Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 33

On the Optimal Design of Genetic Variant Discovery Studies

Iuliana Ionita-Laza* Nan M. Laird†

*Columbia University, ii2135@columbia.edu

†Harvard School of Public Health, laird@hsph.harvard.edu

On the Optimal Design of Genetic Variant Discovery Studies*

Iuliana Ionita-Laza and Nan M. Laird

Abstract

The recent emergence of massively parallel sequencing technologies has enabled an increasing number of human genome re-sequencing studies, notable among them being the 1000 Genomes Project. The main aim of these studies is to identify the yet unknown genetic variants in a genomic region, mostly low frequency variants (frequency less than 5%). We propose here a set of statistical tools that address how to optimally design such studies in order to increase the number of genetic variants we expect to discover. Within this framework, the tradeoff between lower coverage for more individuals and higher coverage for fewer individuals can be naturally solved.

The methods here are also useful for estimating the number of genetic variants missed in a discovery study performed at low coverage.

We show applications to simulated data based on coalescent models and to sequence data from the ENCODE project. In particular, we show the extent to which combining data from multiple populations in a discovery study may increase the number of genetic variants identified relative to studies on single populations.

KEYWORDS: species problem, variant discovery studies, sequencing technologies

*We are grateful for the comments provided by two reviewers that helped improve the paper.

Introduction

New developments in sequencing technologies have brought along substantial reductions in cost and increases in genomic throughput by more than three orders of magnitude (Shendure and Ji, 2008; Tucker et al., 2009; Metzker, 2010). These improvements have in turn contributed to the launch of the 1000 Genomes Project (http://www.1000genomes.org/page.php), and to an increasing number of smaller scale sequencing studies. These studies have as one of their goals the identification of the genetic variation in the region under study, mostly low frequency variants, with the ultimate goal of performing association testing between rare variants and complex diseases (Li and Leal 2008, Madsen and Browning 2009).

The number of genetic variants identified in these studies depends on several factors. Among these the most important are the number of individuals sequenced, and the sensitivity to call variants present in a sequenced individual. With the new sequencing technologies the sensitivity to detect a variant is influenced directly by the depth of coverage at the variant position. Therefore when designing a variant discovery study researchers need to decide the number of individuals that can be sequenced, and the average depth of coverage to be used for each individual given a fixed study cost in order to maximize the expected number of variants to be discovered. The natural tradeoff is between lower depth of coverage for more individuals, and higher coverage depth for fewer individuals. In this paper, we propose analytical methods that allow estimation of the number of genetic variants we expect to identify in sequencing studies performed at reduced coverage, and that provide a natural solution to this tradeoff.

We frame the problem of estimating the genetic variation in a population as a species problem in ecology, where one is interested in estimating the number of species in a closed population (Sekar and Deming, 1949; Efron and Thisted, 1976; Bunge and Fitzpatrick, 1993). Here we extend the approach in Ionita-Laza et al. (2009) to account for imperfect probability of variant detection in an individual, and this extension allows us to address the optimality issue of the current sequencing studies. The approach is appealing in that it makes few and reasonable assumptions, and it makes use of already generated sequence data in the region of interest to empirically estimate the underlying distribution of the variant frequency. In the remaining of the paper we present details of the approach and show applications to simulated data based on a coalescent model and to the ENCODE data.

Methods

We present here a statistical framework for estimating the number of genetic variants we expect to identify in a fixed number of individuals. We start by introducing several assumptions on the underlying frequency distribution of variants in the population, and on the process of calling variants in a set of sequenced individuals. We assume here that variants are in linkage equilibrium, and that their population frequencies follow a Beta distribution (Wright, 1959), and hence the observed frequency counts follow a Beta-Binomial distribution. This assumption in addition to theoretical justification has also proved to work well in practice.

In the context of current sequencing technologies, the probability of calling a truly present variant is a function that depends on several factors, including the number of reads observed at the variant location (i.e., the coverage depth) and the number of reads consistent with the non-reference allele at the location (we assume that the reference allele is known). We focus here on the discovery of true variants, and assume that the probability of calling a non-existent variant is negligible. Note that this latter probability can be made as small as desired by imposing a more stringent variant calling rule.

