ABSTRACT

Summary: Three recent publications have examined the quality and completeness of public database single nucleotide polymorphism (dbSNP) and have come to dramatically different conclusions regarding dbSNPs false positive rate and the proportion of dbSNPs that are expected to be common. These studies employed different genotyping technologies and different protocols in determining minimum acceptable genotyping quality thresholds. Because heterozygous sites typically have lower quality scores than homozygous sites, a higher minimum quality threshold reduces the number of false positive SNPs, but yields fewer heterozygotes and leads to fewer confirmed SNPs. To account for the different confirmation rates and distributions of minor allele frequencies, we propose that the three confirmation studies have different false positive and false negative rates. We developed a mathematical model to predict SNP confirmation rates and the apparent distribution of minor allele frequencies under user-specified false positive and false negative rates. We applied this model to the three published studies and to our own resequencing effort. We conclude that the dbSNP false positive rate is \( \sim 15\% - 17\% \) and that the reported confirmation studies have vastly different genotyping error rates and patterns.

Contact: dcutler@jhmi.edu

INTRODUCTION

Now that the final draft of the human genome reference sequence has been completed, the next major challenge facing the human genomics community is to harness this resource to help understand complex disease (Chakravarti, 1999). To this end, an international consortium has formed to describe systematically and catalog human genetic variation (Couzin, 2002; Wade, 2002). Because this multi-year, multi-million dollar project and numerous other smaller disease mapping projects begin their work by genotyping single nucleotide polymorphisms (SNPs) contained in the public database dbSNP (Sachidanandam et al., 2001), the quality and completeness of dbSNP is of paramount importance. Three recent reports (Gabriel et al., 2002; Carlson et al., 2003; Reich et al., 2003) have explored dbSNP, and the conclusions drawn vary dramatically. Gabriel et al. (2002) and Reich et al. (2003) estimate dbSNP’s false positive rate at 2–10%, while Carlson et al. (2003) estimates it at 35–36%. Additionally, the former two reports predict that 73–74% of the SNPs that do confirm are common (minor allele frequency >10% ) while the latter predicts that 83–89% of dbSNPs are common. We hypothesize that these differences in confirmation rates and allele frequency distributions are due, in large part, to implicit and explicit decisions regarding acceptable cut-offs for false positives and false negatives, specifically in calling heterozygotes. In the conversion of any quantitative measure into a dichotomous outcome, a decision must be made as to the level of the threshold. A higher threshold will result in more false negatives and fewer false positives than a lower threshold. Different sequencing technologies employ different measures of quality and, even when the same technology is employed, the researcher must make individual decisions as to minimal acceptable quality scores and their implementation.

Gabriel et al. (2002) genotyped 400 chromosomes from four ethnic groups (Yoruban, African American, European and Asian) using Sequenom SpectroCHIP assays for over 2000 SNPs identified by The SNP Consortium (TSC) across 51 autosomal regions. Assays were rejected if fewer than 75% of genotype calls could be made and markers were eliminated if they violated Hardy–Weinberg equilibrium. Using Mendelian inheritance checking, the genotyping error rate was estimated at 0.4%.
Carlson et al. (2003) sequenced 50 genomic regions spanning 564 kb in 47 European and African individuals using an ABI 3700, requiring a minimum Phred score (Ewing and Green, 1998) of 25. Polymorphic sites were first identified using PolyPhred (Nickerson et al., 1997), and then genotype calls were individually observed by analysts who accepted or rejected the PolyPhred call based on personal inspection of the trace reads. Fifty-nine of the highest minor allele frequency sites were later re-genotyped using Taqman assays, and an error rate of 0.1% for these common SNPs was estimated.

Reich et al. (2003) sequenced 173 kb across 17 loci in 150 chromosomes from individuals of European and West African ancestry using an ABI 3700, requiring a minimum Phred score of 25. Calls were reviewed and accepted or rejected by four analysts. Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry was later used to re-genotype 302 of the SNPs, and a genotyping error rate of 1% was estimated. Quality of SNP detection was broadly assessed by calculating overall heterozygosity (π) for the regions sequenced and finding it to be consistent with the expected genome-wide average.

Using custom microarrays from Affymetrix, Cutler et al. (2001) sequenced 50 kb across each of 40 loci in 80 chromosomes of African, Asian, European and native American ancestry, of approximately equal numbers (Collins et al., 1998). The minimum quality threshold was set very high, at approximately Phred score >54, to ensure a very low false positive rate. However, as a trade-off, 20% of bases could not be called. Among bases that were called, the error rate was estimated at <0.2% for segregating sites through restriction fragment length polymorphism (RFLP) genotyping, taqman genotyping and ABI sequencing of one locus.

