sib_2 Aa | Sib_1 aa)}{p(Sib_2 Aa)} \\ &= \frac{\frac{1}{2}p^2 + pq}{2pq} \\ &= \frac{\frac{1}{2}p + (1-p)}{2(1-p)} \\ &= \frac{1 - \frac{1}{2}p}{2 - 2p} \end{aligned}$$

In this example, at $MAF = 0.01$, the relative risk of genotype 'Aa' is 25.25, given information that Sib_1 carries genotype 'aa' at that locus. However, at $MAF = 0.5$, the relative

risk of genotype 'Aa' is 0.75, given information that Sib_1 carries genotype 'aa', explaining that the risk of having the genotype 'Aa' is reduced at this MAF. This may seem counterintuitive, as the risk of carrying a disease allele is actually higher at this MAF, but Sib_2 carrying genotype 'Aa' is lower than in the control population, while the relative risk of carrying the disease allele with genotype 'aa' is higher.

$$\begin{aligned} \Gamma_{aa | Sib_1 aa} &= \frac{p(Sib_2 aa | Sib_1 aa)}{p(Sib_2 aa)} \\ &= \frac{\frac{1}{4}p^2 + pq + q^2}{q^2} \\ &= \frac{\frac{1}{4}p^2 + p(1-p) + (1-p)^2}{(1-p)^2} \end{aligned}$$

At $MAF 0.5$, $\Gamma_{aa | Sib_1 aa}$ is 2.25, demonstrating that it is more likely that a disease allele will be carried by Sib_2 in genotype 'aa' than in the control population given the Sib_1 genotype.

The explicit probability of developing a disease is also altered. If an individual with genotype 'Aa' at a specific locus has a probability p_d of developing a disease by age a , and that individual has a probability p_s of having that genotype given his sibling's genotype at that locus, his probability of developing that disease by age a is $p_s \cdot p_d$. This can easily be extended to multiple independent loci, important for diseases in which a set of common or rare variants dictates disease likelihood [12, 13]. As SNPs are both clinically informative and there is a wealth of supporting allele frequency data, they have been the focus of our analysis,

Table 3: Probability of sib-ship for three pool sizes.

