Identifying Rare Variants With Optimal Depth of Coverage and Cost-Effective Overlapping Pool Sequencing

Chang-Chang Cao, Cheng Li, Zheng Huang, Xin Ma, and Xiao Sun*

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China

ABSTRACT: Genome-wide association studies have identified hundreds of genetic variants associated with complex diseases although most variants identified so far explain only a small proportion of heritability, suggesting that rare variants are responsible for missing heritability. Identification of rare variants through large-scale resequencing becomes increasingly important but still prohibitively expensive despite the rapid decline in the sequencing costs. Nevertheless, group testing based overlapping pool sequencing in which pooled rather than individual samples are sequenced will greatly reduce the efforts of sample preparation as well as the costs to screen for rare variants. Here, we proposed an overlapping pool sequencing to screen rare variants with optimal sequencing depth and a corresponding cost model. We formulated a model to compute the optimal depth for sufficient observations of variants in pooled sequencing. Utilizing shifted transversal design algorithm, appropriate parameters for overlapping pool sequencing could be selected to minimize cost and guarantee accuracy. Due to the mixing constraint and high depth for pooled sequencing, results showed that it was more cost-effective to divide a large population into smaller blocks which were tested using optimized strategies independently. Finally, we conducted an experiment to screen variant carriers with frequency equaled 1%. With simulated pools and publicly available human exome sequencing data, the experiment achieved 99.93% accuracy. Utilizing overlapping pool sequencing, the cost for screening variant carriers with frequency equaled 1% in 200 diploid individuals dropped to at least 66% at which target sequencing region was set to 30 Mb. 

KEY WORDS: cost-effective; group testing; overlapping pool sequencing; rare variants

Introduction

As we all know, the genome-wide association studies (GWAS) is one of the most powerful methods for the studies of common diseases and complex traits [Hirschhorn and Daly, 2005] which have revealed numerous associations between diseases and genetic variants in the past 10 years [Duerr et al., 2006; Hampe et al., 2006; Klein et al., 2005; Ozaki et al., 2002]. However, most variants identified confer only a small proportion of heritability and results also suggest that rare variants may have a large effect as genetic risk factors for complex genetic diseases [Manolio et al., 2009]. Several researches have verified the evolution and functional impact of rare variants on some human traits and showed that screening individuals for rare variants becomes increasing important for associating rare variants with complex traits, investigating target biology and drug response, providing clinically actionable information [Nelson et al., 2012; Tennessen et al., 2012].

Due to the relatively low frequency of rare variants, large sample sizes are required to guarantee efficiency observations. Although over the past decade, massively parallel sequencing (MPS) platforms have become widely available, reducing the cost of DNA sequencing by over two orders of magnitude [Shendure and Ji, 2008]. MPS is still unaffordable for large-scale resequencing in detecting rare variants. Because most functional variants are thought to be located within exons, exome sequencing therefore provides a cost-effective method for identifying most of the functional variants [Nielsen, 2010]. To further reduce the cost of large-scale resequencing for detecting rare variants, another efficient approach is to sequence a large number of individuals together on a single sequence run. Simply sequencing pooled DNA samples is more cost-effective as it can fully make use of the high-throughput of MPS [Druley et al., 2009]. However, pooled sequencing makes it impossible to determine which individual contributed any rare DNA variants that are detected. The most straightforward approach to determine individual genotypes is to ligate a short, sample-specific DNA sequencing called barcode to every sample. However when sample size is large, this approach will be infeasible due to budget constraints [Patterson and Gabriel, 2009]. Borrowing the idea of group testing (GT) [Ding-Zhu and Hwang, 2000], samples are pooled regularly and then sequenced,
overlapping pool sequencing allows using fewer pools to identify rare variant carriers among thousands of individuals that could vastly reduce the efforts of sample preparation as well as the costs [He et al., 2011].

The concept of GT can date back to World War II, which was first proposed by Dorfman [Dorfman, 1943], for the problem of determining which blood samples contain the syphilis antigen (defective also called positive samples) for numerous soldiers. For now, many biological applications benefit from the GT such as blood testing [Ding-Zhu and Hwang, 2000; Dorfman, 1943; Sobel and Groll, 1959], HIV testing [Hughes-Oliver, 2006; Kim et al., 2007; Westreich et al., 2008], clone library screening [Balding and Torney, 1997; Bruno et al., 1995; Knill et al., 1996], protein–protein interaction mapping [Jin et al., 2007; Jin et al., 2006; Vermeirssen et al., 2007; Xin et al., 2009], drug screening [Jones and Zhigljavsky, 2001; Kainkaryam and Woolf, 2008; Kainkaryam and Woolf, 2009; Wilson-Lingardo et al., 2006] and population genotyping [Erlich et al., 2009a; Erlich et al., 2009b; Patterson and Gabriel, 2009; Prabhu and Pe’er, 2009]. Combine with GT, fewer DNA libraries are required to detect rare variants.

In 2009, Erlich et al. [Erlich et al., 2009a] designed a GT algorithm called DNA Sudoku based on the Chinese remainder theorem, and implemented their strategy in an experiment using 40,000 bacterial clones of amiRNAs targeting Arabidopsis genes and only 1,900 tests. Prabhu and Pe’er [Prabhu and Pe’er, 2009] conducted an overlapping pool sequencing experiment by downloading sequences from the 1000 Genomes project (http://www.1000genomes.org) and simulated pooling in silico but they used a design which was only able to detect and identify “singletons”—a variant occurring in just one sample—with very good reliability. Unfortunately, neither considered the cost of pooled sequencing.

In this paper, we proposed overlapping pool sequencing with the purpose to identify rare variant carriers cost-effectively. We first formulated the cost model of this strategy and the depth model for conducting pooled sequencing successfully. Based on these two models, design parameters can be chosen to minimize sequencing cost when using shifted transversal design (STD) algorithm [Thiéry-Mieg, 2006]. Second, due to the mixing constraint and high depth for pooled sequencing, with repeated blocks design, sequencing data requirement dropped dramatically and the number of blocks can also be optimized to minimize the cost without decreasing efficiency. At last, we applied the strategy with simulated pooled human exome sequencing to detect single base mutations with frequency equaled 1%. The simulation experiment showed the feasibility and potentiality of overlapping pool sequencing.

**Methods**

**Framework and Cost of Overlapping Pool Sequencing**

In the classical GT model we have a population with $n$ samples containing at most $d$ positive samples, denoted as $(n, d)$. The basic problem of GT is to identify all positive samples with a small number of group tests in the case of at most $E$ test errors. Each group test, also called a pool, is a subset of samples. It is assumed a group test $A$ has two possible outcomes: negative and positive. The outcome is positive if and only if $A$ contains at least one positive sample and is negative otherwise.

