Correcting Coalescent Analyses for Panel-Based SNP Ascertainment

James R. McGill, Elizabeth A. Walkup, and Mary K. Kuhner

University of Washington
Department of Genome Sciences
Box 355065
Seattle, WA 98195-5065
Running Title

Correcting Panel SNP Ascertainment

Key Words and Phrases

- panel SNPs
- SNP chips
- ascertainment
- statistical correction
- minor allele frequency

Corresponding Author

Mary K. Kuhner
University of Washington
Department of Genome Sciences
Box 355065
Seattle, WA 98195-5065
Phone: 206.543.8751
Fax: 206.685.7301
mkkuhner@u.washington.edu
ABSTRACT

Single nucleotide polymorphism (SNP) data are routinely obtained by sequencing a region of interest in a small panel, constructing a chip with probes specific to sites found to vary in the panel, and using the chip to assay subsequent samples. The size of the chip is often reduced by removing low-frequency alleles from the set of SNPs. Using coalescent estimation of the scaled population size parameter, $\Theta$, as a test case, we demonstrate the loss of information inherent in this procedure and develop corrections for coalescent analysis of SNPs obtained via a panel. We show that more accurate $\Theta$ estimates can be recovered if the panel size is known, but at considerable computational cost as the panel individuals must be explicitly modeled in the analysis. We extend this technique to apply in the case where rare alleles have been omitted from the SNP panel. We find that when appropriate corrections for panel ascertainment and rare-allele omission are used, the biases introduced by ascertainment are largely correctable, but recovered estimates are less accurate than would be obtained with fully sequenced data. This method is then applied to recombinant multiple population data to investigate the effects of recombination and migration on the estimate of $\Theta$. 
Introduction

Data for genetic studies are commonly obtained by typing only those Single Nucleotide Polymorphisms (SNPs) previously identified using a panel of fully sequenced individuals. Use of a panel to guide SNP genotyping clearly misses some variation in the samples thus reducing the power of the resulting data. Lost variation is not randomly distributed: recent mutations, which are generally of low frequency, drop out at a higher rate than older mutations (Nielsen 2000), and this may bias analysis of the resulting data (Nielsen 2004a). Previous papers (Kuhner 2000, Nielsen et al. 2004b) have addressed this potential bias and concluded that, if the composition of the panel is known, its effects can be accounted for. It may also be possible to reconstruct the relevant properties of the panel if one has access to SNP and sequence data from the same individuals (Albrechtsen et. al. 2010).

SNP collections are then often stripped of rare alleles using various allele-frequency cutoffs (HapMap 2003) in an attempt to reduce the impact of sequencing error (which introduces spurious, low-frequency alleles) as well as to reduce the size, and therefore cost, of the resulting SNP chip. This again biases the SNP collection towards higher frequency alleles. Panel use and allele frequency cutoffs both remove rare alleles, so the potential bias is substantial, and can easily mask the signature of population growth (Coventry et.al. 2010).

In this study, we demonstrate the bias produced by SNP panels, and provide an appropriate correction. We consider the case in which the number of panel sequences
is known but their genotypes are unknown, as well as exploring the consequences of missing or inaccurate information about panel size. We also consider the effects of sequencing error and of the deletion of rare alleles on inference from panel-based SNPs. Our correction is an implementation of the ‘reconstituted DNA’ method (Kuhner 2000), which requires knowing the count of invariant sites from the region of the genome from which SNPs were taken. To model recombination, the location of the SNPs is also required. SNP chips not providing this information cannot be used for recombinant coalescent analyses.

SNP data are used for diverse purposes including gene mapping, association studies, and inference of population parameters. As our test platform for the ability to correctly analyze panel-ascertained SNP data, we use inference of the scaled population size parameter $\theta = 4N_e \mu$, where $N_e$ is the effective population size and $\mu$ is the per-site mutation rate. Some of our experiments also infer recombination rate $r = C/\mu$ where $C$ is the recombination rate per inter-link site per generation, and immigration rates $M = m/\mu$ where $m$ in the immigration rate per generation. To infer these parameters, we use the coalescent genealogy-sampler strategy in the LAMARC program (Kuhner 2006). Inference of $\theta$ in coalescent genealogy samplers is based on the inferred distribution of branch lengths. It is therefore broadly sensitive to the distribution of mutations in the underlying sequences, and reacts strongly to the systematic loss of rare alleles. Inability to accurately estimate $\theta$ from panel-SNP data will skew the search of coalescent trees and therefore perturb recovery of other population parameters.
In this study we measure the biases of single population SNP panels, both with and without the removal of rare SNPs, and develop methods for correcting for these, yielding less biased \( \theta \) estimates within credibility intervals more often containing the truth. We also demonstrate that our method can be used in cases with population subdivision and recombination. Our corrections have been implemented in LAMARC but the principles involved are broadly applicable to coalescent analysis of SNP data gathered with a panel.

In this paper we will consider only the case in which panel members were surveyed from all populations in the analysis. We expect the bias inherent in use of a panel wholly drawn from a different population to be more severe, but we will defer consideration to a later paper.

**Materials and Methods**

We have added two new capabilities to LAMARC: panel corrections and error-aware likelihood analysis. This paper focuses primarily on panel corrections, though the error correction method is also discussed.

**Panel Correction**

Coalescent samplers make use of what has been called the "Felsenstein Equation" (Felsenstein 1988):

\[
P(D|\theta) = \sum_{G} P(G|\theta) \cdot P(D|G)
\]  
(1)
The probability of the observed data $D$ for a given value of $\theta$ is here expressed as a sum, over all possible genealogies $G$, of the combined probability of the genealogy with regard to $\theta$ and the data with regard to that genealogy. As the summation over genealogies is intractably large for realistic data sets, this equation is usually evaluated via Markov chain Monte Carlo methods (see Kuhner 2009 for a review).