Expected number of variants to be identified

We assume that the average depth of coverage across the genome is denoted by λ (per base), which means that each variant position is sequenced λ times on average; the value of λ is known. Accordingly, the number of reads, say m, at a given location (i.e., the coverage depth) is approximately Poisson(λ). The number of reads at a location is an important covariate as it affects the probability that a truly present variant in an individual is detected. In general, the higher the number of reads at a location, the greater the probability of making a call at a variant position. The specific algorithm for calling variants can vary from study to study. For our purposes we assume that a variant is called whenever at least k copies of the same (non-reference) allele are detected among all the reads at a position; however other calling procedures can be accommodated as well, but require more specific information about the sequencing technology used. If we let p_{err} be the error probability for a single read per base (assume the errors are independent), m be the total number of reads at a location, and m_{noerr} be the number of reads without error, then the

DOI: 10.2202/1544-6115.1581

calling probability can be modeled using a binomial distribution, as follows:

$$\begin{aligned} C_{m,k} &= P(\text{call a variant in an individual} \mid \#\text{reads at a location is } m) \\ &= P(m_{\text{noerr}} \geq k), \end{aligned} \tag{1}$$

where m_{noerr} is approximately Binomial$(m, 1-p_{\text{err}})$. A plot of $C_{m,k}$ as a function of m when $k = 3$ is shown in Figure 1 for two different values of the error probability p_{err}, 1% and 10%.

The expected number of variants with frequencies between f_1 and f_2 to be discovered in N_{ind} individuals is given by the following formula:

$$\begin{aligned} E_{N_{\text{ind}},f_1,f_2,\lambda} &= M \int_{f_1}^{f_2} P(\text{detect a variant with frequency } \theta) f(\theta) d\theta \\ &= M \int_{f_1}^{f_2} \left[1 - P(\text{do not detect a variant with frequency } \theta)\right] f(\theta) d\theta \\ &= M \int_{f_1}^{f_2} \left[1 - \left(1 - \sum_m C_{m,k} \theta f_\lambda(m)\right)^{2N_{\text{ind}}}\right] f(\theta) d\theta, \end{aligned} \tag{2}$$

where M is the total number of variants (unknown), f_1 and f_2 are bounds on the frequencies of variants, λ is the mean depth of coverage across the genome,

$$f_\lambda(m) = \frac{e^{-\lambda}\lambda^m}{m!} \text{ (Poisson distribution function)},$$

and

$$f(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \text{ (Beta distribution function)}.$$

Parameters a and b of the Beta distribution can be estimated from available data using maximum likelihood estimation (see Appendix).

Optimal coverage depth in order to maximize the expected number of discoveries

For a fixed study cost we are interested in maximizing the number of variants we expect to discover. Therefore an important aspect in the design of new sequencing studies is choosing the average depth of coverage that one should use per individual, and the number of individuals to sequence. In our framework, this can be rephrased as maximizing the expected number of variants we

expect to discover at a given coverage λ and for a given number of individuals N_{ind}, i.e.:

$$\max_{\lambda, N_{\text{ind}}} \quad E_{N_{\text{ind}}, f_1, f_2, \lambda} \tag{3}$$

with the constraint that the cost of the study is fixed. With the current sequencing technologies, it is reasonable to assume that the cost of the study is linear in the number of individuals, i.e. $\text{Cost} = cN_{\text{ind}} \cdot \text{Cost}(\lambda)$, where c is a constant, and $\text{Cost}(\lambda)$ is the cost of sequencing an individual at an average coverage depth of λ. We assume here that $\text{Cost}(\lambda)$ is linear in λ, although other cost functions can be accommodated as well.

Number of variants missed in a discovery study

We assume now that we have already performed a variant discovery study at average depth λ in N_{ind} individuals, and we are interested in estimating the number of genetic variants that we missed in the study. The number of unseen genetic variants can then be decomposed into two separate components: the number of variants missed in the N_{ind} individuals due to the depth of coverage used (denoted here by M_1), and the number of variants missed due to our sampling only N_{ind} individuals from the entire population (M_2). The first component (M_1) can be estimated as follows:

$$M_1 = E_{N_{\text{ind}}, f_1, f_2, \infty} - E_{N_{\text{ind}}, f_1, f_2, \lambda},$$

where $E_{N_{\text{ind}}, f_1, f_2, \infty}$ is the number of variants that we expect to discover in N_{ind} individuals with perfect coverage.