GT algorithms can be roughly divided into two types, sequential and nonadaptive. A sequential algorithm conducts the tests one by one and the outcomes of all previous tests can be used to set up the later tests. A nonadaptive algorithm specifies that all tests are conducted simultaneously which is more time-saving. Up to now, there has been a number of sequential, nonadaptive GT algorithms [Balding et al., 1996; Ding-Zhu and Hwang, 2000; Dyachkov and Rykov, 1983; Ngo and Du, 2000] and decoding algorithms [Chen and Hwang, 2008].

Due to the feature of some biological experiments, all tests should be conducted simultaneously, therefore, only the nonadaptive GT algorithms suit. Traditionally, a nonadaptive GT is also called a pooling design. A pooling design with $n$ samples and $t$ pools is associated with a $t \times n$ binary matrix $M = \{m_{ij}\}$, in which the rows are indexed by pools $A_1, \ldots, A_t \subset \{1 \ldots n\}$, the columns are indexed by samples $S_1, \ldots, S_n$, and $m_{ij} = 1$ if and only if the $j$th sample is contained in the $i$th pool.

For example, seven samples with one positive, the matrix can be designed as below:

$$M = \begin{bmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1
\end{bmatrix}$$

This matrix consists of three pools: $[4, 5, 6, 7], [2, 3, 6, 7]$ and $[1, 3, 5, 7]$. Suppose the binary outcome vector is $[0, 1, 1]^T$ of which 1 stands for positive and 0 for negative, then we can infer that the third sample is positive.

The objective of traditional GT is to minimize the number of pools $t$ in the presence of a few errors and the whole cost $C$ is a linear function of $t$:

$$C = tP + C_p$$

(1)

where $P$ denotes the price of a group test, $C_p$ denotes the cost of pooling includes tips, robot time and plastics which scale with the number of “1” in the pooling matrix.

Variant carriers are treated as positive samples and normal individual as negative for applying GT in MPS, i.e., overlapping pool sequencing to screen rare variants. DNA samples are pooled and sequenced together and a sample can be classified as positive or negative by analyzing the sequencing reads. The cost $C$ therefore covers library preparation and data production, which is a function of two variables: the number of pools $t$ and the amount of sequencing data $N_d$:

$$C = tP + N_dP + C_p$$

(2)

where $P$ denotes the price of library preparation in MPS and $P_d$ is the price of data production:

$$N_d = \sum_{i=1}^{t} D_i \times R$$

(3)
where $D_i$ is the mean depth of coverage for the $i$th pool and $R$ denotes the length of sequencing region.

Each individual is sequenced multitimes in pooling design, as a result, the data capacity increases dramatically which cannot be ignored any more. Hence, for overlapping pool sequencing, pooling design should be optimized to minimize the whole cost including the library preparation and data production, which is different from the traditional GT (Fig. 1). In the following methods section, a depth model is introduced to obtain the optimal depth of coverage for pooled sequencing, STD and repeated blocks design are employed to obtain the pooling design. Finally, optimal overlapping pool sequencing strategy could be selected among candidate designs with respect of whole experiment cost according to the cost model.

**Optimal Depth of Coverage for Pooled Sequencing**

The basic procedure of overlapping pool sequencing is to mix DNA samples and run pooled sequencing. Due to the nature of shotgun sequencing strategy, sequencing depth must be sufficient to guarantee effective observations of rare variants for pooled sequencing in case of possible sequencing errors. Higher depth unfortunately increases costs and induces more errors, which could cause false-positive observations. Therefore, the design of sequencing depth is a key factor for conducting pooled sequencing successfully [Sugaya et al., 2012].

Given the hypothesis that reads are completely randomly distributed in genome, sequencing depth is a Poisson variable. Unfortunately, because of bias variation in local content, such as the proportion GC and segment duplication, some regions in the genome are more difficult to sequence for certain MPS technologies [Sampson et al., 2011]. Therefore, the sequencing depth follows an over-dispersed Poisson distribution, which has variance greater than mean. As an alternative, Robinson and Smyth [Robinson and Smyth, 2008] explored the use of negative binomial distribution to model sequence depth. Following the study conducted by C. Miller [Miller et al., 2011], we use a negative binomial model to approximate the over-dispersed Poisson distribution to fit the depth distribution (Equation (4):

$$P(\text{depth} = k) = \text{NB} \left( k; \frac{D}{r-1}, \frac{1}{r} \right) \quad (4)$$

where $D$ denotes the mean depth of coverage and $r$ is the variance mean ratio which can be specified depends on sequencing platforms and genomes.

For a negative pool without variant carriers, taking single base mutation as an example, with mean sequencing depth $D$, the number of reads $N_r$ follows negative binomial distribution $\text{NB}(D/(r - 1), 1/r)$. As one base has unequal chances to be mis-sequenced to one of the other three bases, with the mean sequencing error rate denoted as $p_{\text{error}}$, observations of variants that result from sequencing errors follows a binomial distribution $\text{Bin}(N_r, p_{\text{error}})$ when considering the worst-case scenario. We formulate the probability $F_P$ of the situation that at least $T$ observations of variants are found in a negative diagram for the situation that at least $T$ observations of variants are found in a negative
pool (Equation (5)):

\[ F_p = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left( \sum_{(r-1)}^{D \cap (r)} \right) \sum_{(r-1)}^{D \cap (r)} \sum_{(r-1)}^{D \cap (r)} \text{Bin}(E; N, p_{\text{error}}) \]  

For a positive pool with \( N \) diploid DNA samples and \( c \) heterozygous variant carriers, ignoring the bias in mixing samples, the percent of the variant chromosomes is \( p = c/2N \) where we assume that probability of observing samples homozgyous for the rare variants is extremely low that can be ignored. And for haploid samples, the percent \( p \) equals \( c/N \). Reads containing variants consist of two parts: observations of real variants from variant carriers and fake variants stemming from sequencing errors. The number of reads from variant carriers follows the negative binomial distribution \( \text{NB}(pD/(r-1), 1/r) \), and from normal samples follows the negative binomial distribution \( \text{NB}((1-p)D/(r-1), 1/r) \). We also formulate the probability \( F_N \) of the situation that less than \( T \) reads containing the variants are observed in a positive pool (Equation (6)). More details are given in the Appendix about Equations (5) and (6):

\[ F_N = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left( \sum_{(r-1)}^{D \cap (r)} \right) \sum_{(r-1)}^{D \cap (r)} \sum_{(r-1)}^{D \cap (r)} \text{Bin}(x; i, p_{\text{error}}) \times \left( \sum_{(r-1)}^{D \cap (r)} \right) \sum_{(r-1)}^{D \cap (r)} \sum_{(r-1)}^{D \cap (r)} \text{Bin}(x; j, p_{\text{error}}) \]  