N = 100,000										
q	M = 1	M = 10	M = 20	M = 30	M = 40	M = 50	M = 60	M = 70	M = 80	M = 90
0	0.00001	1.00E-05	1.00E-05	1.00E-05	1.00E-05	1.00E-05	1.00E-05	1.00E-05	1.00E-05	1.00E-05
0.05	1.10E-05	2.67E-05	7.11E-05	0.000189	0.000505	0.001345	0.003578	0.009482	0.024886	0.063706
0.1	1.21E-05	6.64E-05	0.000441	0.002923	0.019099	0.114527	0.462126	0.850907	0.974301	0.996045
0.15	1.31E-05	0.000148	0.002194	0.031572	0.325877	0.87757	0.990679	0.999366	0.999957	0.999997
0.2	1.40E-05	0.000287	0.008152	0.190701	0.871059	0.994863	0.99982	0.999994		
0.25	1.47E-05	0.000472	0.021816	0.512966	0.980292	0.999574	0.999991			
0.3	1.52E-05	0.000666	0.042483	0.747176	0.994946	0.999924	0.999999			
0.35	1.55E-05	0.000823	0.063574	0.848341	0.997835	0.999974				
0.4	1.57E-05	0.000924	0.078846	0.88788	0.998637	0.999985				
0.45	1.58E-05	0.000975	0.086919	0.902796	0.998898	0.999989				
0.5	1.58E-05	0.000989	0.089295	0.906621	0.998961	0.999989				
N = 10,000,000										
q	M = 1	M = 10	M = 20	M = 30	M = 40	M = 50	M = 60	M = 70	M = 80	M = 90
0	1.00E-07	1.00E-07	1.00E-07	1.00E-07	1.00E-07	1.00E-07	1.00E-07	1.00E-07	1.00E-07	1.00E-07
0.05	1.10E-07	2.67E-07	7.11E-07	1.89E-06	5.05E-06	1.35E-05	3.59E-05	9.57E-05	0.000255	0.00068
0.1	1.21E-07	6.64E-07	4.41E-06	2.93E-05	0.000195	0.001292	0.008518	0.053991	0.274896	0.715775
0.15	1.31E-07	1.48E-06	2.20E-05	0.000326	0.004811	0.066884	0.515231	0.940333	0.995739	0.999711
0.2	1.40E-07	2.87E-06	8.22E-05	0.002351	0.063279	0.659483	0.982308	0.999372	0.999978	0.999999
0.25	1.47E-07	4.72E-06	0.000223	0.010423	0.332172	0.959166	0.999099	0.999981		
0.3	1.52E-07	6.66E-06	0.000443	0.028705	0.663129	0.992431	0.999886	0.999998		
0.35	1.55E-07	8.24E-06	0.000678	0.052974	0.821712	0.997374	0.999968			
0.4	1.57E-07	9.25E-06	0.000855	0.073378	0.879899	0.998527	0.999984			
0.45	1.58E-07	9.76E-06	0.000951	0.084983	0.900612	0.99887	0.999988			
0.5	1.58E-07	9.90E-06	0.00098	0.088497	0.905783	0.998951	0.999989			
N = 6,000,000,000										
q	M = 1	M = 10	M = 20	M = 30	M = 40	M = 50	M = 60	M = 70	M = 80	M = 90
0	1.60E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10
0.05	1.80E-10	4.44E-10	1.18E-09	3.16E-09	8.42E-09	2.24E-08	5.98E-08	1.60E-07	4.25E-07	1.13E-06
0.1	2.00E-10	1.11E-09	7.35E-09	4.89E-08	3.25E-07	2.16E-06	1.43E-05	9.51E-05	0.000631	0.00418
0.15	2.10E-10	2.47E-09	3.66E-08	5.43E-07	8.06E-06	0.000119	0.001768	0.025594	0.280299	0.852397
0.2	2.30E-10	4.78E-09	1.37E-07	3.93E-06	0.000113	0.003217	0.084701	0.726254	0.987023	0.999542
0.25	2.40E-10	7.87E-09	3.72E-07	1.76E-05	0.000828	0.037674	0.648979	0.988676	0.999758	0.999995
0.3	2.50E-10	1.11E-08	7.39E-07	4.93E-05	0.00327	0.179341	0.935717	0.99897	0.999985	
0.35	2.50E-10	1.37E-08	1.13E-06	9.32E-05	0.007623	0.387598	0.981185	0.999767	0.999997	
0.4	2.60E-10	1.54E-08	1.43E-06	0.000132	0.012063	0.530447	0.990523	0.999897	0.999999	
0.45	2.60E-10	1.63E-08	1.59E-06	0.000155	0.014878	0.595717	0.993092	0.999929	0.999999	
0.5	2.60E-10	1.65E-08	1.63E-06	0.000162	0.01577	0.613392	0.993675	0.999936	0.999999	

