BMC Genomics

BioMed Central

Research article

The theory of discovering rare variants via DNA sequencing

Michael C Wendl* and Richard K Wilson

Address: The Genome Center and Department of Genetics, Washington University, St. Louis MO 63108, USA

Email: Michael C Wendl* - mwendl@wustl.edu; Richard K Wilson - rwilson@wustl.edu

* Corresponding author

Published: 20 October 2009

BMC Genomics 2009, **10**:485 doi:10.1186/1471-2164-10-485

Received: 20 March 2009
Accepted: 20 October 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/485

Abstract

Background: Rare population variants are known to have important biomedical implications, but their systematic discovery has only recently been enabled by advances in DNA sequencing. The design process of a discovery project remains formidable, being limited to *ad hoc* mixtures of extensive computer simulation and pilot sequencing. Here, the task is examined from a general mathematical perspective.

Results: We pose and solve the population sequencing design problem and subsequently apply standard optimization techniques that maximize the discovery probability. Emphasis is placed on cases whose discovery thresholds place them within reach of current technologies. We find that parameter values characteristic of rare-variant projects lead to a general, yet remarkably simple set of optimization rules. Specifically, optimal processing occurs at constant values of the per-sample redundancy, refuting current notions that sample size should be selected outright. Optimal project-wide redundancy and sample size are then shown to be inversely proportional to the desired variant frequency. A second family of constants governs these relationships, permitting one to immediately establish the most efficient settings for a given set of discovery conditions. Our results largely concur with the empirical design of the Thousand Genomes Project, though they furnish some additional refinement.

Conclusion: The optimization principles reported here dramatically simplify the design process and should be broadly useful as rare-variant projects become both more important and routine in the future.

Background

Technological developments continue to dramatically expand the enterprise of DNA sequencing. In particular, the emergence of so-called "next-generation" instruments (NGIs) is opening a new chapter of genomic research [1]. If we characterize sequencing economy by the ratio of project speed to total project cost, NGIs are orders of magnitude superior to their traditional Sanger-based predecessors. Indeed, they are the first systems to demonstrate the economic feasibility of sequencing individual genomes on a large scale [2].

Future efforts will undoubtedly use NGIs to address issues in medical sequencing and personal genomics [3], but these instruments are also poised for major contributions at the population level [4,5]. For example, the Thousand Genomes Project (TGP) is focusing on comprehensive identification of variants in the human population

through cohort-level whole-genome sequencing using NGIs [6,7]. One of its main goals is to discover and characterize rare single nucleotide alleles, basically those present at minor allele frequencies around 1% or less. This region was not accessible to the earlier HapMap Project [8]. Rarer instances are obviously much more difficult to find and necessitate gathering enormously larger amounts of data. Such demands will obviously extend to any future such projects one might envision, including those for model organisms, agriculturally important species, cancer genomes, infectious agents, etc.

The success of such variation projects depends upon adequately understanding the relevant process engineering issues and subsequently crafting a suitable project design. One concern in traditional single-genome sequencing is the so-called "stopping problem" [9-11], which is the proposition of estimating what redundancy will suffice for a desired level of genomic coverage. Variation projects similarly require specification of a total, project-wide redundancy, R. Yet, because they necessarily involve multiple genomes, an essentially new design question also emerges. That is, how does one optimize the number of samples, σ, versus the redundancy allotted per sample, ρ, such that the probability of finding a rare variant, P_v, is maximized? The existence of such optima is intuitively clear. Heavily sequencing only a few samples will tend to miss a variant because it is unlikely to be present in the original sample set. Conversely, light sequencing of too many samples may overlook the variant by virtue of insufficient coverage for any samples actually harboring it. Somewhere between these extremes lie optimum combinations of parameters.

At present, this issue can only be addressed in *ad hoc*, fairly inefficient ways. For example, the TGP conducted both painstaking computer simulations and pilot sequencing phases involving hundreds of genomes to aid in designing the full-scale project [6,7]. While certainly informative, even such seemingly extensive data may not, by themselves, give a complete picture of optimization because combinations of the many underlying variables (Table 1) lead to an enormous solution space. We comment further on this aspect below. Existing theory is also ineffective because sequence coverage has not yet been considered [12].

Here, we examine optimization from a more focused mathematical perspective. Our treatment accounts for sequence errors via the proxy of a variable read covering count [3,13], but it omits secondary, project-specific details like software idiosyncrasies [14], instrument-specific biases [15], and alignment issues [16]. The solution leads to a set of general, though unexpectedly simple optimization principles, which correct some earlier speculation [17] and are useful as first approximations for actual projects. Because these rules appreciably narrow the solution space, they also offer good starting points for even more targeted numerical and empirical searches that might account for secondary effects, if such are deemed necessary.

Table 1: Variables in a Multi-Genome Variant Detection Project

variable†	meaning
P_v	probability of finding a rare variant
$P_{v,\,min}$	minimum acceptable value of P_v for a project
ρ	haploid per-sample sequence redundancy
R	total, project-wide redundancy
φ	frequency of variant in population
σ	number of samples sequenced in project
τ	minimum read coverings for detection
N	minimum variant observations to declare discovery

†Some variables are modified with a "star" superscript to denote optima, for example σ^* is the optimum sample size for a project, ρ^* the optimum per-sample redundancy, and P_v^* the discovery probability under optimal conditions.