Given the principle that a pool with at least \( T \) observations of variants is classified as positive and less than \( T \) observations as negative, \( F_P \) and \( F_N \) are known as the false-positive rate and false-negative rate of tests in pooling design. For a successful pooling design, test errors must be less than the upper bound that can be corrected. Both \( F_P \) and \( F_N \) should be controlled lower than the test error rate \( \alpha \) that can be tolerated. The optimal sequencing depth \( D \) for a pool is the minimum that satisfy the error rate derived from Equations (5) and (6) that is only related to pool size \( N \) once \( \alpha \) is fixed:

\[ D_{\text{optimal}} = \min[D \mid F_N(D, T) \leq \alpha \& F_P(D, T) \leq \alpha, T \in [1, D]] \]  

where \( F_N(D,T)(F_P(D,T)) \) denotes false-negative (positive) rate in the case that mean depth is \( D \) and critical observations are \( T \). Once mean sequencing depth is fixed as \( D_{\text{optimal}} \), we define the critical observations as the minimum that meets test error rate requirement:

\[ T = \min[T \mid F_N(D_{\text{optimal}}) \leq \alpha \& F_P(D_{\text{optimal}}) \leq \alpha] \]  

\( F_P(D_{\text{optimal}}) \) and \( F_N(D_{\text{optimal}}) \) are calculated under the condition that mean sequencing depth is \( D_{\text{optimal}} \).

Cost-Effective Pooling Design

In 2006, Thierry-Mieg [Thierry-Mieg, 2006] introduced a novel pooling design algorithm called STD, which performs similar or better compared to other previous design algorithms. For STD, three input parameters are: sample size \((n)\), maximum number of positive samples \((d)\), and maximum number of test errors \((E)\). The algorithm guarantees that the design will be able to correctly identify up to \( d \) positive samples in the presence of up to \( E \) observation errors.

These input parameters \((n, d, E)\) are used to compute the design parameters of the construction \( q \) and \( k \). STD is a layered construction with \( k \) layers each with \( q \) pools of which \( q \) is a prime number and \( k \) is smaller than \( q+1 \). Each sample appears only once in each layer. The construction algorithm produces a \( t \times n \) binary matrix \( M \) with \( t = q \times k \) rows that indexed by the pools and \( n \) columns that indexed by the samples.

The process of STD algorithm is shown below:

1. Calculate \( q_{\text{max}} = \lfloor \frac{n}{d} \rfloor \), \( \lfloor \rfloor \) means the integer part.
2. Find a prime number \( q \leq q_{\text{max}} \), start from 2.
3. Calculate compression power \( \Gamma = \min \{ r \mid q^{r+1} \geq n \} \) and let \( k = d^\Gamma + 2E + 1 \).
4. If \( k \leq q + 1 \) go to step 5, else repeat step 2 and 3.
5. Design the matrix described in [Thierry-Mieg, 2006] with \( q, k, \) and \( E \).

For \( n > q > q_{\text{max}}, \Gamma \geq 1 \) and \( t = qk > \lfloor \frac{n}{d} \rfloor \times (d^\Gamma + 2E + 1) \geq n \);

\[ \text{For } q \leq n, \Gamma = 0 \text{ and } t = qk = q(2E+1) \geq n \]

Both mean that when \( q \) is greater than \( q_{\text{max}} \), the number of pools is more than that of samples and the “one sample, one test” strategy is more cost-effective.

The objective of traditional pooling design is to minimize the number of pools and the best parameter \( q \) is:

\[ q = \arg \min \{ q(d^\Gamma + 2E + 1) \mid s.t. q \leq q_{\text{max}} \]  

The cost of overlapping pool sequencing mainly comes from library preparation and data production. According to the depth model above, ignoring the bias in mixing samples, we can modify Equation (2) into an advanced version:

\[ C = tp_l + p_l \sum_{i=1}^{t} D(W_i) \times R + C_p \]  

where the function \( D(W_i) \) is the sequencing depth for the \( i \)th pool in which \( W_i \) is the number of samples in the pool known as the pool size.

For a pooling design in MPS, the best parameter \( q \) is the one that minimizes the sequencing cost:

\[ q = \arg \min \{ C \mid s.t. q \leq q_{\text{max}} \]  

Indeed, the design above is under the guarantee requirement, which means the correction of all errors and identification of all positives must be guaranteed. Recent research showed that STD was capable of correcting errors much larger than those that they guarantee for [Thierry-Mieg and Bailly, 2008]. However, substantial simulations must be performed to figure out how many positive samples and errors a design can actually deal with. To facilitate simulations, Thierry-Mieg et al. provided a freely available tool called interpool, which can be utilized to choose candidate designs by performing simulations for wide ranges of values of STD parameters. Besides, a deterministic algorithm was also integrated in
interpool [Thierry-Mieg and Bailly, 2008]. Therefore, optimal design parameters $q$ and $k$ are those pass the interpool simulation and minimize the sequencing cost:

$$ (q, k)_{\text{optimal}} = \arg \min \limits_{(q, k)} C \ s.t. \ (q, k) \in \{(q, k)\text{pass interpool simulation}\} \tag{12} $$

Apparently, the design optimized for a given maximum number of positive samples can identify all the positive samples accurately only when the positive samples are less than the given maximum number. Otherwise there must exist a design with less pools could be efficient and more cost-effective. Because sequencing errors may cause false observations that lead to that the pooling design must be error-tolerance [Ngo and Du, 2000]. For an error-tolerance pooling design, more pools are required to correct more errors. With lower sequencing depth, both the false-positive rate and false-negative rate get high, which require more pools for error correcting. Therefore, it is a trade-off that can be further investigated and there exists a compromise scheme that minimizes the cost and guarantees the efficiency.

### Repeated Blocks Design

Kainkaryam [Kainkaryam and Woolf, 2008] first proposed repeated blocks design, which divided a large population into small blocks and repeated same pooling design strategy for each block. He found the total number of pools decreased by using STD algorithm when only a finite number of samples can be mixed in each pool.