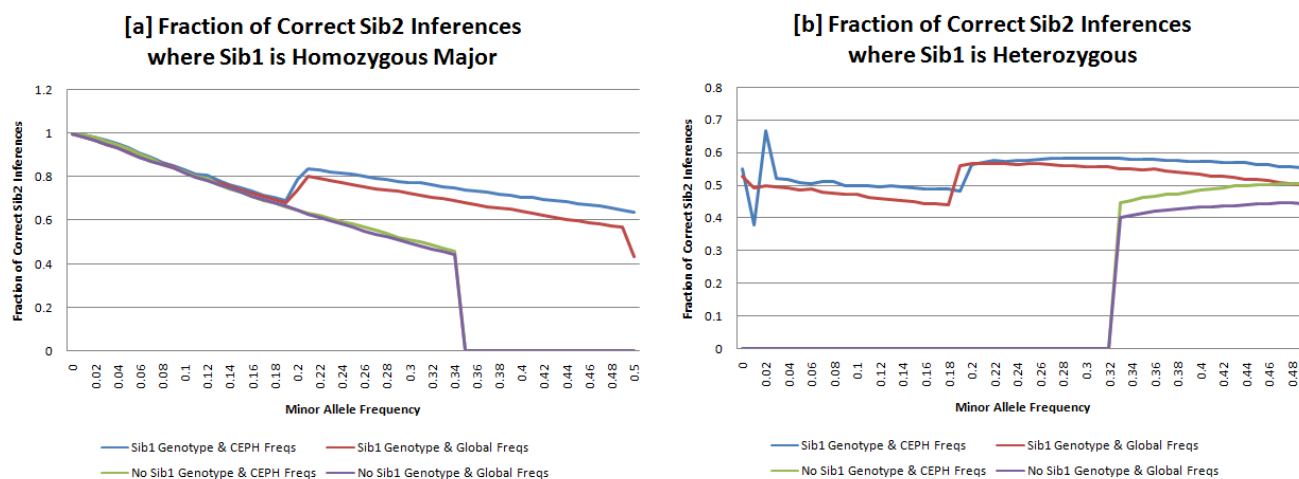
In a sample pool of size N , provided below, the probability that two individuals are siblings given a match at a subset of SNPs is charted as a function of M , the number of independent SNPs that they match at, and the minor allele frequency, q .

however there are other genomic data types which should be considered in a rigorous privacy and propensity analysis, including copy number variant and mutation data.

Discussion

These findings demonstrate that substantial discrimination and privacy concerns arise from use of inferred familial genomic data. While the Genetic Information

Nondiscrimination Act of 2008 (GINA, H.R. 493), recently passed into law, would mitigate the threat of direct discriminatory action by employers or insurers[14], there will continue to be other uses of genomic data that pose privacy risks, including the use of genetic testing in setting life, disability, and long-term care insurance premiums[15]. Familial genotypic sequences can be used to assist in forensic or criminal investigations for indirect

**Figure 4**

Fraction of correct Sib2 inferences: the fraction of Sib2 SNPs that can be correctly identified when Sib1 is (a) homozygous major or (b) heterozygous. Each line represents use of distinct data—inclusion or exclusion of Sib1 genotypes, and use of population-specific or global allele frequency data. Without Sib1 genotypes, homozygous major inferences would always be incorrect at $MAF \geq 0.33$ and heterozygous inferences would always be incorrect at $MAF \leq 0.33$. At many allele frequencies, use of Sib1 genotypes dramatically improves Sib2 inferences.

identification of genotype, increasing the number of people who may be identified[16,17]. Similarly, Freedom of Information Act (FOIA)[18] requests related to federally-funded genome wide association studies could potentially be used to identify research participants and their family members. Clinically, choosing the detail and type of disease propensity information that must be disclosed to patients and their potentially affected family members is also under debate[19,20].

Quantifying the information content of disclosed genomic data will add clarity to the informed consent process when a patient shares genotypic data for research use. For research investigations, it is conceivable that a subject would want to limit the impact of her genomic disclosure on her family members, or be asked to have a discussion with specific family members before proceeding. Providing subjects with different levels of genomic anonymity based on their sequence data, along with an estimate of the probability of re-identification and familial impact for each of those anonymity levels, will allow patients to trade off altruistically motivated sharing[21] with privacy consideration, especially when they volunteer to share all the variants in their genome[22].

While the inference accuracy rates are very high, particularly for inferences where Sib1 has a homozygous major genotype, we would like to caution that some of these findings are not always highly informative. For example, if the MAF is 0.01, where 99% of the alleles in the population are the major allele, the prior probability for a

homozygous major allele is $0.99 \times 0.99 \approx 0.98$. If Sib1 has a homozygous major allele, the posterior probability of observing a homozygous major allele in another sibling is $(1/4 + 1/4 \times 0.99 \times 0.99 + 1/2 \times 0.99) \approx 0.99$. In this case, the difference between prior and posterior probabilities is only 0.01, and knowledge of the Sib1 genotype provides very little information, as most accuracy comes from the allele frequency in the population.