The second component (M_2), i.e. missing variants due to our sampling of only N_{ind} individuals, can be estimated as follows. Let $\Delta_{N_{\text{ind}}, f_1, f_2, \infty}(t)$ be the expected number of new variants to be discovered if we were to do a new discovery study in an additional $t \cdot N_{\text{ind}}$ individuals with perfect coverage. Then

$$\Delta_{N_{\text{ind}}, f_1, f_2, \infty}(t) \quad = \quad E_{(t+1) \cdot N_{\text{ind}}, f_1, f_2, \infty} - E_{N_{\text{ind}}, f_1, f_2, \infty}. \tag{4}$$

Note: Estimating the number of variants to be identified in future sequencing studies at the same depth of coverage λ as used in the initial discovery study follows easily using the above notations, i.e.

$$\Delta_{N_{\text{ind}}, f_1, f_2, \lambda}(t) = E_{(t+1) \cdot N_{\text{ind}}, f_1, f_2, \lambda} - E_{N_{\text{ind}}, f_1, f_2, \lambda}.$$

DOI: 10.2202/1544-6115.1581

Replication probability for variants in the discovery study

Once a discovery study has been performed it is often of interest to assess the chance that we will see a variant detected in the discovery study in other independent datasets. This replication probability depends on estimating the posterior frequency distribution for such a variant. The information about the frequency of each variant detected in the discovery study is given by x, the number of times it appears in the study. Given x, we can estimate the posterior frequency distribution for a variant, i.e.

$$P_x(f \leq p) = P(f \leq p \mid \text{variant is observed } x \text{ times in our study}),$$

as follows:

$$P_x(f \leq p) = \frac{\int_0^p (\sum_m \theta \cdot C_{m,k} \cdot f_\lambda(m))^x (1 - \sum_m \theta \cdot C_{m,k} \cdot f_\lambda(m))^{2N_{\text{ind}}-x} f(\theta) d\theta}{\int_0^1 (\sum_m \theta \cdot C_{m,k} \cdot f_\lambda(m))^x (1 - \sum_m \theta \cdot C_{m,k} \cdot f_\lambda(m))^{2N_{\text{ind}}-x} f(\theta) d\theta} \quad (5)$$

where $f(\theta)$ is the distribution function for the beta distribution. The probability of replicating a variant seen x times in N_{ind} individuals in an additional tN_{ind} individuals can then be calculated using (5). We now show applications to simulated examples, and to real data from the ENCODE project.

Results

Optimal design

In a first set of examples we show applications to the optimal design of variant discovery studies. In particular we show how the expected proportion of rare variants to be discovered in a study varies as a function of the mean coverage depth if we assume a fixed study cost and various values for the error probability, p_{err}, in equation (1). As shown in Figure 1 (second column), for a fixed cost, a mean coverage depth below or above the optimal depth level may lead to substantial reductions in the expected number of discoveries. Also shown is that, using the reasonable assumption that the cost function is linear in both the number of individuals and the average coverage depth, a larger number of individuals at a lower coverage is preferable to a lower number of individuals at higher coverage, if the goal is to maximize the number of variants detected.

On the other hand, if our goal is to discover as many variants as possible in a fixed number of individuals and the cost is not fixed, then a larger coverage depth will naturally result in a larger number of discoveries. However, there are diminishing returns as the coverage increases; and after a certain level is reached, each additional level of coverage leads to only a small increase in additional discoveries (Figure 1, third column).