For a population of $n$ samples dividing into $B$ blocks with each has $n_B$ samples, suppose the frequency of a rare variant is $p_v$, the number of positive samples ($d_B$) in each block follows a binomial distribution $B(n_B, p_v)$ and the probability $p_B$ that the positive samples are no more than $d_B$ is:

$$ p_B = \sum_{i=0}^{d_B} \binom{n_B}{i} p_v^i (1 - p_v)^{n_B-i} \tag{13} $$

When $p_v$ is unknown but there exist at most $d$ positive samples in the population, the number of positive samples in each block follows a hypergeometric distribution $H(n_B, d, n)$ and the probability $p_B$ that the positive samples are no more than $d_B$ is:

$$ p_B = \sum_{i=0}^{d_B} \binom{d}{i} \binom{n-d}{n_B-i} \binom{n}{n_B} \tag{14} $$

Given a probability threshold $p(\alpha)$, the number of positive samples in each block can be taken as $d_B$ when $p_B$ is larger than $p(\alpha)$ and the pooling design for each block can be generated for $(n_B, d_B)$. The cost of repeated blocks design $Cr$ is the sum of the cost for all blocks:

$$ Cr = \sum_{j=1}^{B} t_j + \sum_{j=1}^{B} \sum_{i=1}^{h_j} D(W_{ij}) \times R \tag{15} $$

where $t_j$ denotes the number of pools for the $j$th block and $W_{ij}$ is the size of the $i$th pool for design of the $j$th block.

A pooling design for $(n_B, d_B)$ may only works when positive samples are no more than $d_B$. We call a block failed when positive samples are more than the estimated. Actually, the number of failed blocks $F_B$ should be controlled to ensure the probability that the repeated blocks design succeeds which indicates that $P(F_B = 0)$ must be greater than a threshold $\beta$ to guarantee the correct decoding procedure:

$$ P(F_B = i) \approx \binom{B}{i} (1 - p_B)^i p_B^{B-i} \tag{16} $$

The approximation would be exact if the blocks are independent which is met by Equation (13). According to Equations (15) and (16), we can make an informed best choice about the number of blocks which make repeated blocks design most cost-effective and successfully at the same time:

$$ B = \arg \min \limits_{(B, d_B)} Cr \ s.t. \ (B, d_B) \in \{(B, d_B) \mid P(F_B = 0) \geq \beta\} \tag{17} $$

Repeated blocks design was first proposed in order to reduce the number of pools. In the sequencing experiment, for the reason that pools become smaller for blocks than original population, even with the same sequencing depth per sample, repeated blocks design has the ability to reduce overall sequencing capacity which will further reduce the experiment cost.

### Results

#### Data Increase Due to Pooled Sequencing

In Miller’s research, he found the negative binomial distribution with variance mean ratio $r$ equaling 3 performed much better than Poisson distribution for approximating depth distribution when using Illumina machines [Miller et al., 2011]. Therefore, $r$ was also set to 3 in our experiment.

Taking $p_{\text{error}}$ and $\alpha$ as 0.01, with only one heterozygous variant carried in the positive pool for diploid samples, the optimal sequencing depths of coverage and corresponding critical observations for pooled sequencing of diploid samples were computed based on Equations (7)–(8) and shown in Figure 2. From which, we can infer that the optimal depth of coverage increased rapidly as the enlargement of pool size due to the dilution of chromosomes with variants.

In fact, the optimal depth of coverage was an exponential function of pool size, which means depth per sample also increased with the pool size (supplementary Fig. 1). As described previously, even with the same sequencing depth per sample, repeated blocks design that employs smaller pools is capable of reducing data requirement. On the other hand, depth per sample also drops for smaller pools actually. Therefore, the ability to reduce data requirement of repeated blocks design was underestimated.
Rapid decline in sequencing error rate with the development of sequencing technology will improve the probability of false observation, suggesting the decrease of optimal depth of coverage (Fig. 2C). Therefore, the costs for increased throughput results from pooled sequencing will get lower which broaden the application of overlapping pool sequencing.

Performance of Overlapping Pool Sequencing

With the price of MPS dropped to about $500 per library preparation and $5,300 for 100 Gb (giga-bases) sequencing data production [Sboner et al., 2011], we evaluated the performance of overlapping pool sequencing from the aspect of cost. Cost of pooling procedure accounts for little fraction of whole cost that was neglected in our calculation [Kim et al., 2010]. Due to the limitation of computational capability, pool size was constrained to be smaller than 40 diploids (same in the following analysis).

The simulation tool interpool was utilized to perform series simulations with wide ranges of values for STD parameters $q$ and $k$ for various numbers of variant carriers, sequencing region sizes and sample sizes to obtain the candidate designs. The designs that got 100% accuracy under 1% false-negative errors and 1% false-positive errors were chosen as candidate designs. As described previously, the candidate designs that optimized for a given number of variant carriers can identify all the carriers accurately only when the carriers are less than the given value (e.g. supplementary Fig. 2). Optimal depths for pooled sequencing of diploid samples were computed based on Equation (7) in which $P_{\text{error}}$ and $\alpha$ were set to 0.01. Finally, most cost-effective design was chosen as optimal for further analysis based on Equation (10). The experiment was also implemented by utilizing DNA Sudoku, which was proposed by Erlich et al. [Erlich et al., 2009a].

Comparison of the costs of these two overlapping pool sequencing designs to sequencing separately strategy was shown in Figure 3A. Results indicated that overlapping pool sequencing employing STD performed better than DNA Sudoku in terms of cost. Besides, STD has the advantage of tolerant more test errors. But costs of both increased rapidly as the number of variant carriers. Therefore, we calculated the highest frequencies of rare variant carriers at which using overlapping pool sequencing for screening is more cost-effective subsequently (Fig. 3B, supplementary Fig. 3). From which we can conclude that compared with DNA Sudoku, STD can be applied to screen for higher frequency variants at the same cost especially when the sequencing region was limited to be very small. Due to the huge depth for pooled sequencing, cost for overlapping pool sequencing increased rapidly as the region size, resulting into the decline of the highest frequencies of rare variant carriers. Results also indicated that overlapping pool sequencing can handle higher frequency variants carriers for smaller sample sizes, which also verified the significance of repeated blocks design.

From Figure 3, we can infer that overlapping pool sequencing is only suitable for very rare variants and short target region. For a population size of 200 diploid individuals, overlapping pool sequencing is cost-effective only when the variant carriers frequency is lower than 1.5% at which the target region is set to 30 Mb. This situation appears not only because of the limited efficiency of the pooling design algorithms but also due to the high depth for pooled sequencing. Especially for large sequencing regions, dramatically increased sequencing data requirement will offset the cost reduction in library construction. However, major rare variants in human population have minor allele frequency (MAF) lower than 0.5%, which means rare variants are observed in only few samples in a population size of hundreds [Nelson et al., 2012; Tennessen et al., 2012]. In these cases, the overlapping pool sequencing strategy will be very valuable and feasible. Combined with target sequence capture technology that shrinks the sequencing region significantly, the method will be more cost-effective [Chou et al., 2010].
Figure 3. (A) The cost of overlapping pool sequencing for identifying variant carriers in 200 diploid samples where sequencing region is fixed as 30 Mb (approximate to the length of human exome). (B) The highest percent of variant carriers at which using overlapping pool sequencing for screening is more cost-effective than sequencing each sample separately where sequencing region ranges from 0.1 to 50 Mb. We set 0.1–10 Mb as the range of size for candidate gene sequencing, and 30–50 Mb as the range of size for human exome sequencing.