For DNA data the term $P(D|G)$ can be computed under a variety of mutational models (Felsenstein 2004). Use of fully ascertained SNP data requires modification of this term to include the probability of omitted invariant sites. Kuhner et al (2000) gives a correction for SNP data, termed the “reconstituted DNA” method. This correction assumes that the panel captures all variation in the sequences (as would happen if the samples themselves were used as the panel) and that the number of invariant sites is known, but their sequence is not. Using this correction, $P(D|G)$ becomes as follows where $s$ ranges over the SNP (variant) sites and $u$ over the unobserved (invariant) sites.

\[
P(D|G) = \prod_s P(D_s|G) \cdot \prod_u P(D_u|G)
\]  

For non-recombinant genealogies, the rightmost product term in Equation (2) above simplifies as follows, where $I$ represents invariant data at a single site and $|u|$ is the number of invariant sites. The equation for recombinant genealogies would be the product of several such right-side terms, one for each non-recombinant marginal tree.

\[
\prod_u P(D_u|G) = [P(I|G)]^{|u|}
\]
Summing over all possible values of $I_x$ (the refinement of $I$ to a specific base, $x$) this becomes

$$\prod_u P(D_u|G) = \left[ \sum_{x \in \{a,c,g,t\}} P(I_{x}|G) \right]^{|u|} \quad (4)$$

This basic correction is not sufficient for panel-ascertained data, as it assumes that every unobserved site is known to be invariant. When data are ascertained based on a panel, some variable sites in the sample will be missed because they did not vary in the panel. Treating them as if they are invariant will bias estimates of $\theta$ downward.

Kuhner et al. (2000) proposed, but did not implement, a correction for panel-based SNPs that relies on explicitly representing panel sequences as proxy samples in each sampled coalescent genealogy. Thus if ten haplotype samples were collected using a SNP panel originally taken from six haplotypes, each genealogy would contain sixteen haplotypes: ten sample haplotypes and six panel-proxy haplotypes. We will refer to all of these haplotypes as “tips” as they are at the tips, or leaves, of the genealogy.

Calculating the likelihood of the data on such an expanded genealogy is a little more complicated than given above as we now have four different classes of site data based on the division of sites into assayed and un-assayed and of haplotypes into sample and panel. These four classes are as follows:
• \( D_s \): Data samples at assayed sites. Variation here is fully captured.

• \( D_u \): Data samples at un-assayed sites. Nothing is known about these sites’ data.

• \( D_s' \): Panel proxies at assayed sites. Bases are unspecified but known to vary within the set of proxies.

• \( D_u' \): Panel proxies at non-assayed sites. Bases are unspecified but known to be invariant across the set of proxies.

These data can be combined to calculate \( P(D|G) \) with the following decomposition.

\[
P(D|G) = \prod_{s \in S} P(D_s \cup D_s'|G) \cdot \prod_{u \cup u'} P(D_u \cup D_u'|G)
\]

(5)

Calculation of data likelihood for the assayed and un-assayed sites is similar, with Figure 1 giving an overview of the procedure. For assayed sites, one first calculates \( P(D|G) \) with sample tips as measured and panel tips unknown, and then subtracts each of four likelihoods with sample tips as measured and panel sites all invariant for one of the four possible bases. For un-assayed sites the procedure is to sum each of four possible likelihoods in which the panel sites are invariant for a single base, while assuming all sample data is unknown. Likelihood calculation is done with the tree-peeling algorithm of Felsenstein (1981), as previously implemented in LAMARC. The calculation must be done five times for assayed sites and four times for un-assayed sites.
Figure 1: Pseudocode for Calculating $P(D|G)$ with panel correction

Calculate $P(D_s \cup D_s'|G)$, data likelihood on tree $G$ for assayed sites:

- Calculate $L_N$, likelihood of assayed sites on $G$ with panel proxy values unknown
- Assign data tips $D_s$ likelihood consistent with assayed values:
  - An observation of a base $x$ at a given position has likelihood 1.0 for base $x$
  - All other bases $\neq x$ have likelihood 0.0

- Assign panel proxy tips $D_s'$ likelihoods consistent with unknown data
  - All bases are simultaneously given likelihood of 1.0
  - Propagate likelihood through $G$ to the root and save the value as $L_N$

For each $x \in \{a, c, g, t\}$, calculate $L_x$, likelihood for invariant panel with base $x$
- Assign data tips $D_s$ likelihood consistent with assayed values as above
- Panel proxy tips $D_s'$ are assigned likelihood of 1.0 for $x$
- Panel proxy tips $D_s'$ are assigned likelihood of 0.0 for all bases $\neq x$
- Propagate likelihood to the root of $G$ and save the value as $L_x$

\[
P(D_s \cup D_s'|G) = L_N - (L_a + L_c + L_g + L_t)
\]

Calculate $P(D_u \cup D_u'|G)$, data likelihood for unassayed sites

- For each $x \in \{a, c, g, t\}$, calculate $L_x'$, likelihood that panels all had value $x$
  - Sample tips $D_u$ are assigned likelihood 1.0 for all four bases (unknown data)
  - Panel proxy tips $D_u'$ are assigned likelihood of 1.0 for $x$
  - Panel proxy tips $D_u'$ are assigned likelihood of 0.0 for all bases $\neq x$

- Propagate likelihoods to the root of $G$, and retain likelihood $L_x'$

\[
P(D_u \cup D_u'|G) = L_a' + L_c' + L_g' + L_t'
\]

Multiply $P(D_s \cup D_s'|G)$ and $P(D_u \cup D_u'|G)$ to get complete data likelihood $P(D|G)$

Figure 1 gives the method for calculating $P(D|G)$ with panel correction. For simplicity’s sake, sequencing error is ignored here. Assayed and un-assayed sites are calculated in separate steps, each simultaneously handling data samples and panel proxy members.

Likelihood calculations are performed as in previous LAMARC versions.
The algorithm for searching among genealogies -- the $P(G|\theta)$ term in Equation (1) -- does not change to accommodate panels, other than inclusion of the panel tips. Instead, adding panel proxies to the genealogy allows us to approximately integrate over the unknown relationship between the panel sequences and the sample sequences. This in turn allows us to calculate the probability of the data given the genealogy while conditioning on panel ascertainment.