Results

The term "rare variant" is routinely taken to mean a rare allele, although it can also mean a rare SNP genotype. Take φ to be the variant frequency, i.e. the minor allele frequency or the rare homozygous genotype frequency, as appropriate. We assume the TGP convention whereby samples are sequenced separately to uniform depths [6,7], instead of being pooled first. The general theory then encompasses the multiple-genome population sequencing problem and its subsequent design optimization.

Analytical Characterization of Discovery in Multiple Genomes

Theorem 1 (Allele Variants). Let D_A be the event that a rare allele is detected, i.e. found by the investigator in a sequenced diploid genome sample. Its probability is

$$P(D_A) = \phi P(C)\big(2 - \phi P(C)\big), \tag{1}$$

where

$$P(C) = 1 - \sum_{k=0}^{\tau-1} \frac{1}{k!}\left(\frac{\rho}{2}\right)^k e^{-\rho/2} \tag{2}$$

is the coverage probability of spanning the allele's genomic position on a chromosome with at least τ sequence reads. Let σ independent, randomly-selected samples each be sequenced uniformly to haploid depth ρ. Then, if K is the random variable representing the number of samples the variant is found in and if N is the minimum number of observations necessary to declare the var-

iant as being "discovered", the discovery event is defined as $K \geq N$ and its probability is

$$P_v(K \geq N) = \sum_{k=N}^{\sigma} \binom{\sigma}{k} P^k(D_A)\left[1 - P(D_A)\right]^{\sigma-k}. \quad (3)$$

Theorem 2 (Genotype Variants). The probability of D_G, the event that a rare genotype is detected in a sample, is

$$P(D_G) = \phi P^2(C), \quad (4)$$

and its discovery probability is again given by Eq. 3, except where D_G, replaces D_A.

Statement of the Optimization Problem

Variant discovery is a constrained optimization problem [18], which can be stated as follows. Given the biological parameter φ and project-specific design parameters R, σ, τ, $P_{v,\,min}$, and N, maximize the objective function P_v, subject to both the equality constraint

$$R = \sigma \cdot \rho, \quad (5)$$

and to the auxiliary constraint

$$P_v \geq P_{v,min}. \quad (6)$$

In practical terms, we want to most efficiently discover a variant at the lowest possible cost, as represented by R.

Although the problem is framed in terms of finding a single variant, actual projects are apt to be specified according to discovering a certain average number of rare variants. These scenarios are equivalent, as Eq. 6 also quantifies the expected fraction of variants that will be found in the project. For example, $P_{v,\,min} = 0.95$ indicates finding 95%, on average, of the variants occurring at some value of φ.

Optimizing for Single and Double Variant Observations

Leaving aside the optimization of ρ versus σ for a moment, the least obvious of the project-specific parameters to specify is arguably N. Higher values may exceed the actual number of instances in the sample set, resulting in *a priori* failure of the project. We will therefore concentrate on the experimentally relevant special cases $N = 1$ and $N = 2$. The former is clearly a minimum requirement, while the latter serves to better discern between a rare population variant and a SNP that is unique to an individual sample (a "private SNP").

Because we have an explicitly-defined equality constraint in the form of Eq. 5, the number of design variables can be reduced by one [18]. Specifically, substituting $\rho = R/\sigma$ into Eq. 2 allows us to write a constrained form of the coverage probability, which in turn furnishes constrained expressions for the probabilities of events D_A and D_G. It is expedient at this point to switch from the event-based notation of probability used up until now to the Eulerian (functional) convention for the calculus-based aspect of optimization. Specifically, let $f_{\tau,\,i}$ with $i \in \{A, G\}$ represent the now-constrained probabilities of D_A and D_G. (A detailed explanation of the switch in notation appears in "Mathematical Preliminaries".) We now state the following important optimization conditions.

Theorem 3 (Optimal Conditions). The optimum number of samples in a multiple-genome variation project for $N = 1$ is governed by the differential equation

$$\ln(1 - f_{\tau,i}) - \frac{\sigma}{1 - f_{\tau,i}} \cdot \frac{\partial f_{\tau,i}}{\partial \sigma} = 0, \quad (7)$$

and for $N = 2$ by the differential equation

$$\ln(1 - f_{\tau,i}) + \frac{f_{\tau,i}}{1 + (\sigma - 1) f_{\tau,i}} + \left[\frac{\sigma - 1}{1 + (\sigma - 1) f_{\tau,i}} - \frac{\sigma - 1}{1 - f_{\tau,i}}\right] \frac{\partial f_{\tau,i}}{\partial \sigma} = 0. \quad (8)$$

In particular, the roots of these equations in σ indicate maxima in P_v for rare alleles and genotypes. Each setting of the independent variables has one such optimum, σ^*, which is necessarily a global optimum.