Table 1. Candidate pooling designs for \((200, 2)\). \(q\) and \(k\) are two parameters of STD. Sequencing region is fixed as 30 Mb in the calculation of cost. Results for \(q\) greater than 29 are not shown. \(P_l = $500\) and \(P_d = $5300/100\) Gb for Equation (10)

| \(q\) | \(k\) | Max pool size | Number of pools | Data requirement (Gb) | Cost ($) |
|------|------|--------------|----------------|----------------------|----------|
| 7    | 5    | 29           | 35             | 2,070                | 127,202  |
| 11   | 5    | 19           | 55             | 1,687                | 116,890  |
| 13   | 5    | 16           | 65             | 1,589                | 116,698  |
| 17   | 3    | 12           | 51             | 862                  | 71,211   |
| 19   | 3    | 11           | 57             | 828                  | 72,579   |
| 23   | 3    | 9            | 69             | 771                  | 75,350   |
| 29   | 3    | 7            | 87             | 706                  | 80,935   |

Because cost of library preparation increases when incorporates target capture and cost of data production falls fast, experimental design should be optimized based on the cost model. With higher ratio of cost for library preparation/data production, designs that require fewer pools will be optimal and vice versa (supplementary Table S1).

Further Cost Reduction by Repeated Blocks Design

We first used the process described above to choose the most cost-effective pooling design that can tolerate 1% false-negative and 1% false-positive errors for 200 diploid samples with two variant carriers at which the length of sequencing region was set to 30 Mb. Results showed that the most cost-effective design necessitated 51 pools and 862 Gb data where optimal depth for each pool was computed base on Equation (7) in which \(p_{\text{error}}\) and \(\alpha\) were 0.01 (Table 1).

Utilizing repeated block design, suppose the threshold probability of success \(\beta\) was 0.7; the number of samples and positive samples in each block are shown in Table 2 when the population is divided into \(1\sim10\) blocks. Repeated same most cost-effective strategy with the ability to correct 1% false-negative and 1% false-positive errors for each block, the total number of pools and sequencing data requirement for repeated blocks design were shown in Table 2. Taking advantage of repeated blocks design, only half of the data are required compared with design for initial population, which further reduced the cost to 80%, illustrating that repeated blocks design was worth being used in practice.

Given a population, even when the variant carriers exceeds the highest value at which using overlapping pool sequencing for screening is cost-effective, repeated blocks design could still be efficient. Take exome sequencing project of Tennesse
et al., example, they found 86% variants had a MAF less than 0.5% in a population with 2,440 individuals, namely at most 12 carriers in the population [Tennessee et al., 2012]. For the population (2,440, 12), overlapping pool sequencing cannot be cost-effective anymore, but utilizing repeated blocks design could reduce the cost to only 67% (supplementary Table S2 and Fig. 5). As long as the variant carriers are not more than 12 in the population, they can be identified correctly.

Compared with the design for original population, designs for smaller blocks have smaller pools, which translate into much fewer data requirements. Furthermore, due to the vastly reduction in data requirement, repeated blocks design also facilitates short reads alignment, data storage, downstream analyses which further reduces the cost.

### Simulation Experiment

We simulated an experiment to apply overlapping pool sequencing to detect single base mutations among 200 simulated individuals by using human exome sequencing data from Gracia-Aznarez’s research [Gracia-Aznarez et al., 2013]. First, exome sequencing data of seven individuals from seven different families were downloaded. Based on which we simulated 200 samples where samples stemmed from the seventh family were taken as negative control (Table 4). Next, Bowtie 0.12.7 [Langmead et al., 2009] was used to map reads back to human exon sequence and single base mutations were called by using SAMtools0.1.18 [Li et al., 2009] based on the mapping results. To control the quality, mutations with coverage lower than 15 were filtered. Subsequently, we selected the mutations that happened only once in these seven families, among which mutations that belonged to the seventh family was deleted as the seventh family was taken as negative control. Finally, 1,375 mutations were left which can be considered as rare variants because each happened exactly in two individuals among these 200 simulated individuals (Table 4).

First, we applied sequencing separately strategy to detect mutations, which meant sequencing each sample independently. Genome sequencing data were simulated by taking reads randomly from the dataset. Optimal depth for single sample sequencing was computed from Equation (7) in which $N$ was 1, $p_{\text{error}}$ and $\alpha$ were 0.01, and critical observations $T$ was computed from Equation (8) subsequently. The optimal depth is 27 and critical observations are 3 according to the results. Consequently, this strategy necessitates 200 libraries, 162 Gb sequencing data (region was set to 30 Mb) and the cost was about $108,586. Next, Bowtie was used to map reads and Perl scripts to count the reads mapped at mutations’ loci. At last, all the 1,375 mutations were assigned to the samples correctly.

Next, we simulated overlapping pool sequencing. Because each mutation belongs to at most two samples, the parameter maximum positive samples $d$ was set to 2 for STD. From Table 1, we found the cost got lowest when $q$ equaled 17 which necessitated 51 pools and 862 Gb data and the whole cost was $71,211, which was only 66% of the cost for sequencing separately. Pooled sequencing was conducted by taking reads randomly from the dataset and mixing in silico. Considering up to 5% average noise on the DNA quantities of each sample in pooling procedure, the number of reads for each sample was revised with a random coefficient following normal distribution $N(0, 0.05^2)$ to make pooled sequencing closer to reality [Shental et al., 2010]. With the same procedure to discover mutations, a pool was classified positive if the observations of mutations are more than the critical observations. Next, we used $\text{interpool}$ to carry out error tolerant decoding procedure. Results showed that for all but one mutations were assigned to the samples correctly.

Last, we divided 200 samples into blocks randomly. Took $\beta$ as 0.7, from previous results (Table 2), we found the cost got lowest when divided into three blocks. Then 200 samples were randomly divided into three blocks with 66, 67, 67 samples, respectively. Same procedure was applied to choose best parameters to design the pools for (66, 1) and (67, 1) that guaranteed the ability to correctly identify up to one positive samples in the presence of up to 1% false-negative errors and 1% false-positive errors (Table 3). Finally, 66 pools and 453 Gb data were needed theoretically and the cost was only 52% of the cost for sequencing separately. With the similar procedure, the results showed that all mutations were decoded correctly.