Table 1 lists each of the confirmation studies, the number of chromosomes examined, the confirmation rate and the observed fraction of SNPs with minor allele frequency >10%. Figure 1 shows the percent of TSC and bacterial artificial chromosome (BAC) overlap SNPs found to be monomorphic and the fraction of SNPs with minor allele frequency <10% among those that confirmed in each of the four studies. Each study’s confirmation rate is significantly different from all others. The fraction of confirmed SNPs found to be rare by Cutler et al. (2001), Gabriel et al. (2002) and Reich et al. (2003) are not significantly different from one another, but each of the three is significantly different from Carlson et al. (2003). The p-values for all pairwise comparisons, obtained using Fisher’s exact test, can be found in Table 2 for TSC SNPs and BAC SNPs.

In an attempt to understand the discrepancies in confirmation rates and apparent allele frequency distributions, we created four error models, which are shown in Table 3 and described in the Methods section. Model 1, which does not allow for errors in dbSNP or in the confirmation studies, and model 2, which allows for errors in dbSNP but not in the confirmation studies, were rejected at the 0.01 level. Model 3 allows for false positives in dbSNP and for genotyping error in the confirmation studies. Parameters include ε₁, ε₂ and ε₃, where ε₁ is the false positive rate of dbSNP, ε₂ is the probability of miscalling the major allele as the minor allele in the confirmation study and ε₃ is the probability of miscalling the minor allele as the major allele in the confirmation study. In model 3, ε₃ is independent of the genotypes of other individuals in the sample. In model 4, ε₃ is replaced with ε₄, which is equal to zero if the minor allele is common enough to be found in homozygous state in the sample. If the minor allele is rare (i.e. not found in homozygous state in the sample), ε₄ can be non-zero. We allowed error rates in the following ranges: 0 ≤ ε₁ ≤ 0.4, 0 ≤ ε₂ ≤ 0.02, 0 ≤ ε₃ ≤ 0.4, 0 ≤ ε₄ ≤ 1.0 and fit all four datasets simultaneously. For the studies that examined both BAC and TSC dbSNPs (Cutler et al., 2001; Carlson et al., 2003, Reich et al., 2003), each study provides four observations: the fraction of BAC SNPs that confirm, the fraction of TSC SNPs that confirm, the fraction of BAC SNPs that are common and the fraction of TSC SNPs that are common. Gabriel et al. (2002) provides two observations, as only TSC SNPs were examined. All together,

---

**Table 1. SNP confirmation studies included in analyses**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Number of chromosomes</th>
<th>Source of SNPs</th>
<th>Number of SNPs in regions</th>
<th>Confirmation rate</th>
<th>Percentage of confirmed SNPs with MAFa &gt;10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlson et al. (2003)</td>
<td>94</td>
<td>BACb</td>
<td>510</td>
<td>0.64</td>
<td>0.89</td>
</tr>
<tr>
<td>Cutler et al. (2001)</td>
<td>80</td>
<td>TSCc</td>
<td>176</td>
<td>0.65</td>
<td>0.85</td>
</tr>
<tr>
<td>Gabriel et al. (2002)</td>
<td>380</td>
<td>TSC</td>
<td>2242</td>
<td>0.90</td>
<td>0.73</td>
</tr>
<tr>
<td>Reich et al. (2003)</td>
<td>150</td>
<td>BAC</td>
<td>69</td>
<td>0.93</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSC</td>
<td>54</td>
<td>0.98</td>
<td>0.74</td>
</tr>
</tbody>
</table>

aMinor allele frequency.

b dbSNPs identified by BAC overlap.

c dbSNPs identified by The SNP Consortium (TSC).

False positive rate of dbSNP

---

1023
there are 14 observations and 10 parameters: $\varepsilon_1$ for TSC dbSNPs, $\varepsilon_1$ for BAC dbSNPs, $\varepsilon_2$ and $\varepsilon_3$ or $\varepsilon_4$ for each confirmation study. Thus, there are four degrees of freedom. We covered the parameter space by performing a 10-dimensional grid search in increments of 0.01 for $\varepsilon_1$ and $\varepsilon_3$ or $\varepsilon_4$ and 0.001 for $\varepsilon_2$.

RESULTS

Figure 2 shows the expected minor allele frequency distribution for a 100-chromosome dbSNP confirmation study under each of the four error models for selected values of $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$ and $\varepsilon_4$. From these graphs, it is clear that different rates and patterns of error affect not only the fraction of SNPs that are expected to be monomorphic in the confirmation study, but also the expected frequency spectra of the SNPs that do confirm. Note in particular that error model 4, with a high $\varepsilon_4$, leads to a deficit of rare alleles. Lower values of $\varepsilon_4$ do not create this deficit and it is not seen with any of the other error models.

