BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 13 2012, pages 1745–1751
doi:10.1093/bioinformatics/bts263

Genetics and population analysis Advance Access publication May 3, 2012

SEQCHIP: a powerful method to integrate sequence and genotype data for the detection of rare variant associations

Dajiang J. Liu[1,*] and Suzanne M. Leal[2,*]

[1]Department of Biostatistics, Center of Statistical Genetics, University of Michigan, Ann Arbor, MI, 48109 and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Next-generation sequencing greatly increases the capacity to detect rare-variant complex-trait associations. However, it is still expensive to sequence a large number of samples and therefore often small datasets are used. Given cost constraints, a potentially more powerful two-step strategy is to sequence a subset of the sample to discover variants, and genotype the identified variants in the remaining sample. If only cases are sequenced, directly combining sequence and genotype data will lead to inflated type-I errors in rare-variant association analysis. Although several methods have been developed to correct for the bias, they are either underpowered or theoretically invalid. We proposed a new method SEQCHIP to integrate genotype and sequence data, which can be used with most existing rare-variant tests.

Results: It is demonstrated using both simulated and real datasets that the SEQCHIP method has controlled type-I errors, and is substantially more powerful than all other currently available methods.

Availability: SEQCHIP is implemented in an R-Package and is available at http://linkage.rockefeller.edu/suzanne/seqchip/Seqchip.htm

Contacts: dajiang@umich.edu or sleal@bcm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 2, 2012; revised on April 8, 2012; accepted on April 27, 2012

1 INTRODUCTION

There is solid evidence that rare variants play an important role in the etiology of complex traits (Bodmer and Bonilla, 2008; Cohen *et al.*, 2004; Cohen *et al.*, 2006; Fearnhead *et al.*, 2004; Ji *et al.*, 2008; Kryukov *et al.*, 2007; Romeo *et al.*, 2007; Romeo *et al.*, 2009). The development and implementation of next-generation sequencing in genetic studies of complex traits have made possible the detection of rare variant associations. However, it is still very expensive to sequence a large number of samples at high coverage depth, which is necessary to accurately detect rare variants for association studies. Instead of sequencing all samples, a two-stage strategy can be applied, where a subset of the sample is sequenced to discover variants, and the identified variants are then genotyped in the remaining sample. Genotyping currently is considerably less expensive than sequencing, therefore given a fixed budget, a two-stage design can be much more powerful than a one-stage design, where only sequencing is used. This is because for the same financial expenditure, genetic information from a much larger number of samples can be extracted and analyzed in a two-stage study. In fact, the two-stage study design has been widely applied. Many rare variant associations were identified for a number of clinically important traits, including colorectal adenomas (Fearnhead *et al.*, 2004), age-related macular degeneration (Raychaudhuri *et al.*, 2011), lipids level (Sanna *et al.*, 2011) and inflammatory bowel disease (Rivas *et al.*, 2011).

In a two-stage study, candidate genes that were previously implicated in the etiology of complex trait through genome-wide association studies or functional studies may be sequenced to identify rare variants. It was demonstrated that sequencing >500 cases can uncover variants that can explain over 80% of the locus population attributable risk (Liu and Leal, 2010b). When more sophisticated methods are used for selecting samples, e.g. calculating expected number of causal variants for each individual and sequencing the individuals with the maximal counts (Edwards *et al.*, 2010), it is potentially possible to enrich for causal variants using a smaller sample size than if samples are randomly selected. To make the two-stage-design cost effective for large scale studies, the commercially available exome chip can also be customized and up to 30 000 variants can be added. Alternatively, custom genotyping arrays can also be developed.

For the two-stage study design integrating sequencing and genotyping, it was shown that for a fixed number samples sequenced, sequencing only cases can be more powerful than sequencing a balanced number of cases and controls, for detecting associations with causal variants that are enriched in cases (Longmate *et al.*, 2010). This is because a larger portion of low frequency causal variants can be identified by sequencing only cases. However, it has also been shown (Li and Leal, 2009) that type-I errors will be inflated for two-stage studies where only cases are sequenced to discover variants and naïve analysis is implemented, which directly integrates sequence and genotype data and compares aggregated variant frequencies between cases and controls. A straightforward correction is to exclude the sequence data used for variant discovery and use only the genotyped samples in the downstream association analysis (GSO – genotype samples only). However, this approach does not make full use of the available data and is underpowered, especially when a large portion (e.g. >50%) of affected individuals are sequenced for variant discovery. In Longmate *et al.* (2010), the authors suggested removing one variant carrier from the

*To whom correspondence should be addressed.

sequenced sample per variant nucleotide site (ROPS). When data are integrated using the ROPS method, rare variant association tests have controlled type-I error. However, this method is not without problems: (i) when there are many variant sites within the analyzed region, the ROPS method may result in removing a large number of variant carriers from the analysis, which is highly inefficient; and (ii) When covariates are included in the analysis, variant carriers may not be interchangeable under the null hypothesis, which makes it statistically invalid to apply the ROPS method.