Results of the simulation experiment suggested that overlapping pool sequencing can reduce costs significantly and simultaneously keep the correct decoding rate (Table 5).

---

Table 3. Candidate pooling designs for (66, 1) and (67, 1). Target sequencing region is set to 30 Mb in the calculation of cost. Results for $q$ greater than 17 are not shown

| Population | $q$ | $k$ | Number of pools | Data requirement (Gb) | Cost($) |
|------------|-----|-----|-----------------|-----------------------|--------|
| (66, 1)    | 5   | 3   | 15              | 297                   | 23,260 |
|            | 7   | 3   | 21              | 262                   | 24,395 |
|            | 11  | 2   | 22              | 148                   | 18,836 |
|            | 13  | 2   | 26              | 139                   | 20,390 |
|            | 17  | 2   | 34              | 142                   | 24,505 |
| (67, 1)    | 5   | 3   | 15              | 304                   | 23,589 |
|            | 7   | 3   | 21              | 268                   | 24,686 |
|            | 11  | 2   | 22              | 151                   | 19,001 |
|            | 13  | 2   | 26              | 142                   | 20,546 |
|            | 17  | 2   | 34              | 144                   | 24,645 |

Table 4. DNA samples for 200 individuals in the simulation experiment

| Family ID | Family   | Simulated samples | Number of mutations | SRA ID |
|-----------|----------|-------------------|---------------------|--------|
| 1         | 073240   | 62th, 131th       | 105                 | ERR166303 |
| 2         | DAD1_1   | 130th, 132th      | 640                 | ERR166305 |
| 3         | F2887_24 | 6th, 183th        | 68                  | ERR166308 |
| 4         | L1408    | 124th, 143th      | 142                 | ERR166315 |
| 5         | NA12813  | 10th, 71th        | 35                  | ERR166327 |
| 6         | RUL036_7 | 157th, 138th      | 385                 | ERR166332 |
| 7         | RUL153_3 | Others            | N/A*               | ERR166337 |

* The accession number of SRA database (http://www.ncbi.nlm.nih.gov/sra).
  * 188 simulated samples are from RUL153_3.
  * These individuals are taken as negative control.

---

Genetic Epidemiology, Vol. 37, No. 8, 820–830, 2013
blocks design could still be efficient. However, selecting the carriers exceeds the highest value at which using overlap-
as reads alignment, data storage. Especially, when the variant and reduces the costs that stem from data management such
plifies pooling procedure, further decreases data requirement
specific genomic regions or genes, e.g., exons.

pool sequencing necessitates fewer data requirements and
which shrinks sequencing region significantly, overlapping
error rates translating into less cost for data production, the
probability of each kind of sequencing errors. Furthermore,
is overestimated which can be improved by considering the
depth model. Therefore, the optimal sequencing depth

depths for pooled sequencing hinders the application of over-
 pooling sequencing. However, due to the inequality in the
 chances for a base being mis-sequenced to one of the
 other three bases, the worst-case scenario is considered in
 the depth model. Therefore, the optimal sequencing depth is
 overestimated which can be improved by considering the probability of each kind of sequencing errors. Furthermore,
with the development of the third-generation sequencing technology [Clarke et al., 2009; Eid et al., 2009] which has
potential longer read lengths, shorter sequencing time, lower error rates translating into less cost for data production, the
application of overlapping pool sequencing will broaden.

As for practical aspects, current overlapping pool-
sequencing strategy suits better for short sequencing region
and very few variant carriers. For a population with 200
diploid individuals, overlapping pool sequencing is cost-
effective only when the variant carriers frequency is lower
than 1.5% at which the target region is set to 30 Mb. High
depths for pooled sequencing hinders the application of over-
lapping pool sequencing. However, due to the inequality in the
 chances for a base being mis-sequenced to one of the
 other three bases, the worst-case scenario is considered in
the depth model. Therefore, the optimal sequencing depth is
overestimated which can be improved by considering the
probability of each kind of sequencing errors. Furthermore,
with the development of the third-generation sequencing
technology [Clarke et al., 2009; Eid et al., 2009] which has
potential longer read lengths, shorter sequencing time, lower
error rates translating into less cost for data production, the
application of overlapping pool sequencing will broaden.

Besides, taking advantage of sequence capture technology
which shrinks sequencing region significantly, overlapping
pool sequencing necessitates fewer data requirements and
will be more worthwhile for those studies that focused on
specific genomic regions or genes, e.g., exons.

Utilizing repeated blocks design, pools shrink, which sim-
plifies pooling procedure, further decreases data requirement
and reduces the costs that stem from data management such
as reads alignment, data storage. Especially, when the variant
carriers exceeds the highest value at which using overlap-
pool sequencing for screening is cost-effective, repeated
blocks design could still be efficient. However, selecting the

Table 5. Comparison of three strategies for screening rare variant carriers

| Strategy                  | Number of libraries | Data requirement (Gb) | Cost ($)     | Accuracy (%) |
|---------------------------|---------------------|-----------------------|--------------|--------------|
| Sequencing separately     | 200                 | 162                   | 108,586      | 100          |
| Overlapping pool sequencing| 51                  | 862                   | 71,211       | 99.93        |
| Repeated blocks design    | 66                  | 450                   | 57,003       | 100          |

Discussion

In this paper, we proposed overlapping pool sequencing to
screen rare variants with a corresponding cost model and a
depth model for conducting pooled sequencing successfully.
According to the depth model, optimal depths of coverage
were computed for various pool sizes. Optimal design para-

eters can be chosen to minimize the cost and guarantee
correct decoding procedure based on these two models. Un-
der the condition that optimal depth increased exponentially
as the pool size, we verified that it was more efficient to di-
vide a large population into smaller blocks, which requires
much less data. In the repeated blocks design, the number of
blocks could also be optimized to make the experiment more
cost-effective. At last, the simulation experiment to detect
single base mutations in human exome showed the poten-
tiality of overlapping pool sequencing to screen rare variants
and proved the efficiency of repeated blocks design.

As for practical aspects, current overlapping pool-
sequencing strategy suits better for short sequencing region
and very few variant carriers. For a population with 200
diploid individuals, overlapping pool sequencing is cost-
effective only when the variant carriers frequency is lower
than 1.5% at which the target region is set to 30 Mb. High
depths for pooled sequencing hinders the application of over-
lapping pool sequencing. However, due to the inequality in the
 chances for a base being mis-sequenced to one of the
 other three bases, the worst-case scenario is considered in
the depth model. Therefore, the optimal sequencing depth is
overestimated which can be improved by considering the
probability of each kind of sequencing errors. Furthermore,
with the development of the third-generation sequencing
technology [Clarke et al., 2009; Eid et al., 2009] which has
potential longer read lengths, shorter sequencing time, lower
error rates translating into less cost for data production, the
application of overlapping pool sequencing will broaden.

