File S1 for *Predicting variant discovery rates in sequencing studies*

Simon Gravel,
on behalf of the NHLBI GO Exome Sequencing Project
Department of Human Genetics and Genome Quebec innovation Centre,
McGill University, Montreal, Quebec, Canada

March 13, 2014

Jackknifes and naive linear bounds

We can obtain both upper and lower bounds for the number of undiscovered variants by linear combinations of the $\phi(d)$. To do this, we note that the equations for the number of missed variants

$$V(N) - V(n) = \int_0^1 \left((1-f)^n - (1-f)^N\right) \Phi(f) df$$

and for the number of variants at a given allele frequency

$$\phi_n(j) = \int_0^1 \binom{n}{j} f^j (1-f)^{n-j} \Phi(f) df$$

have a very similar form. The only difference is a 'weight factor' before Φ. If the weight function $w_{n,N}(f) = (1-f)^n - (1-f)^N$ can be approximated by functions of the form $b(f, \vec{\alpha}) = \sum_{i=1}^d \alpha_i f^i (1-f)^{d-i}$ then we can approximate $V(N) - V(n)$ in terms of the observed $\phi_n(i)$. In fact, this is exactly what the jackknife estimates do–A jackknife estimator would correspond to a function

$$J(f) = \sum_{i=1}^d \beta_i \phi(i),$$

with the $\vec{\beta}$ chosen such that $V(N)-V(n) = \int_{0^+} w_{n,N}(f)\tilde{\Phi}(f)$, for a particular d-parameter family of models $\tilde{\Phi}(f)$, thought *a priori* to be a reasonable proxy for $\Phi(f)$. This interpretation of the jackknife provides intuition about the behavior of jackknife estimators when the underlying model is not within $\tilde{\Phi}(f)$; comparison of the jackknife weight $J(f)$ and the correct weight $w(f)$ (Figure S4) provides an idea of the general accuracy of the jackknife estimate, and an idea of the frequencies that are more (or less) sensitive to errors.

However, we can also use the similarity between the expressions to obtain strict bounds on $V(N) - V(n)$, by choosing functions $b(f, \vec{\alpha}) = \sum_{i=1}^{d} \alpha_i f^i (1-f)^{n-i}$ that are strict bounds to $w_{n,N}(f)$. The best such bounds will be attained when the approximating function $b(f, \vec{\alpha})$ touches but does not cross $w_{n,N}(f)$

We can show that the best upper bound with $d = 2$ is $V(N) - V(n) < (N/n - 1)\phi(1)$. There is a one-dimensional family of lower bounds which are optimal for at least one function $\Phi(f)$, parameterized by the contact point $0 \leq f_0 \leq 1$ where

$$\begin{aligned} b_2(f_0, \vec{\alpha}_{f_0}) &= w_{n,N}(f_0) \\ b_2'(f_0, \vec{\alpha}_{f_0}) &= w_{n,N}'(f_0). \end{aligned} \tag{1}$$

To see that these $\vec{\alpha}_{f_0}$ exist and define lower bounds, consider the first, second, and third derivatives of the function $\frac{w_{n,N}(f) - b(f,\vec{\alpha})}{(1-f)^{n-2}}$.

For each f_0, we can solve for $\vec{\alpha}_{f_0}$, and thus obtain a lower bound to $V(N) - V(n)$. Given a sample, one can calculate all bounds and use the tightest. Figure 1 and Table 3 show results using this approach with simulated data. It is easy to derive bounds with higher d, but the process of establishing the optimal bound is more challenging. Extrapolations based on upper bounds with $d = 3$ are shown on Table 3.

As in the case of jackknife estimates, higher order for the bounds means reduced bias, but also reduced stability in the presence of errors.

