Efficient identification of rare variants in large populations: deep re-sequencing the CRP locus in the CARDIA study

Christina T. L. Chen1,*, Andrew N. McDavid2, Orsalem J. Kahsai1, Ahmad S. Zebari1 and Christopher S. Carlson1

1Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA and 2Division of Vaccine and Infectious Disease, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Received July 18, 2012; Revised January 9, 2013; Accepted January 25, 2013

ABSTRACT

Effect sizes of many common single nucleotide polymorphisms identified in genome-wide association studies generally explain only a modest fraction of the total estimated heritability in a variety of traits. One hypothesis is that rare variants with larger effects might account for the missing heritability. Despite advances in sequencing technology, discovering rare variants in a large population is still economically challenging. Sequencing pooled samples can reduce the cost, but detecting rare variants and identifying individual carriers is difficult and requires additional experiments. To address these issues, we have developed a rare variant-detection algorithm V-Sieve to screen for rare alleles in pooled DNA samples which, in combination with a unique pooling strategy, is able to efficiently screen a candidate gene for idiosyncratic variants in thousands of samples. We applied this method to 2283 individuals, and identified >100 polymorphisms in the C-reactive protein (CRP) locus at an allele frequency as low as 0.02%, with a positive predictive rate of 93%. We believe this algorithm will be useful in both screening for rare variants in genomic regions known to associate with particular phenotypes and in replicating rare variant associations identified in large-scale studies, such as exome re-sequencing projects.

INTRODUCTION

Genome-wide association studies (GWAS) have been successful in identifying single nucleotide polymorphisms (SNPs) associated with a spectrum of phenotypes ranging from quantitative traits, such as adult height (1), bone mineral density (2), age of menopause and age of menarche in women (3–5), to disease states including a variety of cancers (6–9). These polymorphisms are typically common, with minor allele frequencies >5%. Even though dozens of SNPs were found to associate with each of these phenotypes, each common variant confers a small increment of risk, and in combination, the common variants explain a relatively small proportion of heritable variation—the portion of phenotypic variance in a population attributable to additive genetic factors. For discrete traits reported to date, the median odds ratio per copy of the risk allele is 1.33, with only three SNPs showing odds ratios >3.0 (10). One hypothesis for the large unexplained residual phenotypic variance is that in addition to the common polymorphisms identified by GWAS, rare variants with <5% frequency may also play a role in determining phenotypes. Advances in sequencing technology have made it possible to sequence the entire genome in an individual, but sequencing large populations remains impractical owing to the cost per sample. Rather than sequence individuals one at a time, an alternative approach that can reduce costs is to sequence pooled samples containing multiple individuals. There are two major challenges in detecting rare variants in pooled samples: first, distinguishing experimental noise from true polymorphisms with low frequencies in the pool is difficult, and second, determining the identity of carriers of these rare variants within the pooled samples without additional experiments is not trivial.

To address the obstacles in identifying rare variants in a pool, we develop an algorithm V-Sieve and demonstrate this approach by re-sequencing the C-reactive protein (CRP) locus in individuals in the Coronary Artery Risk Development in Young Adults (CARDIA) cohort (11). V-Sieve leverages the distribution of variant allele frequency across pools to find frequency spectra consistent with rare variants and inconsistent with background noise. We condense this into a single statistic and use the
predicted null distribution to choose a threshold for declaring the presence of a putative polymorphism in a pool. By applying V-sieve in serial over several different pooling schemes, we can identify individual carriers of rare variants in our samples. We verify these putative carriers through additional genotyping in all samples. In this study, we demonstrate that our approach is both sensitive and efficient in identifying rare variants in a pooled sample, as well as in determining the identity of individual carriers for low frequency alleles.

**MATERIALS AND METHODS**

**Sample descriptions and preparations for sequencing experiments**

The CARDIA cohort is a population-based study initiated in 1985 to investigate the evolution of cardiovascular risk factors in a large biracial cohort of young adults (11,12). We used DNA from 2283 individuals (24 plates of 96 samples, with 21 wells containing no DNA) from CARDIA in these experiments. Samples on these 96-well plates were combined into six 384-well plates. For each individual sample, we performed long range polymerase chain reaction (PCR) to amplify a 6 kb genomic region surrounding the *CRP* locus (NCBI36/hg18 chr1: 157 945 884–157 951 857; forward primer: ATGTCTGTG ATCAGGACACATTT; reverse primer: GGCATGTC CCTGAGATAAGAAATC) (Invitrogen). PCR products from 96 selected wells were combined into one pool, nebulized to fragment the pooled PCR products into an average length of 500 bp and labelled with one barcode (Paired-End DNA Sample Prep kit, Multiplexing Sample Preparation Oligonucleotide kit, Multiplexing Sequencing Primers, Illumina). For the first sequencing experiment, each of the six 384-well plates was divided into four separate pools based on the position of the wells (Supplementary Figure S1). This pooling scheme resulted in 24 pools overall, which were sequenced using two lanes on the Illumina Genome Analyzer IIx platform. In the second experiment, we used a combinatorial approach by designing three distinct pooling schemes used to pool samples. This approach aimed to assign each individual to three separate pools, one in each pooling scheme, and thus each individual was present in three separate pools and sequenced in three separate lanes (Figure 1). This yielded 72 libraries that were sequenced in six lanes.

**Read quality assessment**

The first sequencing experiment generated 53 bp forward and 53 bp reverse reads. Base quality values for the reads were in Illumina FASTQ version 1.5+. Raw reads were processed using the MAQ program (13). There were on average 5.7 million reads per pool, ranging from 2 to 8.2 million reads (Supplementary Figure S2). Paired-end reads were mapped to the reference *CRP* sequence with the following MAQ parameters: 800 bp maximum allowed distance between two paired reads (−a800), maximum 2 mismatches in the first 24 bp (−a200), maximum allowed sum of qualities of mismatches (−e200). We kept only pairs with one forward and one reverse reads in the analysis. Mapped paired reads with other configuration (forward–forward, reverse–reverse, reverse–forward) and orphan reads were discarded. On average, 95% of forward–reverse paired-end reads in each pool were successfully mapped back to the *CRP* reference region (range: 89–97%). The percentage of reads with at least one mismatch averaged ~17% per pool (range: 16–19%), with an average of two mismatches per read per pool (range: 1.8–2.3). The average Phred base quality score per mismatch was ~14.3 across pools, indicating that most mismatches were likely sequencing errors. We excluded reads with at least 14 mismatches or a sum of mismatches quality scores of >200 from analyses, removing an average of 0.18% of reads from each pool. Coverage per position in each pool ranged from 334X–1328X, with an average base coverage of 900X. Using bases with Phred quality score of >35, coverage along the *CRP* region is shown in Supplementary Figure S3.