However, homozygous minor alleles are much more informative. With a MAF of 0.2, if Sib1 has a homozygous minor genotype, the probability of Sib2 having the same genotype, given only the reference population is 0.04. Given that Sib1 has a homozygous minor genotype, Sib2 will have a homozygous minor allele with probability of $(1/4 + 1/4 \times 0.2 \times 0.2 + 1/2 \times 0.2) = 0.36$, which is quite different from the prior probability of 0.04.

One limitation of this study is that the population-based estimates for MAF rely on the HapMap study population sizes, which, at present, are small, though these types of sources will continue to expand. For example, the CEPH population contains 90 participants, so each trio child contributes $1/90^{\text{th}}$ of the allele frequency data used in the study. This approach also depends on the independence of the loci considered, and would need to be adapted for SNPs that are in linkage disequilibrium. Extending this study to include linked SNP loci is possible, using the haplotype block information for HapMap populations that is available. To ensure that SNPs are independent, linkage data from the HapMap population can be used to confirm

independence, and SNPs that are far from one another may be selected. Additionally, this approach does not consider the possibility of genotypic errors, which may be common on some platforms. An adjustment using a binomial probability distribution could be used to account for possible errors.

Conclusion

Technologies for sequencing large numbers of SNPs are rapidly dropping in cost, which will help realize the promise of personalized medicine, but pose substantial personal and familial privacy risks. While electronic storage and transmission of genetic tests is not yet a common component of medical record data, these tests will soon be stored in electronic medical records and personally controlled health records[23]. This mandates the need for improved informed consent models and access control mechanisms for genomic data. The increasingly common practice of electronically publishing research-related SNP data requires a delicate balance between the enormous potential benefits of shared genomic data through NCBI and other resources, and the privacy rights of both sequenced individuals and their family members.

Competing interests

The authors declare there are no competing interests.

Authors' contributions

CC conceived of the study design, carried out the statistical analysis, generated the figures, and drafted the manuscript. BS carried out experiments using HapMap data and imputed family data. KM helped draft and revise the manuscript, and helped perform the statistical analysis. ZK assisted in conception of the study and critical review of the manuscript. All authors read and approved the final manuscript.

Appendix

HapMap CEPH and global population SNP genotypes and allele frequency data

The demographic data used in this project are population-specific SNP allele frequencies from the CEPH HapMap population, Utah residents with ancestry from northern and western Europe, and the global SNP allele frequencies (from all populations that participated in the HapMap)[10]. The HapMap project has compiled allele frequency values for a large selection of SNPs – loci in the genome that account for a great deal of genetic variability in populations. Within the CEPH population, there are 30 familial trios, each containing one mother, father, and child. Additionally, the individual genotypes of the 90 CEPH trio participants are directly used in this study. One limitation of this population specific allele frequency database is the small size of each HapMap population – the CEPH population contains 90 participants, and as

such, each trio child contributes 1/90th of the allele frequency data that are used in the study.

Inferring sibling genotypic sequences from HapMap trio children

Here, we explore a specific example of sibling genotypic inference in greater depth, considering the case where one sibling's genotype is known to be 'AA', and the goal is to determine the probability that the second sibling's genotype will also be 'AA' at that locus. The conditional probability expression that sums over the nine possible parental genotypic combinations (for example, maternal genotype 'Aa' with paternal genotype 'AA') at a single SNP, with each specific parental genotypic combination denoted as i can be used:

$$p(\text{Sib}_2\text{AA} \mid \text{Sib}_1\text{AA}) = \sum_{i=1}^9 p(\text{Sib}_2\text{AA} \mid \text{parental comb. } i) p(\text{parental comb. } i \mid \text{Sib}_1\text{AA})$$

$$= \sum_{i=1}^9 \frac{p(\text{Sib}_2\text{AA} \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} p(\text{parental comb. } i \mid \text{Sib}_1\text{AA})$$

where Sib_1AA and Sib_2AA refer to Sib_1 and Sib_2 genotypes 'AA' at a selected SNP, respectively.