When all datasets were fit with model 1, 2 or 3, all combinations of error parameters were rejected at the 0.01 level. Using
model 4 for Carlson et al. (2003) and model 3 for all others, or using model 4 for Carlson et al. (2003) and Gabriel et al. (2002) and model 3 for Cutler et al. (2001) and Reich et al. (2003), or using model 3 for Cutler et al. (2001) and model 4 for all others, a small region of the 10-dimensional parameter space contained combinations of error rate estimates that could not be rejected at \( p = 0.01 \). Within this space, \( \epsilon_1 \) was tightly bound near 0.15 for TSC SNPs and 0.17 for BAC SNPs. All other combinations of models 3 and 4 for the four datasets were rejected at the 0.01 level. The best-fitting models and ranges of acceptable parameter values can be found in Table 4.

**DISCUSSION**

It is interesting that the predicted error rates for Carlson et al. (2003) and Reich et al. (2003) are dramatically different in spite of using the same technology and the same minimum Phred score. We believe that this is due to the low positive predictive value in calling heterozygotes associated with a minimum Phred score of 25. On average, only 15 and 11% of heterozygous PolyPhred calls for dye-primer and dye-terminator reactions, respectively, are true heterozygotes (Nickerson et al., 1997). Analysts must review each heterozygous call and determine whether or not they believe it to be true. The false positive rate can be reduced by removing heterozygous calls believed to be spurious. What is unknown and most likely varies between laboratories, however, is the fraction of true positives that are removed along with the false positives. It is likely that Carlson et al. (2003) had a higher standard for accepting heterozygous calls than Reich et al. (2003) and, therefore, removed a substantial percent of their true heterozygotes along with the false heterozygotes.

It is important to understand that for the sake of simplicity we have not explicitly modeled missing data, but that missing data will influence our estimates of \( \epsilon_3 \) and \( \epsilon_4 \). If the minimum acceptable quality score is set very high, a disproportionate number of heterozygotes will be ‘no calls’, reducing the confirmation rate, and missing data will have the effect of increasing our estimates of \( \epsilon_3 \) and \( \epsilon_4 \). To a first approximation, these parameters can be thought of as something close to the sum of the minor allele to major allele genotyping error rate plus the rate at which the minor allele is missing data.

With this in mind, our best estimates of the error rates for the four confirmation studies are in line with the investigators’ estimates of their own error and missing data rates. Carlson et al. (2003) re-genotyped a subset of common SNPs that confirmed in their dataset, estimating their false positive rate at 0.001. We estimate Carlson’s \( \epsilon_2 \) at 0.000. Cutler et al. (2001) tuned their allele-calling algorithm to guarantee a false positive rate close to zero but obtain 10–20% missing data. This agrees with our estimates of \( \epsilon_2 \) at 0.000 and \( \epsilon_3 \) at 0.12. Gabriel et al. (2002) estimate their genotyping error rate to be less than 0.004 and claim a missing data rate <25%. We estimate their \( \epsilon_2 \) at 0.001 and their \( \epsilon_3 \) at 0.12. Finally, Reich et al. (2003) estimate a discordance rate of less than 0.01, which is in line with our estimates of their \( \epsilon_2 \) at 0.007 and \( \epsilon_3 \) at 0.000.

Note that, although the best-fitting \( \epsilon_4 \) value for Carlson et al. (2003) is 0.99, we are not asserting that the dataset has an overall 99% false negative rate. Rather, it appears that when the minor allele was rare enough not to occur in homozygous state, the allele was missed in heterozygotes 99% of the time. In model 3, \( \epsilon_3 \) errors are made regardless of genotype, hence the false negatives are in some sense spread over the entire allele frequency space. On the other hand, in model 4, errors

---

**Table 2.** P-values for pairwise comparisons of confirmation rates (lower left corner) and fraction of SNPs with minor allele frequency <10% (upper right corner) for TSC SNPs and BAC overlap SNPs in dbSNP