**Error-Aware Likelihood Analysis**

Neither the sequencing of the original panel, nor application of the resulting SNP chip to samples is an error-free process. Sequencing error typically manifests as observation of a novel allele in a single sample. Such errors can bias population parameter estimation (Clark and Whittam 1992, Johnson and Slatkin 2008). In this study we make use of a correction for sequencing error proposed by Felsenstein (2004), and briefly described below. This correction has been implemented and was released in LAMARC version 2.1.5.

In a naive (error-unaware) likelihood analysis, an observation of nucleotide $c$, for example, is represented at the tip with likelihood 1.0 under the hypothesis that the underlying nucleotide was indeed $c$ and likelihood 0.0 under the hypotheses that it was any of $\{a, g, t\}$. Under a uniform, randomly distributed sequencing error of rate $\varepsilon$, the likelihoods become $(1 - \varepsilon)$ for $c$ and $\varepsilon/3$ for each of $\{a, g, t\}$. That is, when the true base is $c$, there is a $(1 - \varepsilon)$ probability that $c$ will be observed and for each of $\{a, g, t\}$ as the true base there is a $\varepsilon/3$ probability that $c$ will be observed.
Our current implementation and experiments assume that error is randomly distributed and that sequencing error affects both the panel and the subsequent samples with the same error rate. The error rates of DNA sequencing and SNP detection will generally differ. Accommodating this would be a straightforward enhancement, as modeling error only requires one-time calculation of error-aware tip data likelihoods.

Throughout this paper, whenever the data were simulated with sequencing error, errors were introduced at a rate of 0.001 per observed nucleotide, and were random with respect to the base introduced. LAMARC’s error correction feature was used on all error-containing data, except in cases where MAFC (described below) was performed.

MAFC and Effective Panel Size

It has been common practice when constructing a SNP chip to discard minor alleles observed in less than 1% - 3% of the panel sequences (HapMap 2003). We know of no standard name for this practice. We will call it the Minor Allele Frequency Cutoff (MAFC) procedure. MAFC methods are used both to reduce the impact of sequencing error and to reduce the size, and thus the cost, of SNP chips. This approach necessarily discards some real information, particularly recent (and thus rare) mutations. Users expect that the chip’s power to detect variation will only be slightly reduced since the omitted SNPs are likely to be rare in the analyzed samples as well. We examine this assumption in Table 5 in the Results.
Removing rare alleles from a panel has an effect similar to using a smaller panel created earlier in the history of the population. A requirement that a site appears three times in the panel to be included in a SNP chip is similar to a requirement that a mutation happened long enough ago to appear in at least three contemporary panel members. Mutations arising recently are unlikely to have three descendants, and are lost. Thus the MAFC procedure makes it difficult to resolve and evaluate the most tip-ward portions of the genealogies.

We have chosen to model this effect by determining an “effective panel size” for the MAFC procedure. Our technique is very simple: we estimate the size of a non-MAFC panel that would be expected to yield the same number of retained SNPs as our MAFC panel. Two potential alternatives we did not explore are determining an effective panel size to match SNP frequency spectrums, as has been done to correct summary statistics (Nielsen et al. 2004b, Adams and Hudson 2004) and principal component analyses (Albrechtsen et. al. 2010), and using a coalescent genealogy sampler that allows for multiple time points (such as the BEAST sampler, Drummond and Rambaut 2007) to model the MAFC panel as a panel taken in the past.

To estimate the effective panel size, we convert Watterson’s estimator (Watterson 1975) from per-locus $\theta$ to per-site $\Theta$, and use it “backwards” to estimate count of segregating sites from $\Theta$, sequence length, and panel size. We then similarly adapt Ewens’ formula (Ewens 1972; Ewens 2000, Eq. 3.83, p114) for the distribution of $n$ samples with a per-locus $\theta$ under the infinite alleles model to the simpler case of bi-allelic loci, and switching to per-site $\Theta$. This provides the probability that a given site has
$j$ copies of the minor allele among $m$ samples. Combining the two formulae, we produce the equation below. Determining a suitable effective panel size corresponds to finding a value, $p$, which minimizes the expression below, where $m$ is the number of original MAFC panel members, and $c$ is the highest minor allele size discarded by the MAFC correction. Small $\Theta$ terms factor out and some of the summation terms cancel. The full derivation is given in File S1.

$$\left| \frac{1}{m-1} \sum_{i=p}^{m-1} \frac{1}{i} - \left( \sum_{j=1}^{c} \frac{m}{j(m-j)} \right) \right| \quad (6)$$

Note that while Watterson’s estimate relies on the infinite sites model and Ewens’ equation on the infinite alleles model, the two should be reasonably comparable for low values of $\Theta$. This is supported by the low counts of SNPs with three or more alleles found in our simulated data (see Table 1). Additional effort to model the process may be wasted as one is forced to choose an integer value for the effective panel size, $p$, the number of panel-proxy tips to include.

Panel correction adds tips to the genealogy corresponding to the panel sequences. When an effective panel size as given in Equation (6) is used, the tips added to the genealogy do not correspond to real sequences. Instead, Ewens’ formula and Watterson’s estimator combine to give an effective panel size that should discover a similar number of SNPs as the original panel size with a given MAFC. Since nothing is
known about the panel sequences as individuals, there is little information lost by this approximation.

Calculating effective panel sizes for multiple population cases is more difficult as each panel proxy tip must be assigned to a specific population while MAFC considers minor allele count across all populations. For our two-population case, we calculated the effective panel size as if all SNPs surviving the MAFC cutoff did so by meeting the MAFC minimum within their sub-population of origin. This will tend to result in too-large effective panel sizes, as SNPs occurring fewer times than the cutoff in a single populations might still meet the cutoff when all populations are considered. A more accurate correction would require modeling the migration rates and we cannot rely on that if we are attempting to recover those rates.

As the formulae we used involve constant population sizes, neither the single-population nor multi-population MAFC corrections will be appropriate for data sets including growth, shrinkage, or population divergence.