The second sequencing experiment initially generated 31 bp single-end reads. The reads were mapped to the reference sequence with the following MAQ parameters: number of mismatches in the first 24 bp set to 2 (−n2) and maximum allowed sum of qualities of mismatches set to 200 (−n200). We filtered out reads with >5 mismatches or with a sum of qualities of mismatches >150. A second run of the same libraries generated 31 bp forward and 31 bp reverse paired-end reads, and we used the same mapping parameters as the first sequencing experiment to map back to the reference sequence. We obtained an average of 2.4 million reads per pool, ranging from 0.25 to 4.6 million (Supplementary Figure S4). Percentage of reads with at least one mismatch averaged ~12.6% per pool (range: 8.4–28%). Average number of mismatches per read is 1.3 mismatches across all libraries. The average quality score per mismatch ranged from 11.4 to 24.5, with an average of 20.9 across all pools. We excluded reads with at least six mismatches or a sum of mismatches quality scores of >300 from later analyses, removing an average of 0.03% of reads from each pool. The filtered reads from both runs were combined for variant discovery, resulting in an average coverage per base position of 705X in each pool. Using bases with quality score of >Q30, coverage along the *CRP* region is shown in Supplementary Figure S5.

**Variant calling**

Positions along the *CRP* reference region were screened for rare variants independently. To determine whether a position was polymorphic in the first sequencing experiment, we used only reads in eligible pools. A pool was eligible if it contained at least 5000 reads covering that position with a base quality score of at least Q35. Using only bases with high quality lowered the probability of identifying a sequencing error incorrectly as a polymorphic variant. Using pools with a base coverage of at least 5000X at that position ensures that, even if only 1 of 192 chromosomes in the pool is polymorphic, we would
expect to observe the rare allele >26 times. An eligible pool in the second sequencing experiment was defined as having >3000 reads with a base quality score of at least Q30 for a given position. A lower threshold for read depth was chosen because the read length for the second experiment was 31 bp, instead of 53 bp in the first experiment. Even at this read depth, we would expect to observe the rare allele >15 times in a pool if present.

For each base, we considered the major allele to be the most commonly observed allele with highest frequencies in each pool across all eligible pools. The minor allele was called in two ways; first, the allele with the second highest average frequency across all eligible pools and second, the allele which was the second most common allele across majority of the available pools. In most circumstances, these two approaches resulted in the same minor allele; however, they gave different results when the variant was potentially tri-allelic (Supplementary Figure S6). We imposed a stringent criterion requiring the minor allele to have a frequency of at least 0.5%, which equals a single copy of rare allele in a pool of 96 individuals. In cases where minor allele calls differ between these two approaches, suggesting the base may be tri-allelic, both minor alleles were considered separately. We considered $F_i$ be the frequency of the minor allele under consideration in the $i^{th}$ pool. For each potential variant, we calculated the following metric, $STAT: \frac{(X - Y)}{Z}$, where $X$ is the range of $F_i$ over all pools, $Y$ is the sum of pairwise absolute differences in $F_i$ between pools and $Z$ is the standard deviation of $F_i$ over all pools. When $STAT$ was greater than the threshold generated by the noise model (described in the next paragraph), the minor allele was deemed valid and hence the base position was called polymorphic.

For any given base, we set $F_{i(2)}$ to be the frequency of the second-most frequent allele in pool $i$. Given pools of 96 individuals (192 chromosomes), the lower bound for $F_{i(2)}$ was 0.5% when there was only one rare allele present in the pool. Values of $F_{i(2)} < 0.5\%$ were assumed to be due to experimental noise. Therefore, we constructed a noise model, $F_{noise(2)}$, from the distribution of $F_{i(2)}$ across all positions from all pools where $F_{i(2)} < 0.5\%$. $F_{noise(2)}$ was different by reference alleles, CG and AT (Figure 2). We modelled the distribution of noise separately for CG and AT using two Gaussian models in R (function: nls). After the Gaussian models had been constructed, we performed 10000 simulations based on the assumption that only one pool contained a variant among various numbers of total pools. In each iteration, the null model randomly drawn $F_{i(2)}$ from the given $F_{noise(2)}$ distribution for each pool $i$. For the variant model, $F^{(2)}_i$ for the pool containing the variant was set to 0.5% while $F_{i \neq i}^{(2)}$ was drawn from $F_{noise(2)}$. After each iteration, we calculated the metric $STAT$. For a given number of total pools, the threshold for $STAT$ was defined as the 1% quantile of these 10000 iterations. The behaviour of the metric $STAT$ was investigated by altering different values for the number of total available pools and the pre-determined minor allele frequencies.
Verification of variants

To validate the rare variants we identified from sequence data, we used an allele-specific primer extension assay on the BeadXpress platform (VeraCode Golden Gate Genotyping kit, Illumina) to genotype 81 loci for all samples (Supplementary Table S1). Eight of the 81 loci were predicted to be noise by our algorithm and were included to assess the specificity. The predicted variants were selected first, followed by choosing presumably monomorphic sites that were technically compatible with the set of predicted variants. As a result, the predicted monomorphic sites were not sampled randomly over the distribution of distances between their \( \text{STAT} \) values and the threshold values generated by the noise model (Supplementary Figures S7 and S8). We failed to genotype seven variants owing to incorrect primers. In addition, 11 variants exhibited genotype patterns that were consistent with both being monomorphic and experimental failures. Thus, we designed primers for these 18 loci, 1 potentially tri-allelic locus, and 3 more loci that have been confirmed on BeadXpress to be sequenced using 454 Sanger sequencing. A small number of individuals were selected based on their pooling pattern for each variant. We inspected the traces to determine whether the Golden Gate assay failed or that the variants were incorrectly predicted by our method.