Discussion

Finding rare variants is clearly an important aspect of both population and medical genetics [19]. The discovery process was not feasible before the advent of NGIs, but is now being actively prototyped through efforts like the TGP [6,7] and will likely become more routine in the future. This eventuality motivates examination of the problem from a general perspective, similar in spirit to theoretical treatments of single genomes [20]. The following sections report on both some of the broad trends across the design variable spectrum, as well as optimal conditions for the important special cases of $N = 1$ and $N = 2$.

General Trends

Fig. 1 shows P_v versus σ for variants appearing at 1% frequency for thresholds of $N = 1$ and $\tau = 2$. The latter appears to have emerged as the *de facto* choice to better control for sequencing errors [3,13]. Aside from the expected trend that performance improves as more data are gathered, the curves show two notable properties. First, σ^*, the sample size at which the maximum P_v occurs, increases with the project redundancy. This dependence means that a project cannot generally be optimized by

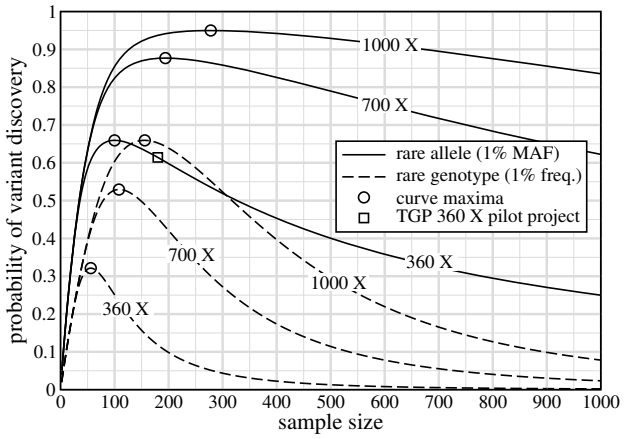

Figure 1
Probability of discovering variants at φ = 1% as a function of sample size for N = 1 and τ = 2. The single square datum represents the TGP pilot project at R = 360×. Circles indicate maxima for each curve.

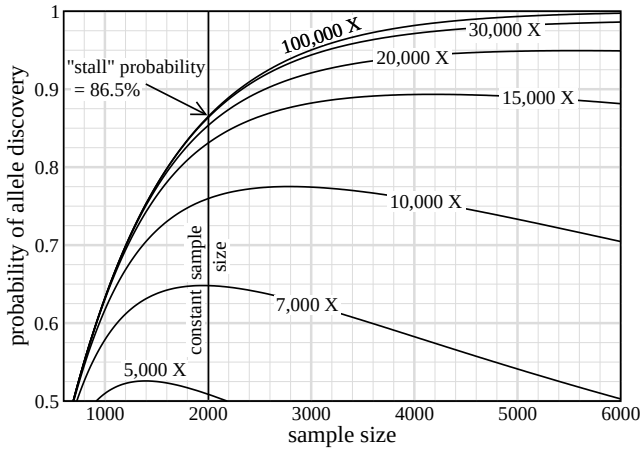

Figure 2
Gibbs' scenario [17]**of using a fixed 2,000 unit sample size to discover extremely rare alleles, φ = 0.0005, under N = 1 and τ = 2**. This hypothetical project is plotted for 5,000 ≤ R ≤ 100,000 and shows the conspicuous "stalling" effect that occurs under increasingly non-optimal conditions.

selecting σ in advance of other factors. Put another way, outright specification of σ almost certainly assures that the discovery process will not be optimal. We expand further upon this point below.

Fig. 1 also shows that curves are not symmetric with respect to σ^*. The rate of drop-off of P_v for a given project-wide redundancy is much more severe for $\sigma<\sigma^*$, implying that it is better to err in sequencing too many samples rather than too few. It is interesting to examine one of the TGP sequencing pilot phases in this context, which specifies 2× data collection for each of σ = 180 samples [6,7]. Here, $R = 2 \cdot 180 = 360$, which is one of the curves plotted in Fig. 1. Using the above thresholds, this design yields $P_v \approx 61\%$, whereas the optimal configuration returns $P_v \approx 66\%$ for only about 100 samples. Despite using almost twice as many samples as is optimal, this design remains relatively good, precisely because of the non-symmetric behavior.

Constant Sample-Size Designs and the Stalling Effect

The above discussion suggests that investigators should consider abandoning the idea of choosing σ outright. An earlier projection offers an interesting case study to further illustrate this point. Gibbs [17] postulated that σ = 2,000 samples would be a good way of discovering extremely rare variants occurring at 0.05%. (This number may simply have been an expeditious choice, as further details were not specified, nor was there any description of how this prediction was made.) Fig. 2 shows the implications of such a σ-based design. As R increases, σ^* marches to the right on the abscissa, eventually passing through the pre-selected σ = 2,000 at around R = 7,000. It continues rightward, leaving our fixed sample datum in the left-side wake of the optimum ($\sigma<\sigma^*$, as mentioned above), where the associated probability is now heavily penalized. In fact, the probability stalls at a value of roughly $P_v \approx 0.85$, regardless of the amount of additional data poured into the project.