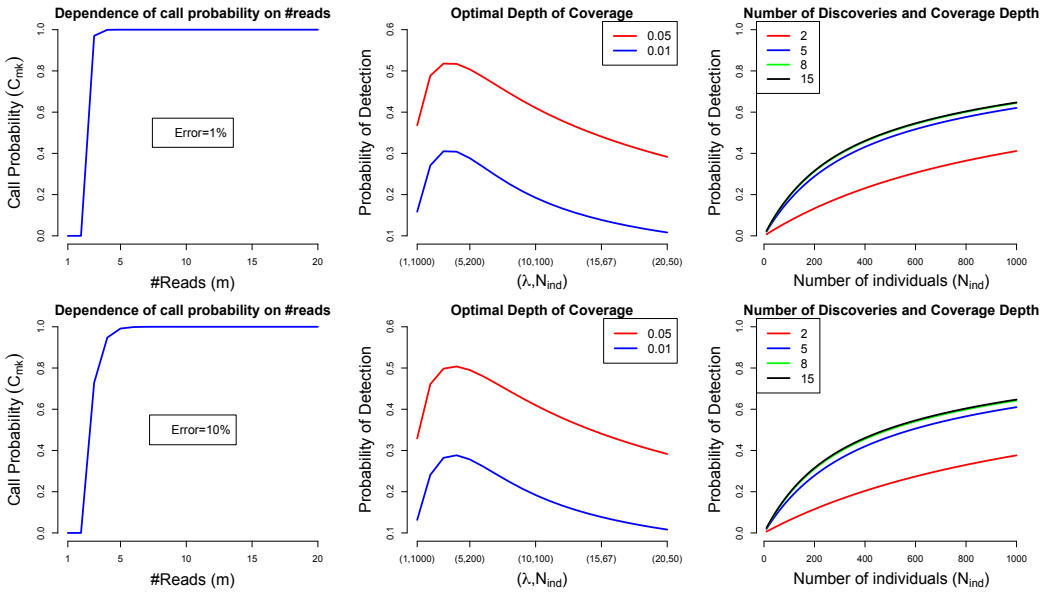

Figure 1: Optimal design of variant discovery studies. The error probability per base, p_{err}, is 0.01 (upper panel), and 0.10 (lower panel). The first column shows the probability of calling a variant, $C_{m,k}$, when the number of reads at a location is m and the number of reads needed without error, k, is 3. The middle column shows the expected proportion of rare variants (frequency less than 0.01 or 0.05) to be discovered as a function of mean depth of coverage, λ (and, implicitly, number of individuals, N_{ind}), when the cost of the study is fixed, and assumed to be equal to $\lambda \cdot N_{\text{ind}}$. The third column shows the expected proportion of rare variants (frequency less than 0.01) to be discovered as a function of the number of individuals sequenced, for various levels of the average coverage depth ($\lambda = 2,\ 5,\ 8,\ 15$) and assuming a cost function that increases linearly with the number of individuals.

DOI: 10.2202/1544-6115.1581

Probability of replication

Once a discovery study has been performed in a number of individuals, N_{ind}, it is of interest to inquire about the probability of replicating a particular variant in future datasets. In Figure 2 we show the probability of replicating a variant that was seen in only a few individuals in a discovery study (i.e. singletons, doubletons etc.). As shown, for singletons many more individuals than present in the original study are needed in order to achieve a high probability of replication (e.g. 10 times more the discovery dataset size for a replication probability of 93%). In contrast, for variants that occur twice or more the number of individuals required for replication is substantially smaller (e.g. 3 times more the discovery dataset size for a replication probability of 95%).

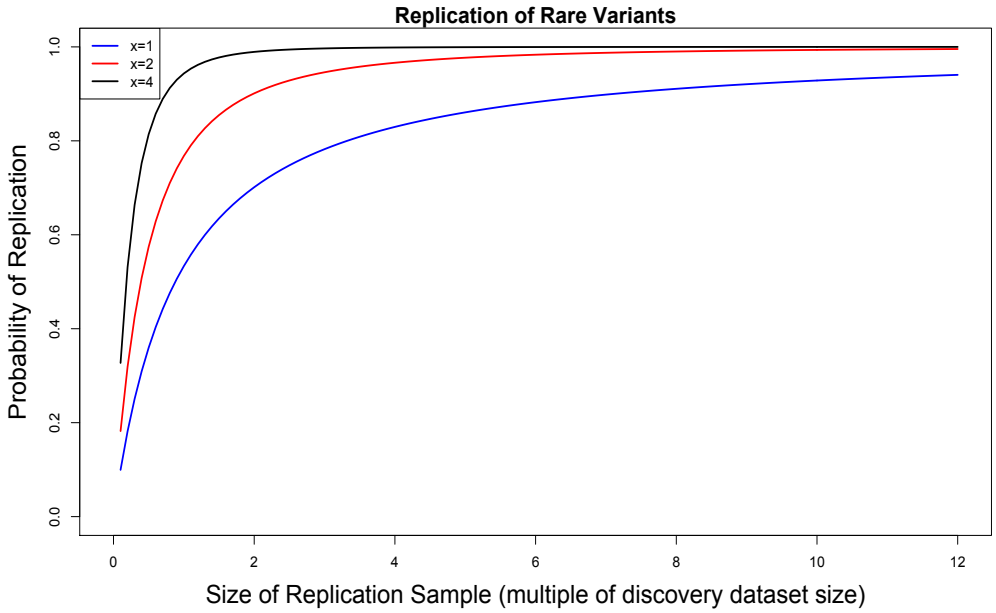

Figure 2: Probability of replicating a variant in a new set of individuals, given its observed frequency count ($x = 1,\ 2,\ 4$) in the original discovery dataset of size 50 individuals; $p_{\text{err}} = 0.10$ and $\lambda = 20$.