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TSC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlson et al. (2003)</td>
<td>—</td>
<td>0.024</td>
<td>0.004</td>
<td>0.088</td>
</tr>
<tr>
<td>Cutler et al. (2001)</td>
<td>5.75 \times 10^{-5}</td>
<td>—</td>
<td>0.429</td>
<td>0.867</td>
</tr>
<tr>
<td>Gabriel et al. (2002)</td>
<td>1.74 \times 10^{-18}</td>
<td>3.65 \times 10^{-10}</td>
<td>—</td>
<td>1.000</td>
</tr>
<tr>
<td>Reich et al. (2003)</td>
<td>6.34 \times 10^{-8}</td>
<td>2.87 \times 10^{-4}</td>
<td>0.056</td>
<td>—</td>
</tr>
<tr>
<td>BAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlson et al. (2003)</td>
<td>—</td>
<td>1.07 \times 10^{-4}</td>
<td>—</td>
<td>4.03 \times 10^{-3}</td>
</tr>
<tr>
<td>Cutler et al. (2001)</td>
<td>4.70 \times 10^{-8}</td>
<td>—</td>
<td>—</td>
<td>0.52</td>
</tr>
<tr>
<td>Reich et al. (2003)</td>
<td>1.93 \times 10^{-7}</td>
<td>2.72 \times 10^{-3}</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*P*-values were obtained using Fisher’s exact test.

**Table 3.** Error models

<table>
<thead>
<tr>
<th>Model</th>
<th>dbSNP False SNP rate</th>
<th>Confirmation study</th>
<th>Minor to major for rare minor alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Major to minor</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>( \epsilon_1 )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>( \epsilon_1 ) ( \epsilon_2 ) ( \epsilon_3 )</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>( \epsilon_1 ) ( \epsilon_2 )</td>
<td>—</td>
<td>( \epsilon_4 )</td>
</tr>
</tbody>
</table>

*A minor allele is ‘rare’, if it does not appear in homozygous state in the dataset.*
only occur in the absence of the rare homozygote. Thus, all the errors are 'compacted' into a smaller portion of frequency space. By integrating over all possible minor allele frequencies, we find that a 99% $\varepsilon_4$ translates into an overall false negative rate of $\sim 12\%$, which is comparable with the other three studies examined. Thus, it is the pattern of missing data and errors, not the magnitude of the error rate that results in the low confirmation rate and the distorted expectation of minor allele frequency distribution in this dataset, as rare alleles have likely been disproportionately missed. The quality-control measures employed by Carlson et al. (2003) would not pick up these errors, as they re-genotyped a subset of common

**Fig. 2.** Expected minor allele frequency distribution, for a 100-chromosome dbSNP confirmation study under each of our four error models. (a) Model 1: $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 0$; (b) model 2: $\varepsilon_1 = 0.15, \varepsilon_2 = \varepsilon_3 = 0$; (c) model 3: $\varepsilon_1 = 0.15, \varepsilon_2 = 0.001, \varepsilon_3 = 0.15$; (d) model 4: $\varepsilon_1 = 0.15, \varepsilon_2 = 0, \varepsilon_4 = 0.50$; (e) model 4: $\varepsilon_1 = 0.15, \varepsilon_2 = 0, \varepsilon_4 = 0.99$. 

1026
SNPs that confirmed in their dataset. This essentially measures the false positive rate (our $\varepsilon_2$), but provides no information about the false negative rate. In order to estimate directly the false negative rate, one could use a different technology to re-genotype a subset of dbSNPs that did not confirm. The fundamental idea is that different technologies have different false positive and negative rates. Using another technology, perhaps one that is inherently better at calling heterozygotes, would provide a more objective view of the SNPs that did not confirm.

Thus, we conclude that all four studies are consistent with the following view of dbSNP: $\sim15$–$17\%$ of the SNPs discovered by the major SNP discovering efforts are false positives, and $\sim80\%$ of the SNPs which are not false positive
have minor allele frequency \( > 10\% \). Published dbSNP confirmation studies appear to show divergent patterns and rates of genotyping error and missing data.

**METHODS**

We have derived the probability that the minor allele at an SNP will be observed in \( k \) out of \( n \) chromosomes examined in a confirmation study, given that it appears in dbSNP under four error models. Each model is progressively more complex. All four models assume that dbSNP was originally constructed by examining exactly two alleles.

In model 1, we assume no genotyping error in the confirmation study, and no false positives in dbSNP. In model 2, we assume no genotyping error in the confirmation study, but allow for false positives in dbSNP. We denote the fraction of dbSNP that is false as \( \varepsilon_1 \).

In model 3, there can be both false positives within dbSNP (i.e. \( \varepsilon_1 > 0 \)) and there can be genotyping error in the subsequent study. The probability that the major allele is miscalled as the minor allele is \( \varepsilon_2 \), and the probability that the minor allele is miscalled as the major allele is \( \varepsilon_3 \). By assumption, in model 3, the probability of an error in one sample is independent of the genotypes of all other samples.