Although this stalling effect may initially seem counter-intuitive, its explanation is quite straightforward. If we hold σ fixed while letting R increase without bounds, then ρ also grows without bounds (Eq. 5). In the limit, each sample will be perfectly sequenced, i.e. $P(C) \to 1$ in Eq. 2. Discovery is then simply a function of whether or not the variant is present in the original sample set. If so, it is absolutely certain to be discovered. The corresponding probabilities are then simple special cases of the model in Thms. 1 and 2. For example, for N = 1 observation of a rare allele we find

$$P_v(K \geq 1)\Big|_{\substack{\text{constant } \sigma, \\ R \to \infty}} \sim 1 - e^{-2\phi\sigma}, \qquad (9)$$

which is asymptotically identical to what is obtained if coverage is not considered at all [5]. The basic problem associated with constant sample-size designs is immediately apparent in this equation. Given small φ, the exponential term decays very slowly and can only be compensated for by increasing σ. The challenge, of course, is to do this such that P_v attains a maximum.

Remarks on Optimization Methods

We commented above that empirical prototyping and numerical simulation are unlikely to give complete insights to the general optimization problem because of the size of the solution space. Consider that the relationship between two parameters requires only a single curve on an X-Y plot, three parameters require a family of curves on one plot, four a textbook of family-type plots, and so forth. Richard Bellman, who developed the optimization technique of dynamic programming, called this phenomenon the "curse of dimensionality". Table 1 shows that we have 8 variables in our particular problem, however, even this is somewhat misleading because it does not consider the probabilistic nature of the problem. That is, P_v can only be established as an expected value through a sufficient number of repeated trials for each particular combination of the independent variables. This is the basic tactic used in simulation.

The population model in Thms. 1 and 2 improves matters considerably, furnishing P_v explicitly in terms of (τ, R, σ, φ, N). One could march through every combination of these variables, evaluating P_v for each, and log maxima that attain given levels of $P_{v,\,min}$. Though this approach would be enormously more efficient than naïve brute-force simulation, the calculations needed to adequately survey the floating-point "continuum" of the real-valued variables remain basically infeasible. Consequently, we still might not expect to discern any latent general laws.

The Weak Optimization Problem

We resort instead to Thm. 3, whose roots for $N = 1$ and $N = 2$ represent optimal sample sizes, σ^*. Let us first describe some unexpected properties found among the independent variables. These are important in that they furnish a direct solution to what might be called the *weak* optimization problem. This is the proposition that relaxes the condition defined by $P_{v,\,min}$. In effect, weak optimization provides the optimal probability, P_v^*, subject to a predetermined R rather than a given $P_{v,\,min} > 0$.

Fig. 3 shows σ^* versus R for representative parameter settings. Data collapse onto curves according to variant type. In each case, $\sigma^* = \sigma^*(R, \tau)$ and $\sigma^* \propto R$. These observations, coupled with $\sigma^* = R/\rho^*$ from Eq. 5 then imply $\sigma^*(R, \tau) = R/\rho^*(\tau)$. In other words, ρ^* is only a function of τ (Table 2). This is quite a significant finding because it immediately establishes the best sample redundancy for a project. In essence, this observation indicates that optimizable designs for rare variants are based on constant values of ρ rather than constant values of σ [17].

Table 2: Constants Associated with Optimum Designs

rare variant	τ	$\rho^*(\tau)$	$\lambda_\tau^\dagger$	$\beta_\tau^\dagger$
genotype	1	2.5	0.512	2.5
genotype	2	6.4	0.690	6.4
allele	2	3.6	0.537	1.8
allele	1	special case, see Eq. 10		

†See Eqs. 14, 15, 16

We emphasize that the numbers in Table 2 are not based on first-principles and are not strictly encoded in the governing equations. Rather, they are fortuitous empiricisms, restricted to the parameter values that characterize rare-variant projects. Fig. 4 demonstrates that, while ρ^* does indeed only depend upon τ up to allele frequencies of about 1%, it becomes a more complicated function of additional variables for higher frequencies.

Remarks on the Special Case of $\tau = 1$ for Rare Alleles

The case of $\tau = 1$ is conspicuously absent for rare alleles in Figs. 3 and 4 because its optimum sample size is not finite. Unlike the other cases, P_v approaches its maximum as $\sigma \to \infty$, for example

$$P_v(K \geq 1)\big|_{\substack{\text{constant } R, \\ \sigma \to \infty}} \sim 1 - e^{-\phi R}. \tag{10}$$

Here, we have the seemingly contradictory implication that we should spread a finite amount of sequence resources over the largest number of samples, each of which will then have a vanishingly small ρ. Mathemati-

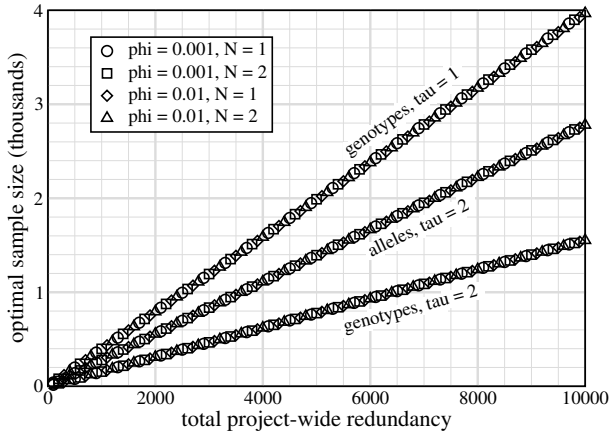

Figure 3
Optimal sample size versus project-wide redundancy for parameters representative of rare-variant projects.