An alternative to the effective panel size approach would be to compute the probability of the observed data conditional on a position being variable under MAFC, as is done in the basic SNP correction (Kuhner et al. 2000). This probability is quite expensive to compute as it involves a sum over all possible data configurations containing fewer than the required number of minor allele occurrences. While this computation can be simplified for reduced mutational models such as Jukes-Cantor, it suffers from combinatorial explosion for more realistic mutational models. We therefore
use the effective panel size approach, which does not restrict our choice of mutational model.

**Generation of Simulated Data**

For each distinct set of population parameters under study, we began by simulating 100 random coalescent genealogies with enough tips to supply the maximum number of samples and panels examined across all studies on that parameter set. DNA sequences were simulated at the tips of each tree. This allowed us to compare the effect of different sequencing and panel schemes on the same “true” underlying data.

Genealogies were created using software based on Hudson’s simulator, *ms* (Hudson 1983; Hudson 2002). DNA sequences were simulated at the tips of each of the trees under the Kimura 2-parameter mutational model (Kimura 1980) with a transition/transversion ratio of 2.0, using the program treedna.c (Felsenstein unpublished). We do not expect our results to be sensitive to the mutational model used.

In some cases we added 0.1% simulated sequencing error. This was done before the SNPs were identified, so it had the effect of increasing the number of SNPs. This error value is somewhat optimistic as the HapMap estimate is 0.5% (Akey et.al.2002), though high quality sequencing can achieve lower error rates (Murray et. al. 2004).

Panel and sample members were randomly chosen from the tips present in each tree. A random ordering was chosen for the panel members and used across all simulations generated for the given tree. Thus, all SNPs included in the panel of size 8
for a given set of conditions are also included in the panel of size 16 over the same conditions. The unused tips were eliminated before the analysis was done. The same choice of panel members was used for each pair of simulations that had the same conditions save for the application of sequencing error.

We identified a site as a SNP if it varied in the panel, or for analyses with the MAFC procedure, if its minor allele frequency in the panel was above the cutoff. Once a site was identified as a SNP in the panel, we assumed it to be fully typed in the sample. This procedure captures slightly more information than the industry standard. In practice, if at a particular site the panel individuals possessed only nucleotides A and C, the SNP chip would contain probes only for A and C, and a T nucleotide in a sampled individual analyzed with the chip would not be identified. For our smallest θ value (0.001), the difference between these approaches is small as sites with more than two nucleotides present are rare. For our highest θ value (0.1), however, our approach would result in correct detection of more variants than the standard SNP-chip approach.

**LAMARC Analysis Conditions**

For the simple, single population experiments, we began with LAMARC (Kuhner 2006) version 2.1.5 and added the panel correction method. LAMARC 2.1.5 had a known bug (corrected in LAMARC 2.1.6) in analysis of recombination, but no recombination was present in these studies. The multi-population experiments were performed with LAMARC 2.1.8, which contains the panel correction method. All analyses used a Bayesian approach with logarithmic priors. Specific prior details, start values, and parameter-sampling strategies are available in Files S3 and S4.
Single-Population Experiments

Our single-population experiments began by generating sets of 100 trees, each containing 148 tips, for each of three different θ values:

- 0.1 to simulate fast mutating populations such as viruses
- 0.001 to approximate human mutation rates
- 0.01 to confirm that the trends seen at the extremes held at intermediate rates

In order to keep information content similar across these, we chose sequence lengths expected to produce the same number of SNPs. Table 1 shows the median and mean SNP counts for these three values as well as the percentage of SNPs with more than two alleles. Straightforward modification of Watterson’s estimate to predict segregating sites from sequence length, sample count, and per-site θ predicts 278.6 SNPs for each of these data. Our observed counts depart from this number. We used a finite-sites mutational model and at the higher values of θ multiple mutations at the same site occur more often, bringing the mean SNP count lower. Mean SNP count for the smallest θ value appears high, but is within 1.75 standard deviations of the expectation of the mean over 100 trials (Watterson 1975, Ewens 2000, p310).

This table also shows that for θ = 0.001, positions with more than two alleles were quite rare. This suggests that results for a conventional SNP chip capable of detecting only 2 alleles would be very similar to the results presented in this study.
All single-population experiments were performed on subsets or variants of these original 300 coalescent trees. The details of the naming of the subsequent data sets are found in File S2.
### Table 1: Comparison of SNP Counts in Data Sets

<table>
<thead>
<tr>
<th>θ</th>
<th>Length of DNA</th>
<th>Median # of SNPs in 100 generated trees</th>
<th>Mean # of SNPs in 100 generated trees</th>
<th>Mean % SNPs with more than 2 alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>500</td>
<td>215</td>
<td>212.8</td>
<td>12.34</td>
</tr>
<tr>
<td>0.01</td>
<td>5000</td>
<td>260</td>
<td>268.6</td>
<td>1.21</td>
</tr>
<tr>
<td>0.001</td>
<td>50000</td>
<td>268</td>
<td>290.1</td>
<td>0.13</td>
</tr>
</tbody>
</table>

For each θ value, 100 trees were generated, each with 148 tips. DNA was simulated for the sequence length given. Sequence lengths were chosen to recover a similar number of SNPs for each θ value. Watterson’s estimator predicts 278.6 SNPs for each of these data sets.
Panel correction and declaration of error: This key experiment measured the effects of the number of panel members, use or non-use of the panel correction, and sequencing error.

Sample size: The effect of using larger samples was investigated by comparing cases with 8 and 20 samples.

Incorrect panel size declaration: The effect of incorrectly declaring the panel size was studied by taking data sets generated with 8 panels and declaring them to have been created using 2, 4, 8, 12, or 16 panels.

MAFC and effective panel size correction: From data sets created with a panel size of we removed SNPS whose minor alleles appeared 2, 4, 8, and 16 or fewer times (frequencies 1.6%, 3.1%, 6.2% or 12.5%). For tri-allelic (or more) sites, the cutoff was applied whenever the sum of all minor allele counts failed to clear the cutoff. Thus when a major allele appeared 98% of the time with two minor alleles each appearing 1% of the time, the site would be included with the 1.6% cutoff, but not for 3.1% or above. These data sets were analyzed with effective panel sizes derived from Equation (6) -- 29, 16, 8, and 4 respectively.