Assigning rare variants to individuals

For each verified variant, pools with \( F^{(2)} \) greater than 95th percentile of \( F_{\text{noise}}^{(2)} \) were deemed to contain carriers. We composed a list of potential carriers from these pools for each variant. Since variants were verified either by genotyping all individuals using the Golden Gate assay on the BeadXpress platform (Illumina) or by genotyping a small subset of individuals using Sanger sequencing, we evaluated these two sets of variants separately. For variants verified using Golden Gate assay, we determined the percentages of individuals with copies of alternate allele in the list of individuals in positive pools. For variants verified using Sanger sequencing, we calculated the expected number of heterozygous individuals by multiplying the average allele frequencies in positive pools by 192 and compared that with the number of individuals we identified.

Simulation of probability of unambiguously identifying carriers of rare variants

We used simulation to estimate the probability that carriers of a rare allele could be unambiguously identified by the pattern of pools positive for the rare allele. Assuming a fixed total population size (\( N \)) and a fixed number of individuals in each pool (\( np \)), if each individual is present in exactly one pool per dimension, the total number of pools per dimension is \( P = N/np \). In our test case, \( N = 2304 \) and \( np = 96 \), so \( P = 24 \). Define a singleton as an allele carried by exactly one person in the population. If a singleton is present, only one pool will be positive for the presence of the rare allele per dimension. To unambiguously assign the identity of a carrier, the number of patterns for positive pools must exceed the number of individuals in the population. The number of possible singleton patterns is \( p^D \), where \( D \) is the number of dimensions in which the population was pooled. Singleton alleles can be unambiguously assigned to an individual in the population as long as \( p^D > N \), so for our example, three dimensions are adequate to unambiguously identify singleton carriers in a population of 2304 individuals because \( 24^3 > 2304 \).

Unambiguously identifying carriers of rare alleles when more than one carrier is present is more challenging, but if the solution space of patterns is sparse enough, then it would be possible. The number of possible patterns increases as a function of the number of positive pools in each dimension, so we suspected that a modest number of additional dimensions would be adequate. Given the carriers of a rare allele, the number of possible patterns is \( \sim p^D \) (assuming no pools carrying more than one carrier in each dimension), but we found it theoretically challenging to directly calculate the likelihood that any doubleton pattern would resolve unambiguously as the union of only one pair of individuals present in the population.

In the absence of a theoretical solution, we simulated data under \( N = 2304 \) individuals, \( np = 96 \) individuals per pool, \( P = 24 \) pools per dimension, \( D \) between three and five pooling dimensions and \( a \) between one and five rare allele carriers. In each pooling dimension, the pool assignments of individuals were made such that they were orthogonal to previous dimensions as possible, producing a three, four or five digit pooling code for each individual. Then, for each value of \( a \), we randomly drew \( a \) individuals from our population. The pooling codes of these \( a \) individuals were combined to create an observed pattern of positive pools, and then we searched to see whether we could find any other combination of \( a \) individuals who were different from the initial individuals but generated the same observed pattern. We repeated for 100,000 iterations and calculated the average frequency with which the observed pattern was consistent with one and only one combination of an individual, yielding unambiguous identity for the carriers. We obtained the mean and standard deviation by repeating 20 sets of these 100,000 iterations.

RESULTS

High coverage sequencing of the \( \text{CRP} \) genomic region

We aimed to discover new rare variants in the \( \text{CRP} \) region by first deeply sequencing this region in individuals in the CARDIA cohort. In each individual, we amplified the \( \text{CRP} \) region and then pooled the amplicons from 96 individuals. The pooled PCR products were used to generate a barcoded library compatible with the Illumina Genome Analyzer (GA) IIx sequencer. We undertook two rounds of sequencing. In the first round, we identified putative polymorphisms from 24 of these pooled libraries containing disjoint samples, but did not attempt to map it to individuals. In the second round, we strived to discover
new rare variants and identify individual carriers simultaneously by re-assorting all individuals combinatorially into three sets of 24 pools for a total of 72 pools (Figure 1). Effectively, each sample was sequenced once in each set of pools and exactly three times across the 72 pools. Thus, the singleton alleles carried by a single individual should be confirmed in three different pools in three separate lanes, and we would be able to identify that individual based on the pattern of positive pools. We describe the quality metrics of the two rounds of sequencing in the METHODS. In both rounds, we filtered reads with low total Phred (basecall) qualities. The coverage per base in each pool averaged 900X in the first round, and 705X in the second round.

A null model of variant allele frequency

A base in a sequencing read can differ from the reference base due to experimental noises introduced at various stages rather than true polymorphisms. The Phred quality score measures the accuracy of sequencing PCR products, but the PCR products themselves could be synthesized with substitutions due to errors introduced earlier, for example, by DNA polymerases. Since we filtered reads based on their Phred qualities, we only removed sequencing errors and not errors introduced prior to sequencing. To differentiate a polymorphic base from these other experimental noises, we needed to first inspect the noise distribution in the sequencing experiment.

We considered $F_i^{(2)}$, the ratio of reads of the second-most frequent allele to the total reads at a given base in pool $i$. A reasonable distribution for $F_i^{(2)}$ across all positions is occasional large values, corresponding to polymorphic positions, superposed on background noise. If the DNA from the 96 diploid samples in a pool were amplified equally, the minimum value $F_i^{(2)}$ obtains for a true polymorphism would be $1/192$, $\sim 0.5\%$. We made this simplifying assumption and constructed the noise distribution, $F_{\text{noise}}^{(2)}$, by truncating the distribution of $F_i^{(2)}$, $i = 1...24$, at 0.5%. Although, $F_{\text{noise}}^{(2)}$ is non-negative, we found that a normal distribution fit reasonably. The distribution was also dependent on whether the reference base was C/G or A/T (Figure 2). The bases that form triple hydrogen bonds (guanine and cytosine) have a 10-fold lower mean error. In the first sequencing experiment, the noise model for A/T reference base had a mean of 0.0021 (standard deviation = 0.00094), and the noise model for C/G reference base had a mean of 0.00038 (standard deviation = 0.00019). For the second sequencing experiment, the noise model for A/T reference base had a mean of 0.0022 (standard deviation = 0.00089) and the noise model for C/G reference base had a mean of 0.00039 (standard deviation of 0.00023).