Given the limitations of existing methods, it is desirable to have a statistical method that can correct for the bias induced by the two-stage study design and can be integrated with existing rare variant association methods (Bhatia *et al.*, 2010; Han and Pan, 2010; Ionita-Laza *et al.*, 2011; Liu and Leal, 2010a; Madsen and Browning, 2009; Neale *et al.*, 2011; Price *et al.*, 2010; Wu *et al.*, 2011). In this article, SEQCHIP, a likelihood-based method for integrating sequence and genotype data was developed to correct for the bias created by two-stage study design. The method corrects for the variant genotypes obtained from sequencing, such that the corrected genotypes approximately follow the same distribution as that of the genotyped samples. The method can be used with any existing rare variant association tests that can analyze uncertain genotypes, e.g. imputed genotypes. In particular, the weighted sum statistics (WSS) have been extended to incorporate imputed genotypes ()(Zawistowski *et al.*, 2010). The extensions of the test of the aggregated number of rare variants (ANRV; Morris and Zeggini, 2010) is straightforward where the genotype coding at each variant site can be replaced by 'dosage' (Guan and Stephens, 2008; Li *et al.*, 2010; Zheng *et al.*, 2011). The variable threshold (VT) test (Price *et al.*, 2010) computes the ANRV test statistics for each frequency threshold and then uses their maximum as the test statistic. To control for multiple testing, *p*-values of VT are obtained empirically. The three tests are used to evaluate the performance of different data integration methods, i.e. ROPS, SEQCHIP and GSO.

We evaluated the performance of SEQCHIP through extensive simulations. Genetic data were generated using a realistic population genetic model (Kryukov *et al.*, 2009). Disease phenotypes were simulated based upon parameters estimated from complex trait studies. We showed that when data are integrated using SEQCHIP, type-I errors for rare variant association tests are well controlled and the power is consistently higher than integrating data through ROPS (Longmate *et al.*, 2010) or analyzing genotyped samples only (GSO) in all the scenarios that were examined.

As an application of the SEQCHIP method, we re-analyzed samples from a case control study of colorectal adenomas (Fearnhead *et al.*, 2004). Sequence data were generated on five genes, which were previously implicated in the etiology of colorectal adenomas (Frayling *et al.*, 1998; Kim *et al.*, 2000; Lamlum *et al.*, 2000; Lipton *et al.*, 2003). In the study, 124 cases were sequenced to discover variants, and the identified variants were followed up and genotyped in 483 controls. Association analysis was carried-out by directly comparing total variant frequencies between sequenced cases and genotyped controls, and therefore the estimated *p*-values could be inflated. We re-analyzed this dataset, where sequence and genotype data were integrated using SEQCHIP and ROPS methods. The ANRV, WSS and VT tests were implemented to detect associations with rare variants. The association signal was statistically significant when the correction for the two-stage design was made using SEQCHIP, but not when ROPS was used. It should be noted that the GSO method could not be applied because all cases were sequenced and only controls were genotyped. The results verified the association with colorectal adenomas using valid methods and established that the SEQCHIP method is essential for integrating sequence and genotype data in cost-effective two-stage association studies.

2 METHODS

2.1 SEQCHIP method

We assume that there are N^{A} affected individuals and N^{U} unaffected individuals in the sample. Among the affected individuals, N^{S} (i.e. individuals 1,2, …, N^{S}) are sequenced to discover variants. An additional $N^{\mathrm{G}}=N^{\mathrm{A}}-N^{\mathrm{S}}$ cases (i.e. individuals $N^{\mathrm{S}}+1$, …, N^{A}) and N^{U} controls are genotyped at the variant nucleotide sites that were uncovered in the sequence sample. The multi-site genotype for an individual i at the candidate gene locus is denoted by a vector, i.e.