The last model, model 4, is intended to model a genotype-dependent error rate. The basic assumption of the model is that calls are more accurate when all three genotypes are present, than when only one homozygote and/or heterozygotes are present. For simplicity, we assume that when all three genotypes are present, no genotyping errors occur at all. Thus, in the presence of all three genotypes, model 4 reduces to model 2. However, when only one genotype is present in the sample (all samples are homozygous for the major allele), the major allele is called the minor allele with probability \( \varepsilon_2 \) (and model 4 reduces to model 3). Finally, when two genotypes are present in the sample (homozygotes for the major allele, and heterozygotes), the minor allele is miscalled as the major allele with probability \( \varepsilon_4 \). Thus, \( \varepsilon_4 \) can be thought of as the rate of ‘dropped’ heterozygotes. This type of error model might be particularly suited for situations where researchers call alleles by eye. The notion is that when all three genotypes are present it may be ‘easier’ to detect (or believe) heterozygotes.

The derivations of all four error models can be found in the Appendix.

We identified the best-fitting models by restricting the parameter space to values for which TSC \( \varepsilon_2 \) is equal to BAC \( \varepsilon_2 \) and TSC \( \varepsilon_3 \) is equal to BAC \( \varepsilon_3 \) (or TSC \( \varepsilon_4 \) is equal to BAC \( \varepsilon_4 \) for model 4) for each of the three studies that examined both TSC and BAC dbSNPs (Cutler et al., 2001; Carlson et al., 2003; Reich et al., 2003). Holding TSC \( \varepsilon_1 \) and BAC \( \varepsilon_1 \) constant, we performed a 10-dimensional grid search to identify values of \( \varepsilon_2 \) and \( \varepsilon_3 \) or \( \varepsilon_2 \) and \( \varepsilon_4 \) that minimized

\[
T = \sum \frac{(a - b)^2}{b} + \sum \frac{(c - d)^2}{d},
\]

where \( a \) and \( b \) are the observed and expected number of SNPs that confirm in each study and \( c \) and \( d \) are the observed and
Table 4. Confirmation study error rates for best-fitting models and range of acceptable values

<table>
<thead>
<tr>
<th>Reference</th>
<th>$\epsilon_2$</th>
<th>Best estimate (range)</th>
<th>$\epsilon_3$</th>
<th>Best estimate (range)</th>
<th>$\epsilon_4$</th>
<th>Best estimate (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive rate ($\epsilon_1$) is 0.16 for TSC SNPs and 0.17 for BAC overlap SNPs ($\chi^2$ value for best-fitting model is 12.53)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlson et al. (2003)</td>
<td>0.000</td>
<td>(0.000–0.000)</td>
<td>0.12</td>
<td>(0.08–0.16)</td>
<td>0.00</td>
<td>(0.00–0.00)</td>
</tr>
<tr>
<td>Cutler et al. (2001)</td>
<td>0.000</td>
<td>(0.000–0.000)</td>
<td>0.00</td>
<td>(0.00–0.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabriel et al. (2002)</td>
<td>0.002</td>
<td>(0.002–0.002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reich et al. (2003)</td>
<td>0.007</td>
<td>(0.006–0.009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positive rate ($\epsilon_1$) is 0.15 for TSC SNPs and 0.17 for BAC overlap SNPs ($\chi^2$ value for best-fitting model is 12.96)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlson et al. (2003)</td>
<td>0.000</td>
<td>(0.000–0.000)</td>
<td>0.12</td>
<td>(0.10–0.15)</td>
<td>0.00</td>
<td>(0.00–0.00)</td>
</tr>
<tr>
<td>Cutler et al. (2001)</td>
<td>0.000</td>
<td>(0.000–0.000)</td>
<td>0.12</td>
<td>(0.11–0.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabriel et al. (2002)</td>
<td>0.001</td>
<td>(0.001–0.002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reich et al. (2003)</td>
<td>0.007</td>
<td>(0.006–0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positive rate ($\epsilon_1$) is 0.16 for TSC SNPs and 0.17 for BAC overlap SNPs ($\chi^2$ value for best-fitting model is 13.18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlson et al. (2003)</td>
<td>0.000</td>
<td>(0.000–0.000)</td>
<td>0.12</td>
<td>(0.11–0.13)</td>
<td>0.00</td>
<td>(0.00–0.00)</td>
</tr>
<tr>
<td>Cutler et al. (2001)</td>
<td>0.000</td>
<td>(0.000–0.000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabriel et al. (2002)</td>
<td>0.002</td>
<td>(0.002–0.002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reich et al. (2003)</td>
<td>0.008</td>
<td>(0.007–0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\epsilon_2$ is the probability that the major allele is miscalled as the minor allele, $\epsilon_3$ is the probability that the minor allele is miscalled as the major allele without regard to minor allele frequency and $\epsilon_4$ is the probability that the minor allele is miscalled as the major allele if the minor allele is rare and does not appear homozygous in the dataset.

Acknowledgements

This work was supported by a grant from the National Institutes of Health (HG2757-1).