**STAT**: Detecting signals for true polymorphisms

We next constructed a metric to determine whether an alternative allele at a given position was likely polymorphic or could be generated from the noise model. Let $F_i^{(n)}$ be the frequency of the $n^{th}$ most frequent allele in the $i^{th}$ pool. For a given position, we calculate the following metric,

$$STAT = (X/Y) \ast Z,$$

where $X$ is the range of $F_i^{(n)}$ over all pools, $Y$ is the sum of pairwise absolute differences in $F_i^{(n)}$ between pools and $Z$ is the standard deviation of $F_i^{(n)}$ over all pools. If no variant exists in any pool, the range and standard deviation of $F_i^{(n)}$ would be small, since $F_i^{(n)}$ would only sample from the noise distribution, and the ratio of range to sum of pairwise absolute difference in $F_i^{(n)}$ would also be small, making the value of $STAT$ small (Figure 3A). If a rare variant exists in one of the pools, the range of $F_i^{(n)}$ over all pools increases, resulting in an increase in the value of $STAT$ (Figure 3B). To consider the case of multi-allelic SNPs, we calculated $STAT$ for $n = 2$, 3 and 4 for all positions in the sequenced region. Next we describe a way of finding a threshold for $STAT$ to approximately control the false discovery rate.

**Null distribution of $STAT$ and its sensitivity in detecting true polymorphisms**

We investigated the sensitivity of $STAT$ to detect true polymorphisms in simulations with various numbers of total pools. For a given number of $n$ total pools, we set one pool to contain one copy of the variant. In other words, the value of $F^{(2)}$ was set to be 0.5% in this positive pool. For the remaining $n-1$ pools, $F^{(2)}$ was drawn from the appropriate noise model. We then calculated $STAT$ using $F^{(2)}$ from all pools and repeated for 10,000 iterations to generate a distribution of $STAT$. We also generated a null distribution of $STAT$ assuming there was no variant in any of the pool. In this simulation, $F^{(2)}$ in all pools were drawn from the noise models. To be conservative, we took the first percentile from the distribution of $STAT$ with one positive pool and compared it with the median from the null distribution of $STAT$. We found that regardless of the number of total pools, the value of $STAT$ calculated from one positive pool was always greater than when none of the pools contained a variant (Figure 4A). Thus, at boundary of detection (one polymorphic allele in the entire set of pools), we have 99% power.

We also investigated the sensitivity of $STAT$ to detect true polymorphisms when $F^{(2)}$ in the positive pool varies. We found that $STAT$ was linearly related to the number of copies of variant in a given pool, regardless of the number of total pools (Figure 4B). These results suggested that $STAT$ was a sensitive and robust metric to use in identifying variants with frequencies as low as a single copy in a pool of 96 samples. We declared position $i$ and allele $n$ to be a candidate variant if its $STAT$ value exceeded the first percentile under the alternate distribution and $F^{(2)}$ in positive pools were >0.5% (Supplementary Table S2).

**Variant discovery and validation**

Assuming all samples in a pool were amplified equally, we used the first percentile $STAT$ values as a threshold and identified 130 putative variants in the 6 kb $CRP$ region. These 130 putative variants all have a minor allele
Figure 2. Distribution of noise by different reference allele $[F_{\text{noise}}^{(2)}]$. For any given position along the CRP reference region, we define $F_i^{(2)}$ as the frequency of second-most frequent allele in pool $i$. Assuming all 96 samples in a pool were amplified equally, the minimum value of $F_i^{(2)}$ is 0.5%. Noise, $F_{\text{noise}}^{(2)}$, is thus defined as any $F_i^{(2)} < 0.5\%$ across all pools and all positions. Distributions of noise differ with respect to the reference allele, regardless of the experiment (A: first sequencing experiment, and B: second sequencing experiment).
frequency of at least 0.5%. In the scenario where samples in a pool were not amplified equally, the minor allele frequency of a single copy variant might not be >0.5%. To take into account this possibility, we declared any position to be a putative variant if its \( STAT \) value exceeded the first percentile under the alternative distribution without requiring any minimum minor allele frequency. We obtained another eight candidates with minor allele frequencies ranging from 0.02 to 0.1%. Together, there were 138 putative variants. Of these 138,
54 polymorphisms were previously documented in the dbSNP database (Table 1). Ninety of the 138 variants are located in the intergenic region flanking CRP; 2 are in the 5' UTR region (chr1: 157,950,900–157,951,003); 8 are in the intron (chr1: 157,950,553–157,950,838); 13 are in exon 2 of CRP (chr1: 157,949,939–157,950,552); and 25 are in 3' UTR region (chr1: 157,948,704–157,949,938). Four putative variants were flagged as potentially tri-allelic. The majority of the variants discovered have a frequency <1%, indicating that rare polymorphisms are much more prevalent than common polymorphisms in CRP (Figure 5).

To assess the predictive power of our algorithm, we genotyped 73 candidate variants (genotyping all 138 was not possible due to oligo restrictions) and eight loci predicted to be monomorphic to verify using either the Illumina Golden Gate genotyping assay or the 454 Sanger sequencing assay (Supplementary Table S3).
Table 1. All variants discovered in sequencing experiments 1 and 2