$$X_i=\left(\begin{pmatrix}x_i^{1,1}\\ x_i^{1,2}\end{pmatrix},\begin{pmatrix}x_i^{2,1}\\ x_i^{2,2}\end{pmatrix},\cdots,\begin{pmatrix}x_i^{K,1}\\ x_i^{K,2}\end{pmatrix}\right),$$

where each entry k represents a site with di-allelic single-nucleotide variations, and $x_i^{k,j}$ is an indicator of whether the j^{th} variant at site k is the minor allele. The total number of minor alleles observed at site k in the sequenced sample follows a truncated binomial distribution, i.e.

$$\begin{aligned}&\Pr\left(\sum_{i=1}^{N^{\mathrm{S}}}x_i^{k,1}+x_i^{k,2}=m\middle|\sum_{i=1}^{N^{\mathrm{S}}}x_i^{k,1}+x_i^{k,2}>0\right)\\&=\frac{\binom{2N^{\mathrm{S}}}{m}(p_k)^m\left(1-p_k\right)^{2N^{\mathrm{S}}-m}}{\left(1-\left(1-p_k\right)^{2N^{\mathrm{S}}}\right)},\end{aligned}\qquad(1)$$

where p_k is the minor allele frequency (MAF) at site k.

It is clear from the above equation that the expected number of minor alleles at site k in the sequenced cases satisfies

$$E\left(\sum_{i=1}^{N^{\mathrm{S}}}x_i^{k,1}+x_i^{k,2}\middle|\sum_{i=1}^{N^{\mathrm{S}}}x_i^{k,1}+x_i^{k,2}>0\right)=\frac{2N^{\mathrm{S}}p_k}{\left(1-\left(1-p_k\right)^{2N^{\mathrm{S}}}\right)}\qquad(2)$$

Therefore, a naïve estimate of allele frequencies (i.e. the mean number of minor alleles per chromosome) will be inflated by a factor of $1/c_k$, where $c_k=1-\left(1-p_k\right)^{2N^{\mathrm{S}}}$.

The idea behind the SEQCHIP method is to correct for the genotypes of the samples that are sequenced, such that corrected sequence genotypes approximately follow the same distribution as that of the genotyped samples. Specifically, auxiliary variables are defined for the samples that are sequenced, i.e.

$$\vec{\tilde{X}}_i=\left(\begin{pmatrix}\tilde{X}_i^{1,1}\\ \tilde{X}_i^{1,2}\end{pmatrix},\begin{pmatrix}\tilde{X}_i^{2,1}\\ \tilde{X}_i^{2,2}\end{pmatrix},\cdots,\begin{pmatrix}\tilde{X}_i^{K,1}\\ \tilde{X}_i^{K,2}\end{pmatrix}\right),$$

where the marginal distribution for each entry element $\tilde{X}_i^{k,j}$ satisfies

$$\begin{aligned}\Pr\left(\tilde{X}_i^{k,j}\middle|X_i^{k,j}=1\right)&=\left(1-p_k\right)^{2N^{\mathrm{S}}}\times\delta\left(\tilde{X}_i^{k,j}=0\right)\\&\quad+\left(1-\left(1-p_k\right)^{2N^{\mathrm{S}}}\right)\times\delta\left(\tilde{X}_i^{k,j}=1\right),\ j=1,2,1\leqslant k\leqslant K\\\Pr\left(\tilde{X}_i^{k,j}\middle|X_i^{k,j}=0\right)&=\delta\left(\tilde{X}_i^{k,j}=0\right)\end{aligned}\qquad(3)$$

It was shown in Supplementary Appendix SA and Supplementary Figure S1 that the modified genotype coding approximately follow the same

distribution as that of the genotyped samples. The allele frequencies in (3) are generally unknown. In practice, they can be estimated from the data. Details for the MAF estimators are displayed in Supplementary Appendix SB.

Instead of analyzing the original dataset $\left(\vec{X}_i, Y_i\right), i=1,\cdots,N^S, N^S+1, \cdots, N^A+N^U$, the 'new' dataset with corrected sequence genotypes are analyzed, i.e. $\left(E\left(\tilde{\vec{X}}_i \middle| \vec{X}_i\right), Y_i\right), i=1,\cdots,N^S$ and $\left(\vec{X}_i, Y_i\right), i=N^S+1, \cdots, N^A, N^A+1,\cdots,N^A+N^U$ $E\left(\tilde{\vec{X}}_i \middle| \vec{X}_i\right)$ can be considered as the 'dosage' for the corrected genotypes conditional on the observed genotypes. In principle, it is also possible to use a mixture likelihood which integrates uncertainties in the modified genotype. However, when multiple rare variants are analyzed, calculating the mixture likelihood can be computationally intensive and numerically unstable. Therefore, the mixture likelihood approach is not pursued.