In addition, we have used the sequence data on 16 CEU individuals that are part of the ENCODE project[1], to show that the estimates for the

[1] http://www.hapmap.org/downloads/encode1.html.en

replication probability are sensible in a real study. To achieve this, we split the dataset on 16 individuals into two datasets, a discovery dataset and a replication dataset, with 8 individuals each, and compare the actual number of replicated discoveries with the estimated one for several classes of variants, e.g. singletons (variants that occur once in the discovery samples). As shown in Table 1, our predictions are fairly accurate in this application.

	n_1	n_2	n_3	n_4	n_5
Discovery Dataset	$1,855$	925	859	763	532
true #replications	993	812	826	711	526
estimated #replications	$1,020$	725	774	729	522

Table 1: Accuracy for the Probability of Replicating Variants detected in a Discovery Study: application to the CEU ENCODE data on 16 individuals. The discovery dataset and the replication dataset each contain 8 individuals. n_x is the number of variants that occur x times in the discovery dataset.

Number of genetic variants missed from a discovery study

We use the software package Genome (Liang et al. 2007) to simulate sequence data on $10,000$ haplotypes according to a coalescent model (Hudson 1990), resulting in a total of $43,305$ SNVs (single-nucleotide variants) in a large genomic region on a single chromosome. We form datasets of N_{ind} individuals (in our examples $N_{\text{ind}} = 50$, or 100 haplotypes) by randomly sampling from the generated haplotypes. For each haplotype the number of reads at a single location is Pois(λ), and the probability that we call a truly present variant depends on the actual number of reads at a location and on the number of reads without error, as explained before in (1). In Figure 3 we show that the beta-binomial model fits the observed frequency counts very well.

With simulated data we are in the unique position to be able to assess the accuracy of the estimates produced by our model. In particular, we use our approach to estimate the number of missing variants in the random sample of 100 haplotypes, and compare it with the true number of missing variants that can be directly assessed based on the remaining $9,900$ haplotypes. As shown in Table 2 the estimates on the missing variation are fairly accurate even if they are based on a sample of only 100 haplotypes.

DOI: 10.2202/1544-6115.1581

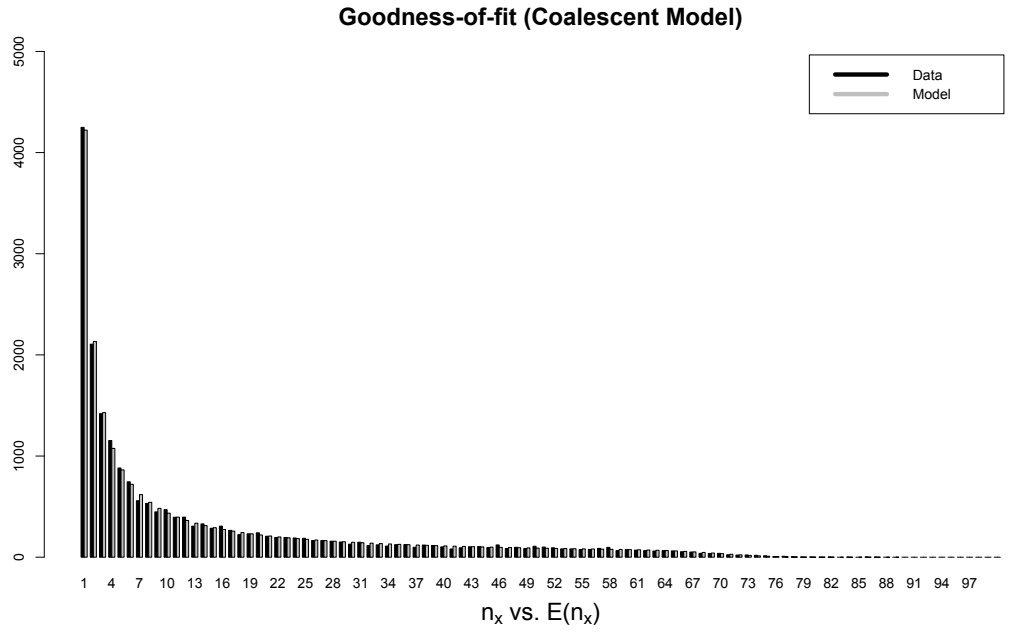

Figure 3: Coalescent-based simulated data with reduced coverage. Fit of the beta-binomial model to the observed frequency counts (n_x). The mean coverage depth is $\lambda = 8$, and the model relating the probability of calling a variant given the number of reads at a location (eq. 1) is specified by $p_{\text{err}} = 0.1$ and $k = 3$.