| Chr 1 location (157, —, —) | Major allele | Minor allele | Experiments | Experiment 1 MAF | Experiment 2 MAF | dbSNP reference cluster number |
|-----------------------------|--------------|--------------|-------------|-----------------|-----------------|--------------------------------|
| 945 884                     | G            | C            | 2           | 2.47%           |                 |                                |
| 945 885                     | G            | T            | 2           | 2.58%           |                 |                                |
| 945 886                     | C            | T            | 2           | 2.43%           |                 |                                |
| 945 889                     | G            | C            | 2           | 1.97%           |                 |                                |
| 945 907                     | C            | G            | 2           | 5.16%           |                 |                                |
| 945 908                     | T            | G            | 2           | 2.87%           |                 |                                |
| 945 910                     | A            | T            | 2           | 5.51%           |                 |                                |
| 945 911                     | G            | A            | 2           | 0.67%           |                 |                                |
| 945 912                     | C            | T            | 2           | 0.55%           |                 |                                |
| 945 913                     | T            | G            | 2           | 0.64%           |                 |                                |
| 945 924                     | C            | T            | 1           | 0.50%           |                 |                                |
| 946 002                     | A            | G            | 2           | 0.27%           |                 |                                |
| 946 004                     | T            | G            | 1, 2        | 1.60%           | 0.28%           | rs192093196c                  |
| 946 009                     | T            | C            | 2           | 0.59%           |                 |                                |
| 946 023                     | G            | A            | 1, 2        | 2.15%           | 0.72%           |                                |
| 946 026                     | G            | A            | 1, 2        | 0.93%           | 0.88%           | rs3093079d                   |
| 946 055                     | C            | T            | 1, 2        | 1.13%           | 0.95%           |                                |
| 946 065                     | A            | G            | 1, 2        | 0.80%           | 0.91%           |                                |
| 946 110                     | G            | A            | 1, 2        | 0.52%           | 0.89%           |                                |
| 946 161                     | C            | A            | 1, 2        | 0.73%           | 0.51%           |                                |
| 946 173                     | G            | A            | 1, 2        | 0.93%           | 1.22%           | rs3093078d                   |
| 946 202                     | A            | C            | 1           | 0.70%           |                 |                                |
| 946 248                     | G            | C            | 1, 2        | 0.58%           | 0.56%           | rs144097876e                  |
| 946 260                     | A            | C            | 1, 2        | 16.12%          | 16.95%          | rs3093077d                   |
| 946 390                     | A            | C            | 1           | 0.64%           |                 |                                |
| 946 391                     | A            | T            | 1           | 0.63%           |                 |                                |
| 946 392                     | T            | C            | 1           | 0.70%           |                 |                                |
| 946 453                     | G            | A            | 1, 2        | 0.72%           | 0.53%           |                                |
| 946 471                     | G            | A            | 1, 2        | 0.75%           | 0.89%           | rs192715884d                  |
| 946 487                     | T            | A            | 1, 2        | 0.63%           | 0.88%           |                                |
| 946 537                     | G            | T            | 1, 2        | 14.50%          | 15.29%          | rs3093075d                   |
| 946 550                     | C            | T            | 1, 2        | 0.57%           | 0.92%           |                                |
| 946 576                     | T            | C            | 1, 2        | 0.79%           | 1.20%           |                                |
| 946 577                     | C            | T            | 2           | 0.94%           |                 |                                |
| 946 579d                    | A            | G            | 1, 2        | 1.14%           | 2.47%           | rs3093074d                   |
| 946 582                     | G            | C            | 1, 2        | 0.85%           | 1.24%           |                                |
| 946 642                     | T            | C            | 1, 2        | 0.63%           | 0.72%           |                                |
| 946 688                     | G            | A            | 1, 2        | 0.88%           | 0.79%           |                                |
| 946 769                     | C            | A            | 1b, 2       | 0.03%           | 1.11%           |                                |
| 946 857                     | C            | A            | 1, 2        | 0.59%           | 0.80%           |                                |
| 946 866                     | A            | G            | 1           | 0.97%           |                 |                                |
| 946 925                     | C            | T            | 2           | 0.62%           |                 |                                |
| 946 980                     | C            | T            | 1, 2        | 1.30%           | 2.51%           |                                |
| 947 001                     | A            | G            | 1, 2        | 0.75%           | 0.90%           | rs3093073d                   |
| 947 039                     | C            | T            | 1, 2        | 0.64%           | 0.85%           |                                |
| 947 053                     | C            | A            | 1, 2        | 0.92%           | 0.98%           | rs3093072d                   |
| 947 123d                    | C            | G            | 1, 2        | 1.16%           | 1.27%           | rs3093071d                   |
| 947 153                     | G            | A            | 1, 2        | 1.16%           | 1.26%           | rs116295897c                  |
| 947 242                     | G            | A            | 1, 2        | 0.63%           | 1.21%           |                                |
| 947 336                     | T            | C            | 1, 2        | 0.72%           | 0.88%           |                                |
| 947 347                     | G            | A            | 2           | 1.25%           |                 |                                |
| 947 350                     | C            | T            | 1, 2        | 0.67%           | 0.75%           |                                |
| 947 355                     | C            | G            | 1, 2        | 0.87%           | 1.21%           |                                |
| 947 390                     | G            | A            | 1, 2        | 2.31%           | 1.16%           |                                |
| 947 441                     | T            | G            | 1, 2        | 1.16%           | 1.50%           | rs3093070d                   |
| 947 474                     | A            | T            | 1, 2        | 0.63%           | 0.68%           | rs112572605d                  |
| 947 485                     | C            | A            | 1, 2        | 0.65%           | 1.06%           | rs28086230d                  |
| 947 492                     | T            | C            | 1, 2        | 24.14%          | 25.12%          | rs143181976c                  |
| 947 548                     | C            | A            | 1b, 2       | 0.03%           | 1.40%           |                                |
| 947 579                     | A            | G            | 1b, 2       | 0.04%           | 0.67%           |                                |
| 947 732                     | C            | T            | 2           | 0.83%           |                 | rs114985520f                  |
| 947 755                     | A            | C            | 1, 2        | 9.14%           | 10.48%          | rs3093069f                   |
| 947 939                     | A            | T            | 1, 2        | 0.84%           | 0.80%           | rs3093068d                   |
| 947 988                     | G            | C            | 1, 2        | 13.45%          | 14.39%          |                                |