In principle, all rare variant association tests that can analyze uncertain genotypes (e.g. imputed genotypes) can be directly applied to analyze the SEQCHIP corrected genotypes. Score tests can be implemented, and standard permutation procedure can be used to obtain empirical p-values.

2.2 ROPS and GSO method

We compared the SEQCHIP method with the ROPS (Longmate *et al.*, 2010) and GSO methods. Specifically, the ROPS method removes one randomly chosen variant carrier for each uncovered variant site. Therefore, if there are K variant sites uncovered in the sample, a total of K samples will be removed and N^A+N^U-K samples will be analyzed following the ROPS approach (Longmate *et al.*, 2010). For the GSO method, all individuals that are sequenced are removed from subsequent association analyses.

2.3 Generation of genetic and phenotypic data

We simulated genetic data using a four-parameter population genetic model (Adams and Hudson, 2004; Kryukov *et al.*, 2009). The model was estimated using sequence data from the Ottawa Obesity Study (Ahituv *et al.*, 2007), and incorporates both demographic change (i.e. bottleneck and exponential expansion) and purifying selection, which are believed to affect rare variant site frequency spectrums. Details for the population genetics model parameters can be found in (Kryukov *et al.*, 2009). Hardy–Weinberg equilibrium was assumed for the general population. Phenotype data were generated according to the following logistic regression model, i.e.

$$\log\left(\frac{\Pr\left(Y_i|\vec{X}_i\right)}{1-\Pr\left(Y_i|\vec{X}_i\right)}\right)=\beta_0+\sum\nolimits_{k\in C}\sum\nolimits_{j=1,2}\beta_k X_i^{k,j}$$

Under the alternative hypothesis, two types of phenotypic models were considered. In the first model, a certain proportion of rare variants sites C are randomly selected to be causal, and affect disease status. The power was investigated when 10, 70 and 90% of the variant sites are causal. In the second model, the causality of variants is determined by the selection coefficient. Power was evaluated, where variants with selection coefficients $>10^{-4}$, 10^{-3} and 10^{-2} are causal and affect the disease risk.

Under both models, it is assumed that each causal variant has an odds ratio of 3 and non-causal variants have an odds ratio of 1, i.e.

$$\beta_k=\begin{cases}\log(3) & k\in C\\ 0 & k\notin C.,\end{cases}$$

as suggested by Bodmer and Bonilla (2008). Under the null hypothesis, all $\beta_k's$ are set to be 0.

2.4 Evaluation of type-I errors and power

Type-I errors of rare variant association tests were evaluated under four different data integration strategies, i.e. (i) naïve method, which directly combines data without corrections,;ii) SEQCHIP, (iii) ROPS and (iv) GSO. Scenarios were considered where 500 cases/500 controls and 1500 cases/1500 controls were analyzed. For each case control dataset, we considered study designs where 10, 50 and 90% of the cases were sequenced to discover variants, and the identified rare variants (with MAF $< 1\%$) were followed-up and genotyped in the remaining samples. For the data integration strategies under which the rare variant association tests have controlled type-I errors, (i.e. SEQCHIP, ROPS and GSO), the power was also compared. One sided tests were performed, i.e. the alternative hypothesis that there is an increased number of rare causal alleles in cases is tested. For the ANRV method, statistical significance was obtained analytically, whereas the p-values for WSS and VT were obtained empirically using 2000 permutations. The power and type-I errors for different tests were evaluated using 10 000 replicates.

2.5 Evaluation of study designs

We also evaluated different two-stage study designs that sequence a portion of the sample to discover variants and genotype identified variants in the remaining samples. Specifically, by applying the SEQCHIP method and performing rare variant association testing using ANRV, WSS and VT, we compared the study design of sequencing only cases with that of sequencing an equal number of cases and controls, and combining sequence and genotype data via meta-analysis methods. p-values for ANRV were evaluated analytically, whereas that for WSS and VT were calculated using 2000 permutations. For meta-analysis, p-values were transformed to Z-score statistics, which were then weighted by the square root of the sample size and combined (Munafo and Flint, 2004). Power at each sequence sample size was obtained based upon 10 000 replicates. A significance level of $\alpha=0.05$ is used.

2.6 The analysis of sequence dataset of colorectal adenomas

In the colorectal adenoma dataset, five genes were first sequenced in cases for variant discovery and the identified variants were then genotyped in controls. In this article, we re-analyze the dataset using two valid data integration methods, i.e. ROPS and SEQCHIP. ANRV, WSS and VT tests were applied to detect associations with rare variants.