Multiple Population Experiment

To explore the effect of panel and MAFC corrections on more sophisticated population models, we generated a single set of 100 independent genealogies with the following parameter values:
• $\theta$ of 0.002 for population “North”
• $\theta$ of 0.003 for population “South”
• No migration into population North from South
• Migration rate $M = m/\mu = 300$ into population South from North
• Recombination rate $r = C/\mu = 0.1$

Where error was modeled the rate was 0.1%. Each population had 64 original panels and MAFC corrections with cutoffs 3, 7 and 15 were applied. SNP data were taken for 8 samples in each population.

For these data we analyzed parameter recovery for the following conditions

• Fully sequenced data, with error, for analyses with and without error modeling
• Data as from a MAFC chip with cutoffs of 3, 7 and 15 (frequencies 2.3%, 5.5%, and 11.7%, respectively), analyzed without any correction for the MAFC procedure
• The same MAFC data, analyzed with an effective panel size correction (10, 5, and 2 proxy panels per population, respectively).

**Results**

**Recovery rate of SNPs using panels**

To explore the loss of SNPs due to the panel process, we compared the number of variable sites in our data sets with the number of SNPs detected using panels of
various sizes. Figure 2 presents the results. Focusing on the mean, one is tempted to conclude that panels of size 8-10 recover 50% of all SNPs and panels of size 28-30 recover 75%. While on average this is true, some individual data sets show severe loss of SNPs due to statistical fluctuation. Note, for example, the case where 80 panel members recovered less than 50% of the SNPs. This occurs when, due to chance, the sample contains one or more divergent lineages not encountered by the panel. The plots of $\theta = 0.1$ and $\theta = 0.01$ and a plot comparing all three $\theta$ values are found in the supporting information (Figures S1, S2, S3).
Figure 2: Proportion of SNPs recovered with increasing panel size

100 trees with 148 tips and $\theta = 0.001$ were generated and DNA simulated for 50,000 bases. For each data set a random ordering of all tips was chosen and the proportion of the total SNPs found as the ordered panel members were added is plotted here. Dotted lines show the edge of the results envelope and the solid line is the mean.
These sequences were simulated without recombination. Recombination would tend to move the number of SNPs detected towards the mean, because a lineage that was unusually divergent for one part of the sequence would not necessarily be divergent elsewhere.

**Panel Correction and Declaration of Error**

To gauge the effectiveness of panel correction, we performed LAMARC analyses on data sets produced with panels of various sizes, with and without simulated sequencing error. These results with sequencing error applied are summarized in Table 2 for all \( \Theta \) values and a graphical representation of the results for \( \Theta = 0.1 \) in the presence of sequencing error are given in Figure 3. (Figures and tables covering the case without sequencing error and the remaining \( \Theta \) values are given in Table S1 and Figures S4-S9 and include plots of Most Probable Estimates for all cases.)

Throughout our experiments, we calculated that for an experiment size of 100 instances, rejecting the truth between 0 and 10 times was consistent with a nominal rejection rate of 5%. Therefore we consider experiments in which LAMARC excluded the simulation value from its 95% credibility interval 10 times or less to be successful.

In the 2-panel data sets for the \( \Theta = 0.1 \) case, only 99 data sets were analyzed as in one case no SNPs were found in the panel. Results for this case are given as a percentage of the number of data sets actually run.
Table 2: Ability of LAMARC to recover Θ values from panel data with sequencing error

<table>
<thead>
<tr>
<th>Θ</th>
<th>Panel correction</th>
<th>Percent of LAMARC runs rejecting simulation Θ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2 panels*</td>
</tr>
<tr>
<td>0.100</td>
<td>No</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>7.1</td>
</tr>
<tr>
<td>0.010</td>
<td>No</td>
<td>74.0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>8.0</td>
</tr>
<tr>
<td>0.001</td>
<td>No</td>
<td>79.0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Percent of LAMARC runs rejecting their generating Θ value, given 0.1% sequencing error and panel correction as shown. All cases used the LAMARC error-aware correction. Except in the cases noted in the paper text, each rejection rate was calculated from 100 LAMARC runs on data obtained from the same set of 100 independently generated trees. Successive panel sizes include the panel members of the smaller sizes.
Figure 3: Effect of panel correction on mean recovered MPE and credibility intervals

The solid line and points are the average Most Probable Estimate (MPE) over 100 data sets and the average 95% credibility interval is enclosed by the dotted lines and outline points. The horizontal dashed line indicates the 0.1 Θ used in generating the data. Data were simulated with 0.1% sequencing error and the LAMARC correction was applied.
Panel and Sample Size

Felsenstein (2005) showed that a small number of samples is sufficient for accurate inference of $\theta$ for a single isolated population of fixed size. Table 3 shows that this is also true for panel-ascertained SNPs. We measured improvement in LAMARC’s ability to recover $\theta$ when switching from 8 to 20 samples for panels of various sizes. Both with and without the panel correction, larger panels captured the true value of $\theta$ more often than smaller ones. Without the correction, however, the truth was rejected at a high rate, especially with larger sample sizes. Scatter plots of MPEs and credibility interval widths (Figure S10) show tighter credibility intervals with increasing sample size both with and without the panel correction. This is expected, as additional information should allow tighter bounds. In uncorrected cases, there is a strong downward bias in the MPE of $\theta$, leading to increased rejection of the truth as the credibility intervals tighten. While Table 3 suggests a trend toward increased rejection of the truth with sample size even in corrected cases, the corrected formula captured the truth an acceptable proportion of the time for all panel/sample combinations tested. A table summarizing similar data without error is found in Table S2.
Table 3: Ability of LAMARC to recover $\Theta$ with different panel and sample sizes