(continued)
Table 1. Continued

| Chr 1 location | Major allele | Minor allele | Experiments | Experiment 1 MAF | Experiment 2 MAF | dbSNP reference cluster number |
|----------------|--------------|--------------|-------------|------------------|------------------|-------------------------------|
| 947996         | T            | G            | 1, 2        | 0.52%            | 0.59%            |                               |
| 948035         | C            | G            | 1, 2        | 0.55%            | 1.16%            |                               |
| 948080         | A            | G            | 1, 2        | 0.88%            | 0.91%            |                               |
| 948217         | T            | C            | 1, 2        | 0.82%            | 0.68%            |                               |
| 948322         | G            | T            | 1, 2        | 0.58%            | 1.11%            | rs1050460866                  |
| 948367         | A            | T            | 1, 2        | 1.24%            | 1.12%            |                               |
| 948391         | A            | G            | 1, 2        | 0.90%            | 1.20%            | rs3093080d                    |
| 948424         | G            | A            | 1, 2        | 0.58%            | 0.06%            |                               |
| 948552         | G            | T            | 1, 2        | 1.07%            | 1.55%            |                               |
| 948642         | G            | A            | 1, 2        | 0.88%            | 0.97%            |                               |
| 948660         | T            | C            | 1, 2        | 4.35%            | 4.66%            | rs2808631d                    |
| 948791         | G            | A            | 1, 2        | 0.65%            |                  |                               |
| 948800         | G            | C            | 1, 2        | 0.02%            | 0.60%            |                               |
| 948819         | T            | 1, 2         | 0.73%       | 1.38%            |                  |                               |
| 948857         | C            | T            | 1, 2        | 24.74%           | 25.22%           | rs1205d                       |
| 948901         | A            | G            | 1, 2        | 0.74%            | 0.72%            |                               |
| 948909         | G            | A            | 1, 2        | 0.86%            | 0.95%            | rs17860477d                   |
| 949036         | G            | A            | 1, 2        | 1.01%            | 1.50%            | rs34457301d                   |
| 949054d        | A            | G            | 1, 2        | 1.97%            | 1.88%            | rs6413465d                    |
| 949112         | T            | C            | 1, 2        | 0.83%            | 0.88%            | rs3093067d                    |
| 949152         | A            | G            | 1, 2        | 2.44%            | 2.21%            |                               |
| 949158         | G            | A            | 1, 2        | 0.57%            |                  |                               |
| 949174         | G            | A            | 2           | 0.62%            |                  |                               |
| 949183         | A            | G            | 1           | 0.70%            |                  |                               |
| 949233         | T            | C            | 1, 2        | 0.94%            | 1.06%            |                               |
| 949256         | A            | G            | 1, 2        | 1.05%            | 0.70%            | rs17860478d                   |
| 949294         | A            | G            | 1, 2        | 1.09%            | 1.26%            | rs17860479d                   |
| 949332         | C            | G            | 2           | 0.56%            |                  |                               |
| 949362         | C            | T            | 1, 2        | 0.91%            | 1.31%            | rs35475549d                   |
| 949715         | G            | A            | 1, 2        | 22.29%           | 22.69%           | rs1130864d                    |
| 949723         | G            | T            | 1, 2        | 9.62%            | 10.82%           | rs3093066d                    |
| 949725         | A            | G            | 1, 2        | 0.83%            | 1.03%            |                               |
| 949831         | G            | C            | 1, 2        | 0.71%            | 1.07%            | rs113028201c                  |
| 949833         | C            | T            | 2           | 0.87%            |                  |                               |
| 949837         | G            | A            | 1, 2        | 1.09%            | 1.32%            | rs3093065d                    |
| 949882         | G            | A            | 2           | 0.83%            |                  |                               |
| 949887         | G            | A            | 1           | 0.80%            |                  | rs35664422d                   |
| 949888         | C            | A            | 1           | 0.71%            | 0.69%            | rs35031149d                   |
| 949895         | G            | A            | 1, 2        | 0.82%            | 1.10%            | rs34034943d                   |
| 949906         | G            | C            | 1, 2        | 1.11%            | 1.29%            | rs35972670d                   |
| 949908c        | G            | C            | 1, 2        | 0.83%            | 1.09%            |                               |
| 950050         | C            | G            | 1, 2        | 0.80%            | 0.51%            | rs1800947d                    |
| 950062         | C            | G            | 1, 2        | 3.60%            | 3.17%            |                               |
| 950068         | A            | G            | 1, 2        | 0.77%            | 1.34%            |                               |
| 950134         | G            | A            | 1, 2        | 0.07%            | 0.58%            |                               |
| 950240         | A            | G            | 1, 2        | 0.69%            | 1.41%            |                               |
| 950264         | C            | G            | 1, 2        | 0.57%            | 0.85%            |                               |
| 950307         | C            | T            | 1, 2        | 0.67%            | 0.91%            | rs146258487c                  |
| 950401         | C            | T            | 1, 2        | 0.09%            | 0.61%            |                               |
| 950411         | C            | T            | 1           | 0.60%            |                  |                               |
| 950420         | C            | T            | 2           | 0.59%            |                  | rs140571492c                  |
| 950438         | G            | A            | 1, 2        | 0.59%            | 0.94%            | rs77832441e                   |
| 950473         | A            | G            | 2           |                  | 0.79%            |                               |
| 950480         | G            | A            | 1, 2        | 0.94%            | 0.83%            | rs116776267c                  |
| 950485         | C            | T            | 1, 2        | 1.46%            | 1.24%            | rs36061058d                   |
| 950520         | C            | T            | 1, 2        | 0.04%            | 1.20%            | rs148534477c                  |
| 950673         | A            | C            | 1, 2        | 1.07%            | 0.63%            |                               |
| 950676         | C            | A            | 1, 2        | 0.99%            | 1.14%            |                               |
| 950681         | A            | C            | 2           |                  | 1.11%            |                               |
| 950682         | C            | A            | 2           |                  | 1.02%            |                               |
| 950683         | A            | C            | 2           |                  | 1.01%            | rs5778129d                    |
| 950711         | C            | A            | 2           |                  | 6.86%            |                               |
| 950712         | A            | C            | 1, 2        | 1.34%            | 5.48%            |                               |
| 950713         | T            | A            | 2           | 3.87%            |                  |                               |
| 950714         | G            | C            | 2           | 2.38%            |                  |                               |

(continued)
Fifty-four of the 73 candidates were verified to be polymorphic by genotyping all CARDIA individuals using the Golden Gate assay. An additional 14 candidates were verified by Sanger sequencing in a subset of individuals. Of the eight loci which were deemed monomorphic by V-sieve, three were found to be polymorphic; two by Golden Gate assay and one by Sanger sequencing. Only one of three false negatives, rs114985520, was reported in the dbSNP database (version 135). Overall, we verified 68 of 73 putative variants, including two tri-allelic loci. This results in a positive predictive value of 93% (Table 2).