λ	#Missing Variants (T/E)	#Missing Variants $f \geq 0.01$ (T/E)
4	$22,187 \pm 104/21,508 \pm 454$	$1,602 \pm 52/1,714 \pm 24$
∞	$20,604 \pm 117/19,961 \pm 549$	$917 \pm 71/939 \pm 26$

Table 2: Accuracy of the estimates of missing genetic variation. T is the true number of remaining variants, i.e. present in the $10,000$ haplotypes and not seen in the 100 haplotypes in the discovery dataset; E is the estimated number of remaining variants. $\lambda = \infty$ corresponds to a perfect sensitivity case. Mean and SD are calculated based on 100 independent samplings of 100 haplotypes; $p_{\text{err}} = 0.1$ and $k = 3$.

Rare variants and coverage

An important class of variants is those that have low observed frequency count (i.e. singletons, doubletons) in a study, since they contain many of the rare variants (population frequency less than 1%) identified in a study. This is especially true with low coverage design and small discovery studies. In Table 3 we report for various values of λ, i.e. the coverage parameter, the proportion of (true) rare variants identified in $100 - 600$ chromosomes that have a small frequency count in the study, e.g. $P(x = 1|f \leq 0.01$ AND identified in the study). As shown, for low coverage, $68\% - 93\%$ of the rare variants identified in a reasonably sized study are expected to be singletons; for a higher coverage study only $39\% - 78\%$ of rare variants identified appear as singletons. Singletons, together with other variants with low frequency count, are therefore important to ascertain if the main goal is to identify rare variants. Such variants can be validated by subsequent genotyping in the discovery study.

N_{ind}	λ	$x = 1$	$x = 2$	$x = 3$	$x = 4$
50	8	0.78	0.17	0.03	0.01
50	4	0.84	0.14	0.02	0.00
50	2	0.93	0.06	0.01	0.00
300	8	0.39	0.22	0.14	0.09
300	4	0.46	0.24	0.14	0.08
300	2	0.68	0.22	0.07	0.02

Table 3: The proportion of rare variants detected in a study which have a small frequency count: $x = 1 - 4$ as a function of the coverage parameter λ (i.e. $P(x = 1|f \leq 0.01$ AND identified in the study)).

Discovery studies in multiple populations

We have applied the methods here to the ENCODE data[2] to investigate how combining different populations in a discovery study affects the overall number of discoveries. Sequence data in ten 500-kilobase regions of the genome were available for 39 unrelated DNA samples: 8 Yoruba (YRI), 16 CEPH European (CEU), 7 Han Chinese (CHB), and 8 Japanese (JPT). These regions were chosen to be representative of the genome in general, including various chromosomes, recombination rates, gene density and values of non-transcribed conservation with mouse.

[2]http://www.hapmap.org/downloads/encode1.html.en

DOI: 10.2202/1544-6115.1581

We have compared the observed number of variants identified by combining individuals from any two populations at a time (shown here are CEU and YRI; CEU and CHB; CHB and JPT) with the estimated number of discoveries, if the discovery dataset consisted of the same number of individuals from only one of the two populations. To estimate the number of new variants we expect to find if we had additional individuals from the same population we used $\Delta_{N_{\text{ind}},0,1,\infty}(t)$ (eq. (4)). In Table 4 we show that, as expected, adding more individuals from a different population (rather than the same population as in the discovery study) may lead to an increase in the overall number of discoveries. This is particularly striking when combining individuals from CEU and YRI, and comparing with only CEU individuals. More precisely, the observed number of discoveries in the combined CEU and YRI datasets is $13,968$ compared to only $10,766$ estimated in the same number of CEU individuals. While many of the additional discoveries may be specific to the YRI population, some will also be polymorphic in the CEU population.