References


Appendix

Following is the derivation of the probability of seeing the minor allele at an SNP in k out of n chromosomes examined, given that the SNP is found in dbSNP using four different error models.
Background

For a new mutation, subject only to the force of genetic drift, the probability that the frequency of the new allele (i.e. the ‘derived’ allele) is between \( x \) and \( x + dx \) (limit as \( dx \to 0 \)) is approximately \( C x^{-1} dx \), where \( C^{-1} \) is half the mean time to fixation or loss of the new allele (Ewens, 1974). This approximation is often called the ‘sojourn time’ density of a neutral allele. The probability that a SNP will be observed among a subsequent study is

\[
\frac{C x^{-1} dx}{\int_0^1 C x^{-1} dx} = 2(1-x) dx, \quad 0 \leq x \leq 1. \quad (1)
\]

Equation (1) is simply the Bayesian posterior estimate of the allele frequency, with prior \( C x^{-1} dx \) and likelihood of observation \( 2x(1-x) \) (Ewens, 1974).

It is generally more practical to discuss the major and minor alleles at a locus, rather than the ancestral and derived alleles. The probability that the minor allele frequency of an SNP observed in a sample of size two is between \( x \) and \( x + dx \) is

\[
\frac{C x^{-1} dx}{\int_0^1 C x^{-1} dx} = 2(1-x) dx, \quad 0 \leq x \leq 1. \quad (1)
\]

Thus, under model 2, the probability that an SNP will fail to confirm is \( \epsilon_1 + 2(1-\epsilon_1)/(n+1) \) and, of the SNPs that do confirm, the proportion with minor allele frequency greater than 0.1 is still \( \sim 80\% \).

Model 3: False positives within dbSNP, genotyping error in subsequent study is independent of the observed genotypes of the other samples in the study. As in model 2, let \( \epsilon_1 \) be the probability that the dbSNP is a false positive. In addition, let \( \epsilon_2 \) be the probability of miscalling the true major allele (or the only allele in the case of a false positive) as the minor allele in the confirmation study. The probability of seeing the minor allele in \( k \) out of \( n \) chromosomes if the SNP is a false positive is

\[
\epsilon_1 \left( \frac{n}{k} \right) \left( \frac{\epsilon_2^k (1-\epsilon_2)^{n-k} + \epsilon_2^{n-k} (1-\epsilon_2)^k}{\epsilon_2^k + \epsilon_2^{n-k}} \right) . \quad (2)
\]

Thus, under model 2, the probability that an SNP will fail to confirm is \( \epsilon_1 + 2(1-\epsilon_1)/(n+1) \) and, of the SNPs that do confirm, the proportion with minor allele frequency greater than 0.1 is still \( \sim 80\% \).

Model 3: False positives within dbSNP, genotyping error in subsequent study is independent of the observed genotypes of the other samples in the study. As in model 2, let \( \epsilon_1 \) be the probability that the dbSNP is a false positive. In addition, let \( \epsilon_2 \) be the probability of miscalling the true major allele (or the only allele in the case of a false positive) as the minor allele in the confirmation study. The probability of seeing the minor allele in \( k \) out of \( n \) chromosomes if the SNP is a false positive is

\[
\epsilon_1 \left( \frac{n}{k} \right) \left( \frac{\epsilon_2^k (1-\epsilon_2)^{n-k} + \epsilon_2^{n-k} (1-\epsilon_2)^k}{\epsilon_2^k + \epsilon_2^{n-k}} \right) . \quad (2)
\]

If no errors occur, the probability of observing \( k \) copies of the minor allele among \( n \) chromosomes examined, given that the SNP was discovered in a sample of size two is

\[
\left( \frac{n}{k} \right) \frac{1}{2} \left[ x^k (1-x)^{n-k} + x^{n-k} (1-x)^k \right] dx. \quad (3)
\]

If the major allele is miscalled as the minor allele with probability \( \epsilon_2 \), and the minor allele is miscalled as the major allele with probability \( \epsilon_3 \), the probability of observing \( k \) copies of the minor allele among \( n \) chromosomes examined, given that the SNP was discovered in a sample of size two is

\[
\left( \frac{n}{k} \right) \frac{1}{2} \left[ p^k (1-p)^{n-k} + p^{n-k} (1-p)^k \right] dx, \quad (4)
\]

where \( p = (1-\epsilon_3)x + \epsilon_2(1-x) \). Reparameterizing (4) in terms of \( p \) yields