With unknown parental genotypes, we would calculate $p(\text{Sib}_2\text{AA})$ considering all nine possible parental genotype combinations, but knowledge that Sib_1 has genotype 'AA' allows exclusion of any parental combinations where either parent has genotype 'aa', as that would require the transmission of at least one copy of the 'a' allele to Sib_1 , if non-paternity and new mutations are excluded.

For example, when the child is homozygous major, all possible parental genotypic candidates that involve one or both parent genotypes of 'aa' are excluded, as it is not possible to have a child with genotype 'AA' if either parent does not have at least one copy of the 'A' allele. In this case, there are four possible parental genotypic combinations:

$$= \sum_{i=1}^4 \frac{p(\text{Sib}_2\text{AA} \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} p(\text{parental comb. } i \mid \text{Sib}_1\text{AA})$$

$$= \left(\frac{p(\text{Sib}_2\text{AA} \cap \text{AA}_{\text{M}}\text{AA}_{\text{F}})}{p(\text{AA}_{\text{M}}\text{AA}_{\text{F}})} \right) p(\text{AA}_{\text{M}}\text{AA}_{\text{F}} \mid \text{Sib}_1\text{AA}) + \left(\frac{p(\text{Sib}_2\text{AA} \cap \text{AA}_{\text{M}}\text{Aa}_{\text{F}})}{p(\text{AA}_{\text{M}}\text{Aa}_{\text{F}})} \right) p(\text{AA}_{\text{M}}\text{Aa}_{\text{F}} \mid \text{Sib}_1\text{AA})$$

$$+ \left(\frac{p(\text{Sib}_2\text{AA} \cap \text{Aa}_{\text{M}}\text{AA}_{\text{F}})}{p(\text{Aa}_{\text{M}}\text{AA}_{\text{F}})} \right) p(\text{Aa}_{\text{M}}\text{AA}_{\text{F}} \mid \text{Sib}_1\text{AA}) + \left(\frac{p(\text{Sib}_2\text{AA} \cap \text{Aa}_{\text{M}}\text{Aa}_{\text{F}})}{p(\text{Aa}_{\text{M}}\text{Aa}_{\text{F}})} \right) p(\text{Aa}_{\text{M}}\text{Aa}_{\text{F}} \mid \text{Sib}_1\text{AA})$$

$$= (1)(p^2) + \left(\frac{1}{2} \right) (pq) + \left(\frac{1}{2} \right) (pq) + \left(\frac{1}{4} \right) (q^2)$$

$$= p^2 + pq + \frac{q^2}{4}$$

$$= p^2 \left[1 + \frac{pq}{p^2} + \frac{q^2}{4p^2} \right]$$

which allows calculation directly from the SNP population frequencies. Before knowledge of the Sib_1 genotype was used, $p(\text{Sib}_2\text{AA})$ would have been the Hardy-Weinberg frequency for major homozygotes, p^2 . However, with the Sib_1 genotype, $p(\text{Sib}_2\text{AA} \mid \text{Sib}_1\text{AA})$, the additional con-

straint increases the probability to $p^2 + pq + (q^2/4)$, increasing inference accuracy by $pq + (q^2/4)$.