Datasets	16 CEU	8 YRI	16 CEU + 8 YRI	16 CEU + 8 CEU	8 YRI + 16 YRI
#discoveries	$10,119$	$10,315$	$13,968$	$10,766^a$	$13,479^a$
Datasets	16 CEU	7 CHB	16 CEU + 7 CHB	16 CEU+ 7 CEU	7 CHB + 16 CHB
#discoveries	$10,119$	$7,802$	$11,466$	$10,700^a$	$9,278^a$
Datasets	7 CHB	8 JPT	7 CHB + 8 JPT	7 CHB + 8 CHB	8 JPT + 7 JPT
#discoveries	$7,802$	$7,804$	$8,613$	$8,790^a$	$8,425^a$

[a]Estimated

Table 4: Discovery Studies with individuals from multiple populations. Data on 8 YRI, 16 CEU, 7 CHB and 8 JPT individuals is available.

Discussion

Variant discovery studies using the new sequencing technologies are now underway. Such studies aim to identify low frequency variants that may explain the yet unaccounted for genetic variance (Manolio et al., 2009) for complex diseases. The amount of the underlying variation identified in such studies is influenced directly by the sensitivity to detect variants present in an individual, and the number of individuals sequenced. In this paper we proposed statistical approaches that can be useful in the context of such studies. In particular, we have been concerned with the optimal design of variant discovery

studies, i.e. how does one choose the optimal depth of coverage for a given study cost in order to maximize the expected number of discoveries. Such a choice inevitably depends on the sequencing technology employed, but we show here that, assuming a cost function linear in both the number of individuals and the coverage depth, sequencing a larger number of individuals at lower coverage depth will generally lead to more discoveries than sequencing a smaller number of individuals at deeper coverage. The assumptions we made on the cost function are reasonable at this time. It is however possible that in the future the cost of sequencing will increase slower than linearly as a function of λ. In Figure S1, we illustrate the case of a logarithmic cost function in λ, and show that the same overall conclusion still holds.

Our approaches are based on several reasonable model assumptions (e.g. linkage equilibrium, Beta-binomial model), and as such results may be sensitive to deviations from these assumptions. Therefore we sought it was important to apply the proposed methods to both coalescent-based simulated data and real sequence data from the ENCODE project. We showed in our applications that the proposed methods work well, and in particular that the predicted number of missing variants is very close to the true number.

With low coverage many of the rare variants will tend to appear as singletons. Since singletons are also enriched for sequencing errors it is important that such variants discovered in a study be confirmed by subsequent genotyping of the discovery data. Discovery studies with individuals in multiple populations can be useful as well. Having data from multiple populations in a discovery study not only ensures a more representative study of the global human variation, but may also increase the overall number of discoveries compared with individual populations. Therefore when there is population stratification, and thus individuals from several subpopulations are included in a study, more genetic variants may be discovered than for the same number of individuals in a single subpopulation.

We have focused here on the estimation of the number of true variants to be identified in a sequencing study, and have assumed that the probability of calling non-existent variants is small. This latter probability can be made arbitrarily small by adopting a more stringent variant calling rule, for example, by requiring a large number of the non-reference allele, k, to be observed at a variant position. Having fixed such a rule, the same methods we already described apply.

DOI: 10.2202/1544-6115.1581

Current sequencing studies focus on sequencing each individual at a time at an expected coverage, λ. An alternative design, suitable mostly when the sequencing error is low, is to sequence pools of DNA samples together (Futschik et al. 2010). DNA pooling could prove more cost-effective for variant detection in the future, and the methodology described here can be adapted easily to accommodate such a design.

The approach we proposed is appealing since it makes few and reasonable model assumptions, is flexible enough to take into account important covariates that influence variant detection (in this case the coverage depth), and at the same time provides fast and accurate estimates for the number of missing variants from a discovery study.

Software R code implementing these methods is freely available on the first author's webpage.