\[
\frac{2}{1-\epsilon_2-\epsilon_3} \left( \frac{n}{k} \right) \frac{1}{2} \left[ p^k (1-p)^{n-k} \right] dp = \frac{2}{1-\epsilon_2-\epsilon_3} \left( \frac{n}{k} \right) \frac{k!(n-k)!(n+1)!}{(n-k+1)!} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{n-k+1} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{k+1} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right) \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{n-k+1} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{k+1} ,
\]

\[
\left( \frac{n}{k} \right) \frac{1}{2} \left[ p^k (1-p)^{n-k} + p^{n-k} (1-p)^k \right] dx, \quad (4)
\]

where \( p = (1-\epsilon_3)x + \epsilon_2(1-x) \). Reparameterizing (4) in terms of \( p \) yields

\[
\frac{2}{1-\epsilon_2-\epsilon_3} \left( \frac{n}{k} \right) \frac{1}{2} \left[ p^k (1-p)^{n-k} \right] dp = \frac{2}{1-\epsilon_2-\epsilon_3} \left( \frac{n}{k} \right) \frac{k!(n-k)!(n+1)!}{(n-k+1)!} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{n-k+1} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{k+1} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right) \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{n-k+1} \times \left( \frac{1}{1-\epsilon_2-\epsilon_3} \right)^{k+1} ,
\]
where \( I_s(a, b) \) is an incomplete beta function (Abramowitz and Stegun, 1972a) defined for integers \( a \) and \( b \) as

\[
I_s(a, b) = \frac{(a + b - 1)!}{(a - 1)! (b - 1)!} \int_0^x t^{a-1} (1 - t)^{b-1} dt.
\]

Simplifying and combining (2) with (5), the probability of observing \( k \) occurrences of the minor allele out of \( n \) chromosomes, given that the SNP was initially found in a sample of size two, allowing for errors in both studies is

\[
\begin{align*}
\varepsilon_1 \left( \begin{array}{c} n \\ k \end{array} \right) & \left[ \varepsilon_2^k (1 - \varepsilon_2)^{(n-k)} + \varepsilon_2^{n-k} (1 - \varepsilon_2)^k \right] \\
+ \frac{2(1 - \varepsilon_1)}{(1 - \varepsilon_2 - \varepsilon_3)(n+1)} \left[ I_{(1+\varepsilon_2-\varepsilon_3)/2}(k+1, n-k+1) \\
- I_{\varepsilon_2}(k+1, n-k+1) + I_{(1+\varepsilon_2-\varepsilon_3)/2}(n-k+1, k+1) \\
- I_{\varepsilon_2}(n-k+1, k+1) \right] \\
\end{align*}
\]

when \( 0 < k < n/2 \). If \( k = 0 \), the expression becomes

\[
\begin{align*}
\varepsilon_1 \left( \begin{array}{c} n \\ k \end{array} \right) & \left[ \varepsilon_2^k (1 - \varepsilon_2)^{(n-k)} + \varepsilon_2^{n-k} (1 - \varepsilon_2)^k \right] \\
+ \frac{2(1 - \varepsilon_1)}{(1 - \varepsilon_2 - \varepsilon_3)(n+1)} \left[ (1 + \varepsilon_2 - \varepsilon_3)(n+1) \varepsilon_2^{(n+1)} \\
- \varepsilon_2^{(n+1)} \right] \\
\end{align*}
\]

and if \( k = n/2 \), it becomes

\[
\begin{align*}
\varepsilon_1 \left( \begin{array}{c} n \\ k \end{array} \right) & \left[ \varepsilon_2^k (1 - \varepsilon_2)^{(n-k)} + \varepsilon_2^{n-k} (1 - \varepsilon_2)^k \right] \\
\times \left[ I_{(1+\varepsilon_2-\varepsilon_3)/2}(k+1, n-k+1) \\
- I_{\varepsilon_2}(k+1, n-k+1) \right] \\
\end{align*}
\]

Model 4: False positives within dbSNP, genotyping error in subsequent study is dependent on the observed genotypes of the other samples in the study.

If the SNP in dbSNP is real and the minor allele is common enough to occur in homozygous state in the confirmation study, we assume that genotyping error will not have a substantial effect on the ratio of \( k \) to \( n \), so this portion of the model includes no genotyping error terms for the confirmation study. The probability of seeing the minor allele in the sample is