In the second sequencing experiment, we identified 157 variants across the 72 pools. We obtained another four candidate variants if we relaxed the minimum frequency requirement on the alternative alleles. The majority of these predicted variants were also discovered in the first sequencing experiment (73%) (Table 3 and Supplementary Figure S9). We used the same set of verified variants to compare the sensitivity of the second sequencing experiment to the first. In total, 65 of 69 predicted variants were confirmed, including two tri-allelic variants. We achieved a 94% positive predictive rate in this experiment.
Identifying individuals with new rare variants from sequencing experiment

Conducting two sets of experiments to discover new rare variants and then to identify individuals carrying these variants is both time consuming and labor intensive. In our second sequencing experiment, we strived to accomplish these two goals simultaneously and evaluated our ability in identifying individuals carrying the 63 verified variants. For each variant, pools with $F^{(2)}$ greater than the 95% $F_{\text{noise}}^{(2)}$ threshold were considered to contain carriers. Since each individual was sequenced three times in three different pools, most individuals in the second sequencing experiment had a unique pooling pattern that we can use to identify them. Based on the pattern of positive pools, we gathered a list of potential individuals carrying each variant. We cannot narrow down 16 of the 63 variants to specific individuals since they were too prevalent (found in 60% of pools and had an average frequency >1% across pools). For the other 47 variants, the lists of potential carriers included at least one confirmed person with the variants.

We further explored whether for a given variant, all confirmed individuals carrying the alternate allele were on the list of potential carriers, a proxy for the sensitivity. For 33 variants verified by Golden Gate assay, we considered both the number of potential carriers (target size) and the percentage of confirmed individuals (success rate). The target size is an estimate of the number of capillary sequencing experiments that we would have performed if we did not use the Golden Gate assay to genotype all samples. We found that for 16 rare polymorphisms with frequencies <0.05% in a pool, we were able to identify all individuals carrying the alternate allele (Figure 6). The median target size was four, suggesting that we only needed to genotype four subjects per variant to identify all individuals carrying the alternate allele using capillary sequencing. For four variants with frequencies between 0.05 and 0.1%, we were able to identify all confirmed individuals, and the median target size was 56 people. For nine variants with frequencies between 0.1 and 0.5%, we were able to identify, on average, 67% of confirmed individuals. The median target size was 99. For four variants with 0.5–1% frequency, we were able to identify 81% of confirmed individuals. The median pool size was 960, indicating that the target size grew nearly exponentially in the minor allele frequency. We performed a total of 59 capillary sequencing experiments in a subset of individuals for 14 variants (Figure 7). On average, we genotyped 4.2 subjects per variant. We calculated the expected number of individuals with a variant by multiplying the average allele frequencies in the positive pools by 192. For nine variants, we identified all individuals carrying that variant by capillary sequencing. In summary, our success in identifying all individuals with a specific variant decreased as a variant became more common. For rare variants with frequencies <0.5% (equivalent to having 11 copies in our cohort), we were able to achieve high success rate in genotyping <100 individuals per variant.

**DISCUSSION**

Common SNPs identified in GWAS to date generally account for small effects, leading to the hypothesis that rare variants might also play a functional role in determining phenotypes. Whole-genome sequencing of individuals is not yet feasible due to the large cost. An alternative approach is to sequence a pool of multiple individuals, identify the polymorphisms present in the pooled sample, and then genotype the discovered
polymorphisms in each individual to assign individual genotypes. A major challenge with a pooled variant-detection approach is discriminating between background technical noise and true rare polymorphisms. To address this, we have developed a detection algorithm, V-sieve, which is able to efficiently screen a candidate gene for idiosyncratic variants in a large population, with a positive predictive rate of 93%. We were able to identify polymorphisms with an allele frequency as low as 0.02%, corresponding to a singleton rare allele carried by one individual in our cohort of 2283.

There are several major advantages with our method. First, the noise model was constructed directly from the experimental data, effectively normalizing for technical noise. The method is robust against false positives and provides a high degree of confidence in the identified variants. Additionally, the V-sieve approach is scalable and can be applied to genome-wide association studies.

Figure 6. Median size of pools containing potential individuals carrying variants and the median percentage of verified individuals with variants of different frequencies.

Figure 7. Number of PCR experiments performed, number of verified individuals with variants and the estimated proportions of individuals identified for each variant.
noise specific to that experiment and requiring no additional experimental control. Second, even though we focused on novel rare variant discovery, we were equally successful in detecting common variants. Thirteen of the 138 variants we identified had an allele frequency >5%, all of them known in dbSNP. Although we assumed equal contribution of each individual in each pool and set allele frequency of 0.5% (1 copy in 192 chromosomes) as a lower bound, the robustness of our method still enables us to discover variants with allele frequencies lower than that. We identified eight singleton SNPs with allele frequencies <0.5% in any specific pooled library, presumably due to imprecise pooling of the original 96 samples in each pool. Four of these putative variants were genotyped in all individuals, and three were confirmed to be real polymorphisms. Finally, our methods identified several tri-allelic polymorphisms.

We included both sites predicted to be polymorphic and monomorphic by V-sieve for verification. The eight presumed monomorphic sites were chosen to be compatible with the predicted variants already in the set. Technical parameters such as primer compatibility defined by the Illumina BeadExpress platform and distances from predicted variant sites were considered. As a result, these eight sites were not sampled randomly over the distribution of distances between their STAT values and the corresponding threshold values generated by the noise model (Supplementary Figure S8). They were in fact biased towards the STAT values close to the threshold, making them far from a comprehensive set of loci to test for false negatives. It was thus not that surprising that three of them turned out to be true rare variants. Two of the three sites have minor allele frequencies <0.1%. We suspect that at such low minor allele frequency levels, the unequal contribution of each sample in the pools becomes a bigger factor in influencing whether a site is called variant by V-sieve. This suggests that for some experiments where identifying as many rare variants is a higher priority than limiting the number of false positives, lowering the threshold values set by the noise models might help uncovering additional true variants.