Known proportion of invariant sites

In the ecology problem, the proportion of individuals or species that have not been observed is unknown; it is the object of the inference. In the genetic context, the total number of sequenced sites L may be known; the object of the inference is to determine the proportion of these sites that would be variable in a larger sample. This does not fundamentally change the inference process:

Jackknife bounds

In the jackknife case, we are provided with one additional function $(1-f)^N$ to try to obtain a linear bound to the weight functions $w_{n,N}(f)$. In the infinite-extrapolation case

($N = \infty$), we now have an upper bound to the number U of undiscovered variants: $U \leq \phi(0)$. This is an inequality because variants with frequency 0 are counted in $\phi(0)$ but not in $U = \int_{0^+}^{1} (1-f)^n \Phi(f)$.

Finite extrapolation bounds can be improved using the knowledge of $\phi(0)$, by following the procedure described in the 'Naive linear bound' section for the optimization of the $\vec{\alpha}_i$. However, we do not study these in detail here.

Linear programming bounds

In the linear programming framework, the observed $\phi(0)$ is easily incorporated as an additional equality constraint stipulating that $\sum_i \Phi(i) = \sum_j \phi(j)$. Intuitively, we expect that the additional constraint will help narrow the confidence interval.

However, when the total sample size is equal to the extrapolation size (i.e., $M = N$), this provides limited information because the additional constraint involves a new variable, $\Phi(0)$, that is not involved in the objective function $V(N)$. Thus, $\Phi(0)$ can be adjusted to satisfy the constraint without affecting $V(N)$. Starting from a vector $\Phi^*(i)$ realizing the upper bound $V_\uparrow^*(N)$ for the problem with $\phi(0)$ unknown, such an adjustment is possible unless $\sum_{i=1}^{N} \Phi^*(i) > \sum_{d=0}^{n} \phi(d)$, in which case $\Phi(0)$ would be negative, violating the constraint $\Phi(0) \geq 0$. In such a case, convexity ensures that the optimal solution must satisfy $\Phi(0) = 0$, and $V_\uparrow(N) = \sum_{d=0}^{n} \phi(d)$. Thus, in general, we simply have the somewhat disappointing result $V_\uparrow(N) = \min\left(V_\uparrow^*(N), \sum_{d=0}^{n} \phi(d)\right)$. The same argument holds for the lower bound, but since $V_\downarrow^*(N) \leq \sum_{d=0}^{n} \phi(d)$, the lower bound is unchanged by the additional information.

This argument does not hold if the population size M is larger than the extrapolation size N because, in that case, $\Phi_M(0) = 0$ does not imply $V(N) = \sum_{d=0}^{n} \phi(d)$. Indeed, we find an improvement of the upper bound that becomes more pronounced as the number of invariant site in the sample of size M is decreased.

Jackknife equivalence

We wish to show that the jackknife expansions A:

$$V(N) - V(n) = \sum_{i=1}^{p} a_i \left(H(N) - H(n)\right)^i$$

, and B:

$$V(N) - V(n) = \sum_{i=1}^{p} b_i H^i(N) - H^i(n)$$

lead to the same predictions. Both expansions can be written in the third expansion form C: $V(N)-V(n) = \sum_{i=0}^{p} c_i(N)H(n)^i$, for different parameterizations of $c_i(N)$. Importantly, these parameterizations do not involve n. In the parameter estimation, we use in the three cases the constraints $V(n) - V(n-j) = \sum_{i=0}^{p} c_i \left(H^i(n-1) - H^i(n)\right)$, for $j = \{1...p\}$. These provide p equations for p unknowns $\{c_i\}_{i\geq 1}$. We can solve for these independently of N. We could equally well expand the c_i in terms of, say, the a_i, solve a linear equation for the a_i, and substitute these back to produce exactly the same expansion. Thus, the expansions A, B, and C are equivalent for $i > 0$. In expansion C, the dependence on N enters only after we impose that $V(N) - V(n)$ must be zero when $N = n$. This imposes $c_0 = -\sum_{i=1}^{p} i c_i H(N)^i$. This simple form of the estimator, made explicit in expansion B, was obscured by the poor parameterization choice of expansion A: whereas the $\{b_i\}_{i\geq 1}$ depend only on n, the $\{a_i\}_{i\geq 1}$ are messy functions of N and n.