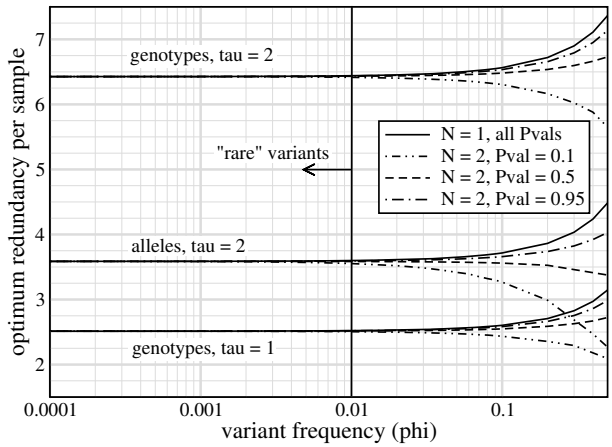

Figure 4
Optimum redundancy per sample, ρ^*, is essentially constant for each value of τ within the conventional range of $\varphi \leq$ 1% for rare variants.

cally speaking, the rate by which the per-sample $f_{1,\ A}$ decreases precisely offsets the favorable rate of increasing sample size, whereby P_v does not asymptotically vanish. However, there will usually be good secondary reasons for limiting σ in practice, e.g. cost of sample procurement. Moreover, conditions approach the limiting value rather quickly, for example setting $\rho = R/\sigma \leq 0.1$ brings P_v very close to the expression in Eq. 10. R is the main factor governing discovery under these conditions and its value can be calculated for a desired P_v by simply inverting Eq. 10.

Optimal Designs for Single and Double Variant Observations

The weak solution specifies constants of ρ^* (Table 2), which simultaneously imply σ^* for any choice of R. These properties subsequently fix certain relationships within the general problem, so that optimization for a desired $P_{v,\ min}$ in Eq. 6 reduces to the task of solving directly for φ (see Methods). Fig. 5 shows the resulting optimal designs for τ = 2, a setting characteristic of recent projects [3,13]. Results are plotted for $P_{v,\ min}$ = 95%, the same threshold set by the TGP. All curves show a special kind of log-log relationship between φ and R^* in which the slope is -1. In other words, optimal designs can be expressed as a family of log-log curves having the form $\varphi\ R^* = C(N,\ \tau,\ P_{v,\ min})$, where C is a so-called *optimization coefficient* for each combination of the variables. Of course, knowing C immediately enables one to compute R^* and subsequently $\sigma^* = R^*/\rho^*$ for a desired φ, which is of enormous practical value for project design. Table 3 shows C for the configurations having well-defined optimum redundancies, although we note that Eq. 10 also follows this form, having C = 3.0. R^* is indicative of the total resources a project requires, so C is also useful in comparing relative costs. For example, requiring two observations of a rare allele instead of just one would only be, somewhat counterintuitively, about 60% more expensive if both schemes were to be done optimally.

Table 3: Optimization Coefficients for 95% Discovery Probability

variant	τ	N	C
allele	2	1	10.0
allele	2	2	15.8
genotype	1	1	14.7
genotype	1	2	23.3
genotype	2	1	27.8
genotype	2	2	44.1

Consider the example of the TGP, whose sizable *ad hoc* design effort was already mentioned above. For $N = \tau = 2$ at the 95% level, Table 3 indicates C = 15.8. Assuming 1% rare allele discovery, optimal processing calls for roughly 440 samples sequenced to 3.6× each, for a project total of R = 1580×. Given the long-standing convention of specifying ρ in whole units, these results largely confirm the TGP design, though in a more precise fashion. That is, TGP has only winnowed the sample size to 400-500 per population cluster, with each sample sequenced to 4× [6,7]. The associated P_v curve is relatively flat in $400 \leq \sigma \leq 500$, but this imprecision, coupled with a round value of ρ, still imposes a degree of cost liability. For instance, on the outer end, the project would expend $4 \cdot 500 = 2000\times$ in data, roughly 25% more than that required for 95% confidence. Project modifications are readily analyzed, for

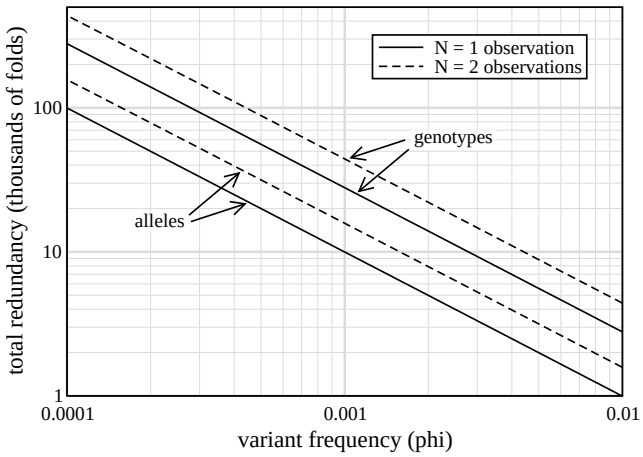

Figure 5
Log-log plot of optimum designs, R^* vs. φ, for discovering rare genotypes and alleles at probabilities of 95% when requiring at least two read coverings (τ = 2).

example, reaching alleles down to $\varphi = 0.5\%$ would simply require doubling the project: about 880 samples with $R = 3160\times$. Analysis of genotypes is now similarly trivial.