The remaining entries in the probability vector, $p(\text{Sib}_2\text{Aa}|\text{Sib}_1\text{AA})$, and $p(\text{Sib}_2\text{aa}|\text{Sib}_1\text{AA})$, can then be calculated just as we have done for $p(\text{Sib}_2\text{AA}|\text{Sib}_1\text{AA})$ above. Again, these probabilities have been generated without any actual knowledge of the parent genotypes. If the Sib_1 genotype were instead 'Aa' or 'aa', the above technique can similarly be used (with a different combination of possible parental genotypes) to calculate the two other probability vectors, $[p(\text{Sib}_2\text{AA}|\text{Sib}_1\text{Aa}), p(\text{Sib}_2\text{Aa}|\text{Sib}_1\text{Aa}), p(\text{Sib}_2\text{aa}|\text{Sib}_1\text{Aa})]$ and $[p(\text{Sib}_2\text{AA}|\text{Sib}_1\text{aa}), p(\text{Sib}_2\text{Aa}|\text{Sib}_1\text{aa}), p(\text{Sib}_2\text{aa}|\text{Sib}_1\text{aa})]$.

Validating the sibling genotype probability vector using parental genotypic data

To validate the results of the refining strategy on inferring the second sibling genotype, the authentic parental genotypes are used to create the probability vector $p(\text{'AA'}), p(\text{'Aa'}), p(\text{'aa'})$ at the SNP being evaluated, for the children the pair would be expected to have. For each of the trio pairs at each of the SNPs being tested, the probability vector was calculated.

Error reduction calculation

The error reduction measurement identifies the extent to which inference error is reduced. For example, where we are trying to infer the probability that Sib_2 has genotype 'AA' at a specific SNP, we calculate the absolute value of the difference between our best inference and the Hardy Weinberg probability for Sib_2 to have genotype 'AA', using population-specific allele frequency data and the Sib_1 genotype, $|p(\text{Sib}_2\text{AA}|\text{Sib}_1\text{genotype}) - p(\text{Sib}_2\text{AA})|$. This value is specifically the improvement to the probability value from the new data, when inferring the specific event that Sib_2 will have genotype 'AA' and Sib_1 will have the specific genotype in question.

Any change to $p(\text{Sib}_2\text{AA})$ must also correspond with the opposite change in the sum of $p(\text{Sib}_2\text{Aa})$ and $p(\text{Sib}_2\text{aa})$. To accurately represent the overall error reduction by Sib_1 genotype, with any of three possible Sib_2 genotypes, the average of the three values is measured. For example, where the Sib_1 genotype is 'AA', the overall average improvement (and error reduction) is the average of $|p(\text{Sib}_2\text{AA}) - p(\text{Sib}_2\text{AA}|\text{Sib}_1\text{AA})|$, $|p(\text{Sib}_2\text{Aa}) - p(\text{Sib}_2\text{Aa}|\text{Sib}_1\text{AA})|$, and $|p(\text{Sib}_2\text{aa}) - p(\text{Sib}_2\text{aa}|\text{Sib}_1\text{AA})|$.

Scoring metric for calculating correct fraction of inferences

To ascertain whether the inferences are helpful for producing correct answers, a scoring metric was used to calculate the fraction of correct SNP inferences, in our empirical inference validation study. For each SNP inference, the

scoring metric provides a full point when the plural entry in the inference vector, (the maximum of $p(\text{'AA'}), p(\text{'Aa'}),$ and $p(\text{'aa'})$, and thus the predicted sib genotype), matches the plural entry in the parental validation vector (the empirical most likely genotype). Given the parental genotype values, it is possible, and not infrequent, that a validation probability vector has two matching plural values, for example, if $p(\text{'AA'}) = p(\text{'Aa'}) = 0.5$. When this is the case, one half point was awarded if the plural value in the inference vector matched one of the two validation choices, to signify that one of the two equally likely candidates was chosen.

There are some conditions that arise from use of a simple scoring metric, where it becomes difficult to score well. For example, a heterozygous Sib_1 will likely result in a 0.5 score for inferences. A score of 1 point would be possible if one parent had a genotype of 'AA' and the other had genotype 'aa', making the probability that the parents would have a child with genotype 'Aa' equal 1. Most remaining parental combinations would not result in the probability of child genotype 'Aa' equal to 1, and would likely result in only a half point. These values can be adjusted using machine learning techniques or more robust decision making algorithms, but those are out of the scope of this work.