\[
\begin{align*}
(1 - \varepsilon_1) \left( \begin{array}{c} n \\ k \end{array} \right) & \int_0^{1/2} \left[ 1 - (1 - x^2)^{n/2} \right] \left[ x^k (1 - x)^{(n-k)} \right] \\
+ x^{(n-k)}(1 - x)^k \right] \frac{dt}{2} \\
= 2(1 - \varepsilon_1) \left( \begin{array}{c} n \\ k \end{array} \right) \int_0^{1/2} \left[ (x^k (1 - x)^{(n-k)} \right. \\
+ x^{(n-k)}(1 - x)^k \right] \frac{dx}{2} \\
- \sum_{i=0}^{n/2} \left( \begin{array}{c} n/2 \\ i \end{array} \right) (-1)^i x^{2i} \left( x^k (1 - x)^{(n-k)} \right. \\
+ x^{(n-k)}(1 - x)^k \left. \right) dx \\
= 2(1 - \varepsilon_1) \left( \begin{array}{c} n \\ k \end{array} \right) \left[ \beta(k + n - k + 1, n - 1) \right. \\
- \sum_{i=0}^{n/2} \left( \begin{array}{c} n/2 \\ i \end{array} \right) (-1)^i \int_0^{1/2} \left( x^{2i+k} (1 - x)^{(n-k)} \right. \\
+ x^{2i+k}(1 - x)^k \left. \right) dx \\
= 2(1 - \varepsilon_1) \frac{n}{n+1} - 2(1 - \varepsilon_1) \left( \begin{array}{c} n \\ k \end{array} \right) \sum_{i=0}^{n/2} \left( \begin{array}{c} n/2 \\ i \end{array} \right) (-1)^i \\
\times \left[ (2i + k)(n-k)! \right. \\
\left. I_{1/2}(2i + k + 1, n - k + 1) \right. \\
+ (2i + n - k)!(k)! \right. \\
\left. I_{1/2}(2i + n - k + 1, k + 1) \right] \right]
\end{align*}
\]

\( \beta(a, b) \) is a beta function (Abramowitz and Stegun, 1972b), defined for integers \( a \) and \( b \) as

\[
\beta(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} dt = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}
\]

When \( k = 0 \), (6) reduces to

\[
2(1 - \varepsilon_1) \left[ \frac{1}{n+1} - \sum_{i=0}^{n/2} \left( \begin{array}{c} n/2 \\ i \end{array} \right) (-1)^i \left( \frac{(2i)! n!}{(2i+n+1)!} \right) \right] \]

\begin{align*}
&\times I_{1/2}(2i + 1, n + 1) + \left( \frac{1}{2} (\frac{(2i+n+1)!}{(2i+n+1)!}) \right] \right]
\end{align*}

(7)
and when \( k = n/2 \), (6) becomes

\[
2(1 - \varepsilon_1)n! \int_0^{1/2} (1 - x^2)^{n/2} \left[ p^k (1 - p)^{(a-k)} \right] dx 
= 2(1 - \varepsilon_1) \sum_{i=0}^{n/2} \binom{n/2}{i} (-1)^i \int_{I(1 - \varepsilon_1)k}^{I(1 - \varepsilon_1)k + 1/n} \frac{n!}{(1 + \varepsilon_4)^{2i+1}} (2i + k)!(2i + n + 1)! I_{I(1 - \varepsilon_1)k}^{I(1 - \varepsilon_1)k + 1/n} I_{I(1 - \varepsilon_1)k}^{I(1 - \varepsilon_1)k + 1/n} \left[ \frac{(2i + k)!}{(2i + n + 1)!} \right] \left[ \frac{I_{I(1 - \varepsilon_1)k}^{I(1 - \varepsilon_1)k + 1/n}}{2i + n + 1} \right].
\]

If \( k = 0 \), (9) becomes

\[
2(1 - \varepsilon_1) \sum_{i=0}^{n/2} \binom{n/2}{i} \left( \frac{(-1)^i}{I(1 - \varepsilon_1)^2i+1} \right) \left[ \frac{(2i)!}{(2i + n + 1)!} \right] + \left[ \frac{(1 - \varepsilon_1)^2i + 1}{2i + n + 1} \right].
\]

If \( k = n/2 \), (9) becomes

\[
2(1 - \varepsilon_1) \sum_{i=0}^{n/2} \binom{n/2}{i} \left( \frac{(-1)^i}{I(1 - \varepsilon_1)^2i+1} \right) \left[ \frac{(2i + k)!}{(2i + n + 1)!} \right] \left[ \frac{(2i + k)!}{(2i + n + 1)!} \right] \left[ \frac{I_{I(1 - \varepsilon_1)k}^{I(1 - \varepsilon_1)k + 1/n}}{2i + n + 1} \right].
\]

Allowing for errors in both the initial and confirmation studies, the probability of observing \( k \) out of \( n \) occurrences of the minor allele under model 4 is then:

\[
(2) + (7) + (10) \quad \text{if } k = 0,
\]

\[
(2) + (6) + (9) \quad \text{if } 0 < k < \frac{n}{2},
\]

\[
(2) + (8) + (11) \quad \text{if } k = \frac{n}{2}.
\]