<table>
<thead>
<tr>
<th>Panel Correction</th>
<th>Percent of LAMARC runs rejecting simulation $\Theta$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 panels 8 samples</td>
</tr>
<tr>
<td>No</td>
<td>47.0</td>
</tr>
<tr>
<td>Yes</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Percent of LAMARC runs from rejecting their generating $\Theta$ value, given 0.1% sequencing error and panel correction as shown. All cases used the LAMARC error-aware correction. Each rejection rate was calculated from runs on the same set of 100 independently generated trees. Successive panel sizes and sample sizes include the members of the smaller sizes. Plots of these data are found in Figure S11.
Incorrect Panel Size Declaration

Given that panel correction greatly improves our ability to recover $\theta$, a natural question is whether one must know the panel size to achieve this improvement. Figure 4 confirms one must know the panel size by comparing the results of SNPs generated with a panel size of 8 and declared panel sizes of 2, 4, 8, 12, or 16 for correction. The values used for this figure are found in Table S3.
100 data sets generated with $\theta = 0.1$ and 8 panels were analyzed using LAMARC with different declared panel sizes. The solid line and points are the average Most Probable Estimate (MPE); the dotted lines and outline points enclose the average 95% credibility interval. The horizontal dashed line indicates the 0.1 $\Theta$ used in generating the data.
MAFC Procedure and Effective Panel Size Correction

We began our exploration of the MAFC data by examining the number of SNPs found with and without sequencing error for several minor allele frequency cutoff levels applied to the 128-member panel. Table 4 shows the effective panel sizes calculated for each MAFC correction, along with the corresponding average SNP counts both with and without sequencing error. Note that at an error rate of 0.1%, the vast majority of SNPs occurring due to error were removed from the panel at the lowest MAFC level.

The trend of higher-than-expected observed mean SNP counts at each MAFC level was consistent with the higher than expected total SNP count seen in Table 1. The smaller observed SNP count for higher MAFC rejection levels on data with sequencing error was not intuitive. We therefore calculated an approximation to the expected distribution of observed minor allele counts under sequencing error. Drawing on earlier work modeling restriction site observations from underlying sequence data (Nei and Tajima, 1985 equation 4; Felsenstein 1992), we modeled the transformation from ‘true’, unobserved minor allele count to observed count with error. The observed decrease in MAFC SNP counts from error-free to error-containing data does not track exactly, but the same pattern is seen.
<table>
<thead>
<tr>
<th>MAFC Percent</th>
<th>Largest minor allele count removed by MAFC</th>
<th>Effective Panel Size</th>
<th>Expected SNP count No sequencing error</th>
<th>Approximate expected SNP count 0.1% sequencing error</th>
<th>Observed mean SNP count No sequencing error</th>
<th>Observed mean SNP count 0.1% sequencing error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0</td>
<td>128</td>
<td>271.3</td>
<td>6248.8</td>
<td>283.02</td>
<td>6259.6</td>
</tr>
<tr>
<td>1.6%</td>
<td>2/128</td>
<td>29</td>
<td>195.5</td>
<td>214.2</td>
<td>210.2</td>
<td>226.0</td>
</tr>
<tr>
<td>3.1%</td>
<td>4/128</td>
<td>16</td>
<td>165.5</td>
<td>166.8</td>
<td>183.0</td>
<td>180.5</td>
</tr>
<tr>
<td>6.2%</td>
<td>8/128</td>
<td>9</td>
<td>132.1</td>
<td>132.2</td>
<td>149.0</td>
<td>144.7</td>
</tr>
<tr>
<td>12.5%</td>
<td>16/128</td>
<td>4</td>
<td>95.5</td>
<td>94.2</td>
<td>102.8</td>
<td>100.7</td>
</tr>
</tbody>
</table>

Effective panel sizes, expected SNP counts without and with error, and mean observed SNP counts without and with error for 100 data sets with $\Theta = 0.001$. 
Table 5 summarizes LAMARC’s ability to recover $\theta$ from MAFC data with and without the effective panel size correction. It also includes the results for smaller panel runs and fully sequenced data for the same eight samples. Figure 5 gives a visual representation of the distribution of recovered $\theta$ values for each of the twelve cases listed in Table 5. The average MPEs and high and low credibility interval bounds for the MAFC data are plotted in the supplement (Figure S12).
Table 5: Performance comparison of different panel corrections

<table>
<thead>
<tr>
<th>MAFC level</th>
<th>Panels declared</th>
<th>Panels correction used</th>
<th>Error correction used in LAMARC analysis</th>
<th>Average run time (seconds)</th>
<th>Mean point estimate</th>
<th>Mean CI width</th>
<th>Percent rejection</th>
<th>Figure 5 letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>--</td>
<td>No</td>
<td>No</td>
<td>864</td>
<td>0.000461</td>
<td>0.000955</td>
<td>49.0</td>
<td>A</td>
</tr>
<tr>
<td>6.2</td>
<td>--</td>
<td>No</td>
<td>No</td>
<td>1,038</td>
<td>0.000672</td>
<td>0.001377</td>
<td>24.0</td>
<td>B</td>
</tr>
<tr>
<td>3.1</td>
<td>--</td>
<td>No</td>
<td>No</td>
<td>1,174</td>
<td>0.000806</td>
<td>0.001637</td>
<td>11.0</td>
<td>C</td>
</tr>
<tr>
<td>1.6</td>
<td>--</td>
<td>No</td>
<td>No</td>
<td>1,349</td>
<td>0.000888</td>
<td>0.001798</td>
<td>8.0</td>
<td>D</td>
</tr>
<tr>
<td>12.5</td>
<td>4</td>
<td>Yes</td>
<td>No</td>
<td>12,591</td>
<td>0.000950</td>
<td>0.002040</td>
<td>20.0</td>
<td>E</td>
</tr>
<tr>
<td>6.2</td>
<td>9</td>
<td>Yes</td>
<td>No</td>
<td>25,391</td>
<td>0.000877</td>
<td>0.001258</td>
<td>11.0</td>
<td>F</td>
</tr>
<tr>
<td>3.1</td>
<td>16</td>
<td>Yes</td>
<td>No</td>
<td>45,552</td>
<td>0.001023</td>
<td>0.001158</td>
<td>10.0</td>
<td>G</td>
</tr>
<tr>
<td>1.6</td>
<td>29</td>
<td>Yes</td>
<td>No</td>
<td>91,148</td>
<td>0.001182</td>
<td>0.001108</td>
<td>11.0</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>True panels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>4</td>
<td>Yes</td>
<td>33,373</td>
<td>0.001023</td>
<td>0.002541</td>
<td>8.0</td>
<td>J</td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>8</td>
<td>Yes</td>
<td>82,430</td>
<td>0.001070</td>
<td>0.001882</td>
<td>8.0</td>
<td>K</td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>16</td>
<td>Yes</td>
<td>244,245</td>
<td>0.001098</td>
<td>0.001550</td>
<td>8.0</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>Complete SNP data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>--</td>
<td>Yes</td>
<td>2,729</td>
<td>0.000962</td>
<td>0.002128</td>
<td>7.0</td>
<td>M</td>
</tr>
</tbody>
</table>