Sequencing pooled samples, followed by genotyping predicted variants in all individual samples, can be more cost-effective than sequencing all individuals in the target population. This approach, however, is labour-intensive and still costly due to the genotyping experiments using Golden Gate assay required to individually identify singleton heterozygotes. Our second approach used a carefully designed strategy to pool samples such that most of samples can be uniquely identified by their presence in exactly 3 out of 72 pools of 96 individuals, enabling us to simultaneously screen for rare variants, and pinpoint individuals carrying them on the basis of which pools were positive for the rare allele. The efficiency of this approach depends on the accuracy of identifying carriers. Unambiguous determination of carrier identity is made possible by combinatorial pooling, where each individual is present in multiple dimensions (pools). In our test case, when only one rare allele carrier is present in the population, we can unambiguously identify carrier identity using three dimensions of parallel pooling, such that every individual is present in exactly three pools (Figure 8). When two carriers are present in the population, with three dimensions we can unambiguously identify both carriers ~53% of the time. In other words, 53% of the time, we do not need additional experiments to verify the carriers. As the number of copies of variant increases, the probability of unambiguously identifying carriers decreases.

However, additional pooling dimensions can overcome these ambiguities. With four dimensions of parallel pooling, even when five carriers are present in the overall population we can unambiguously resolve the identities of all five carriers >75% of the time. This probability increases to 98% with five dimensions of parallel pooling. The drawbacks of our multi-dimensional pooling strategy include the cost of additional sequencing libraries, and the increased amount of DNA required from each sample. Therefore, we suggest that the most comprehensive and efficient strategy for the identification of all variants in a population and the assignment of individual genotype to all individuals in the population, both in terms of cost and labour, may be to use the combinatorial pooling strategy to discover new variants and identify carriers of rare variants with up to five carriers in the population, with subsequent genotyping of more frequent polymorphism in all individuals, to unambiguously assign genotype at the more common variants. We believe this efficient algorithm will be useful in both screening genomic regions known to associate with particular phenotypes from GWAS for rare variants, and in screening large populations for an excess burden of rare variation in candidate genes (14–17).

**ACCESSION NUMBERS**

SNPs have been submitted to dbSNP.
SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1–9.

FUNDING

National Heart, Lung and Blood Institute [1R01HL088531-01A1]; ARRA supplement [3R01HL088531-02S1 to C.T.L.C.]. Funding for open access charge: National Heart, Lung and Blood Institute.

Conflict of interest statement. None declared.

REFERENCES

1. N'Diaye, A., Chen, G.K., Palmer, C.D., Ge, B., Tayo, B., Mathias, R.A., Ding, J., Nalls, M.A., Adeyemo, A., Adoue, V., et al. (2011) Identification, replication, and fine-mapping of loci associated with adult height in individuals of African ancestry. *PLoS Genet.*, 7, e1002298.

2. Duncan, E.L., Danoy, P., Kemp, J.P., Leo, P.J., McCloskey, E., Nicholson, G.C., Eastell, R., Prince, R.L., Eisman, J.A., Jones, G., et al. (2011) Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS Genet.*, 7, e1001372.

3. Elks, C.E., Perry, J.R., Sulem, P., Chasman, D.I., Franceschini, N., He, C., Lunetta, K.L., Visser, J.A., Byrne, E.M., Cousminer, D.L., et al. (2010) Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.*, 42, 1077–1085.

4. Chen, C.T., Fernandez-Rhodes, L., Brzyski, R.G., Carlson, C.S., Chen, Z., Heiss, G., North, K.E., Woods, N.F., Rajkovic, A., Kooperberg, C., et al. (2012) Replication of loci influencing ages at menarche and menopause in Hispanic women: the Women’s Health Initiative SHARE Study. *Hum. Mol. Genet.*, 21, 1419–1432.

5. Stolk, L., Perry, J.R., Chasman, D.I., He, C., Mangino, M., Sulem, P., Barbalic, M., Broer, L., Byrne, E.M., Ernst, F., et al. (2012) Meta-analyses identify 13 loci associated with age at menopause and high-fat diet. *Am. J. Hum. Genet.*, 90, 260–268.

6. Peters, U., Hutter, C.M., Hsu, L., Schumacher, F.R., Conti, D.V., Carlson, C.S., Edlund, C.K., Haile, R.W., Gallinger, S., Zanke, B.W., et al. (2012) Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.*, 131, 217–234.

7. Conde, L., Halperin, E., Akers, N.K., Brown, K.M., Smedby, K.E., Rothman, N., Nieters, A., Slager, S.L., Brooks-Wilson, A., Agana, L., et al. (2010) Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat. Genet.*, 42, 661–664.

8. Goode, E.L., Chen, L., Song, H., Ramus, S.J., Notaridou, M., Lawrenson, K., Widschwendter, M., Vierkant, R.A., Larson, M.C., Kjaer, S.K., et al. (2010) A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat. Genet.*, 42, 874–879.

9. Kote-Jarai, Z., Olama, A.A., Giles, G.G., Severi, G., Schleutker, J., Weischer, M., Campa, D., Riboli, E., Key, T., Gronberg, H., et al. (2011) Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat. Genet.*, 43, 785–791.

10. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA.*, 106, 9362–9367.

11. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R. Jr, Liu, K. and Savage, P.J. (1988) CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J. Clin. Epidemiol.*, 41, 1105–1116.

12. Hughes, G.H., Cutter, G., Donahue, R., Friedman, G.D., Hulley, S., Hunkeler, E., Jacobs, D.R. Jr, Liu, K., Orden, S., Pirie, P., et al. (1987) Recruitment in the Coronary Artery Disease Risk Development in Young Adults (Cardia) study. *Control. Clin. Trials.*, 8, 68S–73S.

13. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18, 1851–1858.

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

15. Madsen, B.E. and Browning, S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, 5, e1000384.

16. Morris, A.P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, 34, 188–193.

17. Wu, M.C., Lee, C.S., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89, 82–93.