Besides, taking advantage of sequence capture technology
which shrinks sequencing region significantly, overlapping
pool sequencing necessitates fewer data requirements and
will be more worthwhile for those studies that focused on
specific genomic regions or genes, e.g., exons.

Utilizing repeated blocks design, pools shrink, which sim-
plifies pooling procedure, further decreases data requirement
and reduces the costs that stem from data management such
as reads alignment, data storage. Especially, when the variant
carriers exceeds the highest value at which using overlap-
pool sequencing for screening is cost-effective, repeated
blocks design could still be efficient. However, selecting the

number of blocks only depends on the cost is not always
suitable. Comprehensive decision can be made by taking all
aspects considered including data requirement, the number of
pools and costs.

One possible difficulty in our approach is that the pool-
ing procedure may be imperfect. Difficulty to obtain equal
amounts of DNA from each individual results in noise in the
design matrix and hinders accurate reconstruction. Another
drawback of our methods stems from the fact that the quan-
titative information in reads, which could reflect the percent
of variant chromosomes in a pool is missed. In our future
work, utilizing the quantitative information, the genotype
(heterozygote and homozygote) of each variant carrier can be
determined accurately. Subsequently, the usefulness will ulti-
ately depend on how the technologies develop. If barcoding
becomes very simple and inexpensive, the way forward will
be to barcode every sample. If, by contrast, sequencing costs
fall rapidly compared with those of barcoding, overlapping
pool sequencing should prove increasingly valuable [Patter-
son and Gabriel, 2009]. Besides, our strategy applies GT in the
most straightforward fashion, with higher efficiency of new
pooling design algorithms or other mathematical methods
such as compressed sensing [Candès et al., 2006; Donoho,
2006], the performance may be further improved and the
potential difficulties could be solved.

Investigating the role of rare variants for complex trait
mapping has led to studies that aggregate rare variants, and
determine the abundance, distribution of rare variants in
populations [Nelson et al., 2012]. At present, rare variants

genotyping has already been applied in screens for rare ge-
netic disorders, e.g., the Israeli ministry of health sponsored
carrier-screening tests for a list of 36 severe genetic diseases in
35 different localities/communities [Zlotogora et al., 2008].
Any rare genetic variants can be detected by our method as
long as it can be discovered by MPS. Error tolerant pool-
ing design based overlapping pool sequencing has the ability
to correct false observations, which can’t be achieved by se-
quencing separately. With the advantage of identifying vari-
ant carriers, overlapping pool sequencing is worthwhile be-
ing used in screening for rare variants in genomic regions or
candidate genes known to associate with diseases which will
empower associating rare variants with complex traits, inves-
tigating target biology and drug response, providing clinically
actionable information and so on. [Koppers et al., 2013; Nel-
son et al., 2012; Newman et al., 2013; Tennessen et al., 2012]

Acknowledgments

This work was supported by the National Basic Research Program of
China (No. 2012CB316501) and the National Natural Science Foundation
of China (61073141).

References

Balding D, Bruno W, Knill E, Torney D. 1996. A Comparative Survey of Non-Adaptive
Pooling Designs, Genetic Mapping and DNA Sequencing. New York: Springer, p.
133–154.

Balding DJ, Torney DC. 1997. The design of pooling experiments for screening a clone
map. Fungal Genet Biol 21:302–307.
Bruno WI, Knell E, Bolding DJ, Bruce D, Doggett N, Sawhill W, Stallings R, Whitaker CC, Tidwell DC. 1995. Efficient pooling designs for library screening. Genomics 26(1):211–20.

Candès EJ, Romberg J, Tao T. 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Inf Theory IEEE Trans* 52:489–509.

Chen HB, Hwang FK. 2008. A survey on nonadaptive group testing algorithms through the angle of decoding. *J Comb Optim* 15:49–59.

Chen L-S, Liu C-S, Noe A, Zhang X, Mao R. 2010. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. *Clin Chem* 56:62–70.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 4:265–70.

Ding-Zhu D, Hwang FK. 2000. Combinatorial group testing and its applications. *SIAM Appl Math* 12:339.

Donoho DL. 2006. Compressed sensing. *Inf Theory IEEE Trans* 52:1289–1306.

Dorfman R. 1943. The detection of defective members of large populations. *Ann Math Stat* 44:460–464.

Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Deninger SW, Hamvas A, Cole FS and others. 2009. Quantification of rare alleles from pooled genomic DNA. *Nat Methods* 6:263–265.

Duer H, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhit AH, Abraham C, Regueiro M, Griffiths A and others. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Sci Signal* 314:1461.

Dyachkov A, Rykov V. 1983. A survey of superimposed code theory. *Probl Control Inform Theory* 12:1–13.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B and others. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

Erlrich Y, Chang K, Gordon A, Ronen R, Navon O, Roos M, Hannon GJ. 2009a. DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res* 19:1243–1253.

Erlrich Y, Shental N, Amir A, Zuk O. 2009b. Compressed sensing approach for high throughput carrier screen. 47th Annual Allerton Conference on Communication, Control, and Computing: IEEE 539–544.

Gracia-Aznarez FJ, Fernandez V, Pita G, Peterlongo P, Dominguez O, de la Hoya M, Duran M, Osorio A, Moreno L, Gonzalez-Neira A and others. 2013. Whole exome sequencing suggests much of non-BRCAl/BRCAl2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PLoS One* 8:e55681.

Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J and others. 2006. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* 39:207–211.

He D, Zaitlen N, Pasaniuc B, Eskin E, Halperin E. 2011. Genotyping common and rare variants in the coding region of ALS-associated genes at 9p21.2 and 19p13.3. *Neurobiol Aging* 34:5) 1518.e5–1518.e7

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–753.

Miller CA, Hampton O, Coarfa C, Molinossi V. 2011. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* 6:e16327.

Nelson MR, Wegmann D, Eth MG, Kessner D, Jean PS, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D and others. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337:100–104.

Newman JR, Leng AH, Boyle RS, Silbun PA, Melllick GD. 2013. Screening for rare variant association in the THAP1 gene in a primary dystonia cohort. *Mov Disord* doi: 10.1002/mds.25471.

Ngo HQ, Du DZ. 2000. A survey on combinatorial group testing algorithms with applications to DNA library screening. *Discrete Math Probl Med Appl* 55:171–182.

Nielsen R. 2010. Genomics: in search of rare human variants. *Nature* 467:1050–1051.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Horii M, Nakamura Y and others. 2002. Functional SNPs in the lymphotaxin-α gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 32:650–654.