Conclusion

Sequence variation is often called the "currency" of genetics [4] and whole-genome sequence variation projects, enabled by continuing advances in technology, will likely become both increasingly important and routine in biomedical research. Although finding common occurrences is no longer considered to be very difficult, rare ones remain challenging because of the significantly larger amounts of data that must be gathered. Process optimization has to be considered much more carefully here. We have reported a general, though remarkably simple set of optimization principles based on analyzing the population sequencing problem. Results largely confirm the design of a special case, that of the TGP, but also permit immediate analysis of a much broader set of possible project requirements. Derivation of optimal conditions for even higher N and/or τ would be a mechanical, albeit not entirely trivial extension of the mathematics shown here, but the experimental feasibility of such designs for future projects remains unclear.

Population structure is another consideration, as rare variants are likely to be associated with particular geographic regions and their sub-populations [4]. A few issues are relevant here. First, some studies target the variation underlying specific phenotypes [21], but variant discovery projects do not place strong preference on the kinds of variation that are sought. Second, ρ^* is not a function of rareness (Fig. 4), meaning that latent population-related differences in frequency will not ruin optimality. One should simply treat each desired sub-population separately, making no differential adjustments to per-sample redundancies. This strategy assures discovery of population-specific variants and, incidentally, is precisely what the TGP is following.

Methods

Mathematical Preliminaries

This section expands on some of the mathematical esoterica involved in establishing the theory.

Chain Rule

This principle enables one to find the derivative of a function that itself depends on another function [22]. In essence, it establishes a product rule for the respective derivatives. For example, if $y = z^3$ and $z = x^2 + 1$, then $dy/dx = dy/dz \cdot dz/dx = 3z^2 \cdot 2x = 6x(x^2 + 1)^2$. Chain Rule is used in the logarithmic differentiation process described below.

Independently and Identically Distributed (IID)

This term means that all random variables in a collection are independent of one another, i.e. they have no mutual influences or relationships, and that each has the same probability as all the others [23]. Coin flipping is a simple example. The current flip is not influenced by past ones, nor does it influence future ones, and each flip has the same probability of showing, say, "heads". This concept is the basis of ultimately establishing the binomial nature of P_v in Theorem 1.

Logarithmic Differentiation

This mathematical device employs the Chain Rule (see above) to differentiate functions whose forms render them difficult to handle using more basic rules. Proof of Theorem 3 (below) requires this treatment because the independent variable being differentiated against appears in the exponent. An illustrative example having precisely the same issue is $y = e^x$, which is readily shown by this procedure to be its own derivative. Applying Chain Rule to the logarithmic form, $\ln y = x$, yields $y^{-1} \cdot dy/dx = 1$, from which $dy/dx = y = e^x$ immediately follows.

Notation

This aspect is complicated by the fact that the theory straddles two different branches of mathematics: probability and calculus. In the former case, notation is primarily concerned with specifying configurations of events, while in the latter, Euler's convention is used to describe functional dependence on a set of independent variables. This necessitates a change in notation as we move from the probabilistic discovery model in Thms. 1 and 2 to the calculus-based optimization process in Thm. 3.

Substituting the constraint in Eq. 5, $\rho = R/\sigma$, into Eq. 2, we can write the *constrained* form of the coverage probability as

$$1 - \sum_{k=0}^{\tau-1} \frac{1}{k!} \left(\frac{R}{2\sigma} \right)^k e^{-R/(2\sigma)}, \qquad (11)$$

which now depends upon τ, R, and σ. In turn, this expression is substituted into Eqs. 1 and 4 to obtain constrained probabilities for events D_A and D_G, respectively, with dependence now extending to φ, as well. From here on, let us consider these event probabilities simply as mathematical functions. For example, $f_{1,G}$ is the expression obtained by setting $\tau = 1$ in Eq. 11, squaring it, and multiplying by φ, i.e. it is the constrained probability of the event D_G originally introduced in Eq. 4. Under this notation, we can then easily represent *all* such functions universally by writing them in a form $f_{\tau,i} = f_{\tau,i}(\varphi, R, \sigma)$, where $i \in \{A, G\}$. This is the convention we follow in both Thm. 3 (above) and its proof (below).

Roots of a Function

Roots are values of the independent variable which cause a function to vanish, i.e. to be equal to zero. For example, $y = x^2 - 9$ can be factored as $y = (x + 3)(x - 3)$, showing that $x = \pm 3$ are the roots for which $y = 0$. This concept is relevant to the proof of Theorem 3 (below) because maxima within the P_ν family of functions occur at roots in σ of the first derivatives. Roots play a similar role in solving Eqs. 15 and 16.