One hundred independent coalescent trees were generated, DNA data simulated for each, and 0.1% sequencing error applied. SNP data were then obtained using the panel sizes and MAFC procedures listed, with LAMARC runs completed using the corrections listed.
Figure 5: Distribution of MPE of $\Theta$ for different panel corrections

Violin plots showing the log base 10 of the distribution of recovered $\theta$ estimates for the runs in Table 5. Medians of analyses using the same corrections and data ascertainment schemes are linked for clarity.
Effect of MAFC Procedure and Effective Panel Size Correction on Multi-Population Data with Recombination

Table 6 and Table 7 summarize LAMARC’s ability to recover population parameters from the two-population case for MAFC data without and with correction as well as fully sequenced data. Figure 6 gives a visual representation of recovered parameter values for each of the seven cases listed in Table 6. No data is reported for the migration rate into population “North” as the simulated value, zero, is not recoverable by LAMARC.
Table 6: Rejection rates for MAF panel corrections on multi-population data

<table>
<thead>
<tr>
<th>MAFC level</th>
<th>Panels declared</th>
<th>Panels correction used</th>
<th>Error correction used in LAMARC analysis</th>
<th>Percent Rejection Theta North</th>
<th>Percent Rejection Theta South</th>
<th>Percent rejection Mig rate into South</th>
<th>Percent rejection Rec Rate</th>
<th>Figure 6 letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.7</td>
<td>No</td>
<td>No</td>
<td></td>
<td>28</td>
<td>30</td>
<td>10</td>
<td>33</td>
<td>N</td>
</tr>
<tr>
<td>5.5</td>
<td>--</td>
<td>No</td>
<td></td>
<td>9</td>
<td>11</td>
<td>7</td>
<td>14</td>
<td>P</td>
</tr>
<tr>
<td>2.3</td>
<td>--</td>
<td>No</td>
<td></td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>12</td>
<td>Q</td>
</tr>
<tr>
<td>11.7</td>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>22</td>
<td>20</td>
<td>18</td>
<td>23</td>
<td>R</td>
</tr>
<tr>
<td>5.5</td>
<td>5</td>
<td>Yes</td>
<td>No</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>14</td>
<td>S</td>
</tr>
<tr>
<td>2.3</td>
<td>10</td>
<td>Yes</td>
<td>No</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>14</td>
<td>T</td>
</tr>
<tr>
<td>Complete SNP data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--</td>
<td>--</td>
<td>Yes</td>
<td></td>
<td>7</td>
<td>1</td>
<td>9</td>
<td>13</td>
<td>U</td>
</tr>
</tbody>
</table>

One hundred independent coalescent trees were generated, DNA data simulated for each, and 0.1% sequencing error applied. SNP data were then obtained using the panel sizes and MAFC corrections listed, with LAMARC runs completed using the corrections as listed.
Table 7: Recovered MPEs for MAF panel corrections on multi-population data

<table>
<thead>
<tr>
<th>MAFC level</th>
<th>Panels declared</th>
<th>Panels correction used</th>
<th>Error correction used in LAMARC analysis</th>
<th>Mean recovered MPE Theta North</th>
<th>Mean recovered MPE Theta South</th>
<th>Mean recovered MPE Mig rate into South</th>
<th>Mean Recovered MPE Rec Rate</th>
<th>Figure 6 letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.7</td>
<td>No</td>
<td>No</td>
<td></td>
<td>0.001243</td>
<td>0.001483</td>
<td>883.8</td>
<td>0.2563</td>
<td>N</td>
</tr>
<tr>
<td>5.5</td>
<td>--</td>
<td>No</td>
<td></td>
<td>0.001736</td>
<td>0.002227</td>
<td>763.4</td>
<td>0.1407</td>
<td>P</td>
</tr>
<tr>
<td>2.3</td>
<td>--</td>
<td>No</td>
<td></td>
<td>0.002049</td>
<td>0.002817</td>
<td>606.0</td>
<td>0.1018</td>
<td>Q</td>
</tr>
<tr>
<td>11.7</td>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>0.004600</td>
<td>0.004576</td>
<td>660.1</td>
<td>0.1034</td>
<td>R</td>
</tr>
<tr>
<td>5.5</td>
<td>5</td>
<td>Yes</td>
<td>No</td>
<td>0.002255</td>
<td>0.002717</td>
<td>542.6</td>
<td>0.0929</td>
<td>S</td>
</tr>
<tr>
<td>2.3</td>
<td>10</td>
<td>Yes</td>
<td>No</td>
<td>0.002487</td>
<td>0.003337</td>
<td>404.3</td>
<td>0.0768</td>
<td>T</td>
</tr>
</tbody>
</table>

Complete SNP data

| --         | --              | No                      | Yes                                      | 0.002517                     | 0.0036601                     | 466.0                                  | 0.0750                        | U              |

One hundred independent coalescent trees were generated, DNA data simulated for each, and 0.1% sequencing error applied. SNP data were then obtained using the panel sizes and MAFC corrections listed, with LAMARC runs completed using the corrections as listed.
Figure 6: Distribution of MPE for two population parameter cases.