Patterson N, Gabriel S. 2009. Combinatorics and next-generation sequencing. *Nat Biotechnol* 27:826–827.

Prabhur SS, Pe’er I. 2009. Overlapping pools for high-throughput targeted resequencing. *Genome Res* 19:1254–1261.

Robinson MD, Smyth GK. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Bioinformatics* 24:321–322.

Sampson J, Jacobs K, Yeager M, Channock S, Chatterjee N. 2011. Efficient study design for next generation sequencing. *Genet Epidemiol* 35:269–277.

Shoner A, Mu XJ, Greenbaum DB, Auerbach RK, Gerstein MB. 2011. The real cost of sequencing: higher than you think! *Genome Biol* 12:125.

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

Shtantal N, Amir A, Zuk O. 2010. Identification of rare alleles and their carriers using compressed sequence mapping. *Nucleic Acids Res* 38:e179–e179.

Sobel M, Groff PA. 1959. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell Syst Tech J* 38:1178–1252.

Sugaya Y, Sugaya Y, Akazawa Y, Saito A, Kamitsui S. 2012. NDesign: software for study design for the detection of rare variants from next-generation sequencing data. *J Hum Genet* 57:676–678.

Tennessen JA, Bigham AW, O’Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G and others. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69.

Thierry-Mieg N. 2006. A new pooling strategy for high-throughput screening: the Shifted Transversal Design. *BMetrica Bioclinics* 7:28.

Thierry-Mieg N, Bailly G. 2008. Interpool: interpreting smart-pooling results. *Bioinformat* 14:696–703.

Vermeiren V, Deplancke B, Barракsa MI, Reece-Hoyes JS, Arda HE, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Brent MR and others. 2007. Matrix and Steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat Methods* 4:59–664.

Westreich DJ, HUDGENS MG, Fiscus SA, Pichler CD. 2008. Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. *J Clin Microbiol* 46:1785–1792.

Wilson-Lindsay L, Davis PW, Ecker DJ, Hébert N, Acevedo O, Spradkle K, Brennan T, Schwarcz L, Freier SM, Wyatt JR. 1996. Deconvolution of combinatorial libraries for drug discovery: experimental comparison of pooling strategies. *J Biol Chem* 39:2720–2726.

Xin X, Raal JL, Hirozane-Kishikawa T, Hill DE, Vital M, Boone C, Thierry-Mieg N. 2009. Shifted Transversal designs for high coverage smart-pooling for high coverage interactome mapping. *Genome Res* 19:1262–1269.

Zlotogora J, Carmi R, Lev B, Shaley SA. 2008. A targeted population carrier screening program for severe and frequent genetic diseases in Israel. *Eur J Hum Genet* 15:791–597.

Appendix: Proof of $F_P$ and $F_N$

$F_P$ is the probability that at least $T$ observations of variants are found in a negative pool where all these observations stem from sequencing errors. Denote $N_r$ as the number of reads and $E$ as the sequencing errors. Simply, $F_P$ can be written as.
Equation (A1)

\[ F_P = \sum_{N_r = 1}^{\infty} P(N_r) \left( \sum_{E = T}^{N_r} P(E | N_r) \right) \quad (A1) \]

where \( P(N_r) \) means the probability that \( N_r \) reads are obtained and \( P(E | N_r) \) means the probability that \( E \) errors happen among there \( N_r \) reads. As the depth follows negative binomial distribution and sequencing errors follow binomial distribution, we can formulate these two probabilities as Equations (A2) and (A3). In Equation (A2), \( D \) and \( r \) are mean depth of coverage for pooled sequencing and variance mean ratio, respectively. In Equation (A3), \( p_{\text{error}} \) is the mean sequencing error rate.

\[ P(N_r) = NB\left(N_r; \frac{D}{r-1}, \frac{1}{r}\right) \quad (A2) \]

\[ P(E | N_r) = \text{Bin}(E; N_r, p_{\text{error}}) \quad (A3) \]

After integrating Equations (A1)–(A3), we can formulate \( F_P \) as Equation (A4)

\[ F_P = \sum_{N_r = 1}^{\infty} NB\left(N_r; \frac{D}{r-1}, \frac{1}{r}\right) \left( \sum_{E = T}^{N_r} \text{Bin}(E; N_r, p_{\text{error}}) \right) \quad (A4) \]

\( F_P \) is the probability that less than \( T \) observations of variants are found in a positive pool. The observations of a variant in a positive pool consist of two parts: real variants from variant chromosomes and fake variants resulting from sequencing errors. Briefly, we can denote \( F_N \) as Equation (A5) where \( O \) denotes the number of observations of a variant. Further, \( P(O) \) can be written as Equation (A6) where \( P_N(x) \) stands for the probability that \( x \) reads containing variants stem from the sequencing results of normal chromosomes and \( P_P(O - x) \) denotes the probability that \( O - x \) reads containing variants from variant chromosomes.

\[ F_N = \sum_{O = 0}^{T-1} P(O) \quad (A5) \]

\[ P(O) = \sum_{x = 0}^{O} PN(x)P_P(O - x) \quad (A6) \]

Similar to the procedure to obtain Equation (A4), we can easily formulate \( P_N(x) \) and \( P_P(O - x) \) as Equations (A7) and (A8). The only difference is the mean sequencing depth of coverage. The percent of variant chromosomes and normal chromosomes are \( p \) and \( 1 - p \), respectively. Therefore, the mean depth of coverage for sequencing variant chromosomes and normal chromosomes are \( p \) and \( 1 - p \), respectively.

\[ P_n(x) = \sum_{i = x}^{\infty} NB\left(i; \frac{(1-p)D}{r-1}, \frac{1}{r}\right) \text{Bin}(x; i, p_{\text{error}}) \quad (A7) \]

\[ P_P(O - x) = \sum_{j = O-x}^{\infty} NB\left(j; \frac{pD}{r-1}, \frac{1}{r}\right) \text{Bin}(j - O + x; j, p_{\text{error}}) \quad (A8) \]

In the same way, we can obtain \( F_N \) by integrating Equations (A5)–(A8), shown as Equation (A9).

\[ F_N = \sum_{O = 0}^{T-1} \sum_{x = 0}^{O} \left\{ \sum_{i = x}^{\infty} NB\left(i; \frac{(1-p)D}{r-1}, \frac{1}{r}\right) \text{Bin}(x; i, p_{\text{error}}) \right\} \times \left\{ \sum_{j = O-x}^{\infty} NB\left(j; \frac{pD}{r-1}, \frac{1}{r}\right) \text{Bin}(j - O + x; j, p_{\text{error}}) \right\} \quad (A9) \]