Proofs of Theorems 1 to 3

Theorem 1: Let A_j and C_j be the events, respectively, that an allele variant exists on chromosome j in a sample at location x and that x is spanned (covered) by at least τ sequence reads. The detection event is $D_A = (A_1 \cap C_1) \cup (A_2 \cap C_2)$. Given the presumed IID (see "Mathematical Preliminaries") nature of alleles and coverage with respect to chromosomes, $\varphi = P(A_1) = P(A_2)$ and $P(C) = P(C_1) = P(C_2)$, from which Eq. 1 follows directly. Eq. 2 is a corollary of diploid covering theory [24]. Finally, with respect to any given sample, D_A is a Bernoulli process: an allele is either detected, or it is not. Given uniform ρ for each sample and the random selection of presumably independent genomes, the process is IID. The distribution of detected variants is then binomial [23], from which Eq. 3 follows directly.

Theorem 2: Let G represent the existence of a rare genotype in a sample. Since both alleles must be discerned, the detection event is $D_G = G \cap C_1 \cap C_2$. Because coverage of x is not a function of whether the genotype is actually present and *vice versa*, G and $C_1 \cap C_2$ are independent, whereby Eq. 4 follows directly.

Theorem 3: The optimization problem is cast by substituting the single-sample detection probability, $f_{\tau,i}$ (see "Mathematical Preliminaries"), into the project-wide discovery probability, $P_\nu(K \geq N)$ in Eq. 3. Noting that $f_{\tau,i}$ and P_ν are both functions of σ (among other variables), but omitting the functional notation, this process gives

$$P_\nu(K \geq 1) = 1 - \left[1 - f_{\tau,i}\right]^\sigma \tag{12}$$

$$P_\nu(K \geq 2) = P_\nu(K \geq 1) - \sigma f_{\tau,i}\left[1 - f_{\tau,i}\right]^{\sigma-1} \tag{13}$$

for the special cases of interest, $N = 1$ and $N = 2$, respectively.

Roots in σ of the first derivatives of these equations are a necessary condition in identifying the extrema of P_ν[22]. Their forms require us to use logarithmic differentiation. (This procedure and the concept of roots are both outlined in the "Mathematical Preliminaries" section above.) Setting the resulting derivatives equal to zero gives the corresponding characteristic equations

$$\frac{\partial P_\nu}{\partial \sigma} = \left(P_\nu - 1\right)\left[\ln(1 - f_{\tau,i}) - \frac{\sigma}{1 - f_{\tau,i}} \cdot \frac{\partial f_{\tau,i}}{\partial \sigma}\right] = 0$$

and

$$\frac{\partial P_\nu}{\partial \sigma} = \left(P_\nu - 1\right)\left[\ln(1 - f_{\tau,i}) + \frac{f_{\tau,i}}{1 + (\sigma - 1) f_{\tau,i}} + \left(\frac{\sigma - 1}{1 + (\sigma - 1) f_{\tau,i}} - \frac{\sigma - 1}{1 - f_{\tau,i}}\right)\frac{\partial f_{\tau,i}}{\partial \sigma}\right] = 0$$

for $N = 1$ and $N = 2$, respectively. In general, $P_\nu \neq 1$ in either case, so the conditions must instead be satisfied by the terms in square brackets. Eqs. 7 and 8 follow directly.

The fact that there is only a single, global optimum, σ^*, for each case is a consequence of P_ν being a unimodal function in σ. In general, P_ν vanishes monotonically for $\sigma > \sigma^*$ because $P(C) \to 0$, and consequently $f_{\tau,i} \to 0$, as σ is increased under finite values of R. The exception is $f_{1,A}$, for which P_ν asymptotically approaches a maximum (Eq. 10).

Solution of the General Optimization Problem

Optimal conditions are described by constants of ρ^*, which can be substituted into the single-sample probability to obtain an optimized $f^*_{\tau,i}$. For $N = 1$, we can then derive the following expression, valid for both alleles and genotypes, directly from Eq. 12

$$\phi = \frac{1 - \left[1 - P_{\nu,min}(K \geq 1)\right]^{\beta_\tau / R}}{\lambda_\tau}, \tag{14}$$

where constants λ_τ and β_τ are given in Table 2. This equation describes the relationship between φ and R under optimal conditions when given user-specified values of τ and $P_{\nu,min}$. For $N = 2$, we cannot readily obtain an explicit optimization rule from Eq. 13. Instead, we cast the relationship as a root-finding problem in φ for genotypes as

$$\left[1 + \left(\frac{R}{\rho^*(\tau)} - 1\right)\phi\lambda_\tau\right]\left[1 - \phi\lambda_\tau\right]^{R/\rho^*(\tau) - 1} + P_{\nu,min}(K \geq 2) - 1 = 0 \tag{15}$$

and for alleles as

$$\left[1+\left(\frac{R}{\rho^*(\tau)}-1\right)\left(2-\phi\lambda_\tau\right)\phi\lambda_\tau\right]\left[1-\phi\lambda_\tau\right]^{R/\beta_\tau-2} + P_{v,min}(K\geq 2)-1=0. \quad (16)$$

That is, given τ, R, and $P_{v,\,min}$ the values of φ under which the process is optimal are the roots of Eqs. 15 and 16.