Acknowledgements

The authors would like to gratefully acknowledge the assistance of Dr. John Tsitsiklis and Dr. Shannon Wieland for discussion of probabilistic techniques and support from the National Library of Medicine, National Institutes of Health grant R01-LM009375-01A1.

References

- Adida B, Kohane IS: **GenePING: secure, scalable management of personal genomic data.** *BMC Genomics* 2006, **7**:93.
- Hoffman MA: **The genome-enabled electronic medical record.** *J Biomed Inform* 2007, **40**(1):44-46.
- Kaiser J: **Genomic databases. NIH goes after whole genome in search of disease genes.** *Science* 2006, **311**(5763):933.
- Thomas DC: **Are we ready for genome-wide association studies?** *Cancer Epidemiol Biomarkers Prev* 2006, **15**(4):595-598.
- Lin Z, Owen AB, Altman RB: **Genetics. Genomic research and human subject privacy.** *Science* 2004, **305**(5681):183.
- Malin BA, Sweeney LA: **Inferring genotype from clinical phenotype through a knowledge based algorithm.** *Pac Symp Biocomput* 2002:41-52.
- Lowrance WW, Collins FS: **Ethics. Identifiability in genomic research.** *Science* 2007, **317**(5838):600-602.
- Malin BA: **An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future.** *J Am Med Inform Assoc* 2005, **12**(1):28-34.
- Brenner CH, Weir BS: **Issues and strategies in the DNA identification of World Trade Center victims.** *Theor Popul Biol* 2003, **63**(3):173-178.
- A haplotype map of the human genome. *Nature* 2005, **437**(7063):1299-1320.
- Olivier M: **A haplotype map of the human genome.** *Physiol Genomics* 2003, **13**(1):3-9.
- Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nat Rev Genet* 2005, **6**(2):95-108.

13. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH: **Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia.** *Nat Genet* 2005, **37(4)**:435-440.
14. Holden C: **Genetic discrimination. Long-awaited genetic non-discrimination bill headed for easy passage.** *Science* 2007, **316(5825)**:676.
15. Hudson KL, Holohan MK, Collins FS: **Keeping pace with the times--the Genetic Information Nondiscrimination Act of 2008.** *N Engl J Med* 2008/06/21 edition. 2008, **358(25)**:2661-2663.
16. Bieber FR, Brenner CH, Lazer D: **Human genetics. Finding criminals through DNA of their relatives.** *Science* 2006, **312(5778)**:1315-1316.
17. Bieber FR, Lazer D: **Guilt by association: should the law be able to use one person's DNA to carry out surveillance on their family? Not without a public debate.** *New Sci* 2004, **184(2470)**:20.
18. **Freedom of Information Act . 5 USC 552 1996.**
19. Kohut K, Manno M, Gallinger S, Esplen MJ: **Should healthcare providers have a duty to warn family members of individuals with an HNPCC-causing mutation? A survey of patients from the Ontario Familial Colon Cancer Registry.** *J Med Genet* 2007, **44(6)**:404-407.
20. Offit K, Groeger E, Turner S, Wadsworth EA, Weiser MA: **The "duty to warn" a patient's family members about hereditary disease risks.** *Jama* 2004, **292(12)**:1469-1473.
21. Kohane IS, Altman RB: **Health-information altruists--a potentially critical resource.** *N Engl J Med* 2005, **353(19)**:2074-2077.
22. Church GM: **The personal genome project.** *Mol Syst Biol* 2005, **1**:2005 0030.
23. Simons WW, Mandl KD, Kohane IS: **The PING personally controlled electronic medical record system: technical architecture.** *J Am Med Inform Assoc* 2005, **12(1)**:47-54.

Pre-publication history

The pre-publication history for this paper can be accessed here:

<http://www.biomedcentral.com/1755-8794/1/32/prepub>

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