A  recovered $\theta$ North values

B  recovered $\theta$ South values

C  recovered migration from North to South

D  recovered recombination rates

Violin plots showing the log base 10 of the distribution of recovered parameter estimates for the runs in Table 6. Medians of analyses using the same corrections and data ascertainment schemes are linked for clarity.
Conclusions

Information is lost whenever a panel is used to guide SNP typing, and additional information is lost when a minor allele frequency cutoff (MAFC) procedure is applied. Chip-based SNP typing is attractive because it is less expensive than full re-sequencing and provides the possibility of performing additional sequencing quickly. MAFC procedures appeal because a chip made with MAFC will see more variation per SNP included than one without, and therefore appears more cost-efficient. When these methods are used without correction, they bias coalescent analysis of the collected data to underestimate Θ. It is not possible to remove error without also removing low-frequency signal. Therefore, knowing which low-frequency data are missing is essential.

We have demonstrated that for coalescent-based inference of population parameters, data ascertained via a panel can be analyzed with reduced bias through use of the panel correction suggested in Kuhner et al. (2000). Information loss is severe with the smallest panel sizes, as seen in the widened credibility intervals for panel sizes 2 and 4 (Figure 3), whereas increasing from 16 panel members to 32 does not have much effect on the credibility intervals.

As previously shown by Felsenstein (2005), the sample size needed for accurate inference of Θ is quite small. This is confirmed in the presence of panel correction by our results showing only modest improvement in credibility interval tightness when
going from 8 to 20 samples. Consequently, researchers undertaking coalescent analysis should consider full re-sequencing of a smaller number of individuals as an alternative to panel-based ascertainment of a larger number, especially in the case where the panel is to be created as part of the study. Panel-based ascertainment will sometimes lose unexpectedly large amounts of information. This is seen in Figure 2, in which a panel of size 80 missed over half the SNPs in a sample. Similar ill luck can befall the selection of samples for a re-sequencing study, but when it happens to a panel, all studies using that panel will be affected.

Use of MAFC is likely a false economy for coalescent analyses. Considerable work goes into sequencing of panel individuals to make the SNP chip and, as we have shown, even a modest MAFC procedure reduces the effective size of the panel drastically. An allele frequency cutoff of 1.6% on a 128-member panel is comparable to a 29-member panel without MAFC. In data with 0.1% sequencing error and a human-like $\Theta$, a 1.6% MAFC removes the vast majority of error SNPs. More aggressive MAFCs likely remove more signal than error. Appropriate statistical procedures can compensate for but not fully correct this loss of information.

Our two-population experiments suggest that $\Theta$ population parameters are more susceptible to panel-introduced bias than other parameters. In our experiments, rejection of the simulated migration and recombination values was comparable among fully sequenced, corrected, and uncorrected MAFC data for all but the most aggressive (11.7%) MAFC procedures.
Population growth raises additional concerns for panel-based methods. In a growing population, branches of the coalescent tree at the tips are elongated compared to the more ancient branches. This increases the proportion of private polymorphisms in the data (Coventry et al. 2010) and it is exactly these polymorphisms that are lost through the use of panel and MAFC approaches. While the proxy-based true panel correction we present remains applicable in the presence of growth, we expect the loss of information and therefore corresponding loss of accuracy will be greater. Our effective panel size correction is not expected to apply in the presence of growth, as the formulae we used assumed a stable population.

We do not expect panel-ascertained and MAFC data to vanish in the near future. If data must be collected in this way, several steps can be taken to retain as much information as possible:

- Panel size must be documented. We have shown that there is no conservative solution when panel size is not known: both guessing low and guessing high lead to bias (Figure 4). While Albrechtsen et al (2010) suggest methods for inferring panel size, this does not justify omitting such information from the documentation of future SNP chips, as exact knowledge must be better than inference. Even if a SNP chip represents a heterogeneous collection of SNPs ascertained from different panel sizes, this information should be made available, as future analyses may be able to take it into account. Without panel size information the correction given here is impossible.
• Information about any MAFC in effect is also essential, for similar reasons. We have shown that misstating panel size leads to bias. Though we did not measure this explicitly, misstating MAFC level is essentially the same as misstating panel size and should therefore also lead to bias.

• Aggressive MAFC should be avoided. Sequencing error can be better handled by a statistical correction on the raw data. If MAFC must be used, the cutoff should be set as low as possible. While the “rare SNPs” are individually rare, they are collectively numerous and contain a substantial part of the information in a population survey. They should not be considered irrelevant or uninteresting. Table 5 shows that aggressive MAFC leads to very poor estimates even with correction.

• Appropriate statistical corrections should be used for panel ascertainment, MAFC, and sequencing error. The panel corrections shown in this paper will be released in LAMARC version 2.1.8 and can be implemented in other coalescent-based analytic tools. The sequencing error correction was introduced in LAMARC version 2.1.5 and can be used in any coalescent or phylogenetic analysis of DNA data, with or without reduction to SNPs, so long as the error rate is known.

• Re-sequencing should be considered as an alternative to use of a SNP chip, particularly in cases with population growth or strong positive selection near the sampled genomic regions. We expect the loss of information to be more severe in such cases.
Avoiding error in one’s data sets is laudable, but insisting that all error be removed will lead to removing valuable information. When possible, error should be dealt with via statistical correction rather than mechanical removal. For coalescent analyses, this implies that collecting a small, high-quality set of data samples including both common and rare variants is superior to a larger sample of common variants only.

Acknowledgments: The authors would like to thank Joe Felsenstein, Josh Akey, and Bob Giansiracusa for their help. This work was supported NIH grant R01 HG004839 (to Mary Kuhner); many of the two-population simulations were completed with support from an AWS in Education Grants Award from Amazon.com, Inc.

Literature Cited


Coventry, A., et. al., 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature Communications, 1: 131


Kuhner, M. K., 2009 Coalescent genealogy samplers: windows into population history. TREE 24: 86-93


Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7: 256-276