Derivatives and Numerical Method

Eqs. 7 and 8 depend upon partial derivatives of $f_{\tau,\,i}$. For rare alleles and genotypes, i.e. $i \in \{A, G\}$, we follow standard rules of differentiation [22] to obtain

$$\frac{\partial f_{2,A}}{\partial\sigma}=\frac{\phi R^2}{2\sigma^3}e^{-R/(2\sigma)}\left(\phi\left[1-\left(1+\frac{R}{2\sigma}\right)e^{-R/(2\sigma)}\right]-1\right) \quad (17)$$

$$\frac{\partial f_{1,G}}{\partial\sigma}=\frac{\phi R}{\sigma^2}\left(e^{-R/\sigma}-e^{-R/(2\sigma)}\right) \quad (18)$$

$$\frac{\partial f_{2,G}}{\partial\sigma}=\frac{\phi R^2}{2\sigma^3}\left[e^{-R/\sigma}\left(1+\frac{R}{2\sigma}\right)-e^{-R/(2\sigma)}\right]. \quad (19)$$

Note that an equation for $f_{1,\,A}$ is absent because the case of τ = 1 for rare alleles does not have a well-defined optimum (Eq. 10).

Eqs. 7, 8, 15, and 16 all depend upon the concept of finding the roots of an equation. (See "Mathematical Preliminaries" above.) Although none is readily factorable, they can be solved by the bisection algorithm, which is straightforward to apply, has reasonably good convergence behavior, and is extremely robust [25].

Abbreviations

NGI: next-generation sequencing instrument; TGP: Thousand Genomes Project.

Authors' contributions

MCW conceived and constructed the mathematical theory and wrote the paper. Both authors approved the final manuscript.

Acknowledgements

The authors wish to thank Ken Chen, Li Ding, Elaine Mardis, and John Wallis for comments on the draft manuscript, as well as the referees for their suggestions related to making the mathematical content more accessible. This work was partially supported by grant HG003079 from the National Human Genome Research Institute (R. K. Wilson, PI).

References

1. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends in Genetics* 2008, **24:**133-141.
2. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nature Methods* 2008, **5:**1005-1010.
3. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, *et al.*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456:**66-72.
4. Chakravarti A: **Population genetics -- Making sense out of sequence.** *Nature Genetics* 1999, **21:**56-60.
5. Zwick ME, Cutler DJ, Chakravarti A: **Patterns of genetic variation in Mendelian and complex traits.** *Annual Review of Genomics and Human Genetics* 2000, **1:**387-407.
6. Kaiser J: **A plan to capture human diversity in 1000 genomes.** *Science* 2008, **319:**395.
7. Siva N: **1000 genomes project.** *Nature Biotechnology* 2008, **26:**256.
8. Gibbs RA, Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, *et al.*: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
9. Wendl MC, Waterston RH: **Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing.** *Genome Research* 2002, **12:**1943-1949.
10. Wendl MC, Barbazuk WB: **Extension of Lander-Waterman theory for sequencing filtered DNA libraries.** *BMC Bioinformatics* 2005, **6:**. article no. 245.
11. Wendl MC: **Random covering of multiple one-dimensional domains with an application to DNA sequencing.** *SIAM Journal on Applied Mathematics* 2008, **68:**890-905.
12. Li B, Leal SM: **Discovery of rare variants via sequencing: Implications for association studies [abstract].** *Genetic Epidemiology* 2008, **32:**702.
13. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhi-jani V, Roth GT, *et al.*: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.
14. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER: **PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data.** *Genome Research* 2007, **17:**659-666.
15. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, *et al.*: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biology* 2009, **10:**. article no. R32.
16. Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C: **An analysis of the feasibility of short read sequencing.** *Nucleic Acids Research* 2005, **33:**. article no. e171.
17. Gibbs R: **Deeper into the genome.** *Nature* 2005, **437:**1233-1234.
18. Vanderplaats GN: *Numerical Optimization Techniques for Engineering Design* New York NY: McGraw-Hill; 1984.
19. Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IPM, Mortensen NJM, Bodmer WF: **Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas.** *Proceedings of the National Academy of Sciences* 2004, **101:**15992-15997.
20. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: A mathematical analysis.** *Genomics* 1988, **2:**231-239.
21. Halushka MK, Fan JB, Bentley K, Hsie L, Shen NP, Weder A, Cooper R, Lipshutz R, Chakravarti A: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.** *Nature Genetics* 1999, **22:**239-247.
22. Courant R: *Differential and Integral Calculus Volume I*. New York NY: Interscience; 1937.
23. Feller W: *An Introduction to Probability Theory and Its Applications* 3rd edition. New York NY: John Wiley & Sons; 1968.
24. Wendl MC, Wilson RK: **Aspects of coverage in medical DNA sequencing.** *BMC Bioinformatics* 2008, **9:**. article no. 239
25. Hamming RW: *Numerical Methods for Scientists and Engineers* New York NY: McGraw-Hill; 1962.