Appendix

Fitting the Beta-Binomial model

We assume that the observed frequency counts can be modeled using a Beta-Binomial model. The probability that at a position the frequency count is x is given by:

$$P_x = \int_0^1 \binom{2N_{\text{ind}}}{x} (\sum_m \theta \cdot C_{m,k} \cdot f_\lambda(m))^x (1 - \sum_m \theta \cdot C_{m,k} \cdot f_\lambda(m))^{2N_{\text{ind}}-x} f(\theta) d\theta$$

for $x \geq 0$. Since we only observe those variants that have frequency count 1 or above, the truncated probabilities are:

$$P_x^t = \frac{P_x}{\sum_{x=1}^{2N_{\text{ind}}} P_x}$$

for $x \geq 1$. The likelihood function can then be written as:

$$L(a,b) = \prod_{x=1}^{2N_{\text{ind}}} (P_x^t)^{n_x}$$

and the log-likelihood function is:

$$LL(a,b) = \sum_{x=1}^{2N_{\text{ind}}} n_x \log(P_x^t),$$

where we assume that the variants are in linkage equilibrium. We maximize $LL(a, b)$ to obtain the maximum-likelihood estimators (MLEs) for a and b. The maximization is carried out through the Newton-Raphson method.

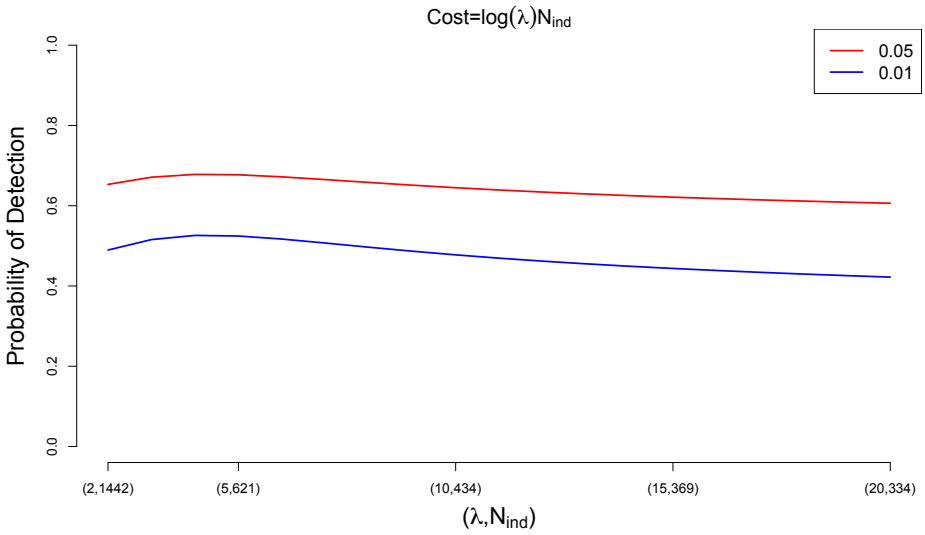

Figure S1: Cost function is logarithmic in λ (the average coverage depth). Shown is the expected proportion of rare variants (frequency less than 0.01 or 0.05) to be discovered as a function of mean depth of coverage, λ (and, implicitly, number of individuals, N_{ind}), when the cost of the study is fixed, and assumed to be equal to $\log(\lambda) \cdot N_{\text{ind}}$.

References

[1] Bromwich T (1955) An Introduction to the Theory of Infinite Series *2nd edition. London: Macmillan*

[2] Bunge J, Fitzpatrick M (1993) Estimating the number of species: a review *J Am Statist Assoc* 88: 364–373

[3] Efron B, Thisted R (1976) Estimating the number of unknown species: How many words did Shakespeare know? *Biometrika* 63: 435–437

[4] Futschik A, Schltterer C (2010) Massively Parallel Sequencing of Pooled DNA Samples–The Next Generation of Molecular Markers. *Genetics* to appear

DOI: 10.2202/1544-6115.1581

[5] Hudson RR (1990) Gene genealogies and the coalescent process. *Oxf. surv. evol. biol.* 7: 1–44

[6] Ionita-Laza I, Lange C, Laird NM (2009) Estimating the number of unseen variants in the human genome. *Proc Natl Acad Sci USA* 106: 5008–5013

[7] Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83: 311–321.

[8] Liang L, Zoellner S, Abecasis GR (2007) GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics* 23: 1565–1567.

[9] Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5: e1000384.

[10] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI et al. (2009) Finding the Missing Heritability of Complex Diseases *Nature* In Press

[11] Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31–46.

[12] Sekar CC, Deming WE (1949) On a Method of Estimating Birth and Death Rates and the Extent of Registration *J Am Statist Assoc* 44: 101–115

[13] Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135–1145

[14] Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* 85:142–154

[15] Wright S (1951) The general structure of populations. *Annals of Eugenics* 15: 323–354