3 RESULTS

3.1 Evaluation of type-I errors

Type-I errors of the ANRV test were displayed in (Table 1) and that of the WSS and VT tests are shown in (Supplementary Table S1), for different strategies of integrating sequence and genotype data. In the naïve analyses where no corrections were made for the sequence data, type-I errors for all tests were highly inflated. In some scenarios, the type-I error can be $>30\%$ under a significance level of $\alpha = 0.05$. For example, when 90% of the cases were sequenced in a sample of 1500 cases and 1500 controls, the type-I errors for the ANRV, WSS and VT tests are, respectively, 30.3, 29.9 and 60.4%.

When sequence and genotype data were combined using SEQCHIP and ROPS, the type-I errors for rare variant association tests were controlled. However, the correction by ROPS can be overly conservative. For example, the type-I error for ANRV, WSS and VT are, respectively, 0.022 (with 95% CI [0.019, 0.025]), 0.015 (with 95% CI [0.013, 0.017]) and 0.017 (with 95% CI [0.014, 0.020]) when 50% cases were sequenced in a sample of 500 cases and 500 controls. For the same scenario, if the SEQCHIP method is used, the type-I error for the three tests are also conservative, but to a lesser extent (i.e. ANRV: 0.035 with 95% CI [0.031, 0.039], WSS: 0.036 with 95% CI [0.032, 0.040] and VT: 0.043 with 95% CI [0.039, 0.047]).

Table 1. Type-I error for rare variant association test ANRV

Sample size[a]	Percentage of cases sequenced	Type-I errors for ANRV[b]			
		Naïve	SEQCHIP	ROPS	GSO
1000	0.1	0.158	0.041	0.037	0.042
1000	0.5	0.217	0.035	0.022	0.048
1000	0.9	0.284	0.037	0.015	0.051
3000	0.1	0.149	0.043	0.037	0.043
3000	0.5	0.234	0.038	0 .021	0.049
3000	0.9	0.303	0.036	0.018	0.047

Scenarios were considered when sequence and genotype data were combined using the naïve method, SEQCHIP method and ROPS method, or when only genotype samples are analyzed.

[a]Total sample size with equal number of cases and controls.

[b]Type-I error was evaluated under a significance level of $\alpha = 0.05$. *p*-values for ANRV were obtained analytically.

3.2 Power comparisons

We compared the performance of different data integration methods, for which rare variant association tests have controlled type-I errors, i.e. SEQCHIP, ROPS and GSO. SEQCHIP method outperforms other methods in most scenarios. In some scenarios the power improvement can be as high as 30%. For example, when variant causality is determined by fitness, (i.e. variants with selection coefficients $>10^{-4}$ are causal with effects $\beta_k = \log(3)$, $k \in C$), if 90% of the cases were sequenced in a sample of 1500 cases and 1500 controls, the power for VT test is 76.5% (Fig. 1f). In the same scenario, if data are integrated by ROPS and GSO methods, the power for VT is only 55.6 and 34.1%, respectively.

ROPS methods can be conservative for correcting the genotypes of low frequencies variants. This is reflected by two observations: (i) when a small number of cases are sequenced and the data are integrated using ROPS, the power may be lower than when only genotype samples are analyzed (i.e. GSO method); (ii) when a larger portion cases are sequenced, the power for rare variant association tests may decrease. For example, when causality is assumed to be independent of fitness, (i.e.70% of the variants are causal with effect $\beta_k = \log(3)$, $k \in C$), if 90% of the cases are sequenced for a cohort of 1500 cases and 1500 controls and ROPS is used to integrate the data, the power for WSS is 61.6%, which is lower than the power (67.4%) when 50% of the cases are sequenced (Fig. 2e). This is because for the ROPS method, by taking out one variant carrier from the sequenced cases, a greater proportion of the carriers are removed from the sequenced cases than from the entire sample, and the variant frequencies may be slightly underestimated. Specifically, when N^S cases are sequenced to discover variants and N^G cases and N^U controls are genotyped, the variant frequencies in cases are estimated by $\hat{p}^A_{ROPS} = \left(M^S + M^G - 1\right)/\left(2N^S + 2N^G - 1\right)$, where M^S and M^G are the number of minor alleles observed in sequenced and genotyped cases for a given site. On the other hand, the variant frequencies in genotyped cases are estimated by $\hat{p}^A_{GSO} = M^G/N^G$. It is easy to verify numerically that when variant frequencies are low, $\hat{p}^A_{ROPS}$ can be smaller in value than $\hat{p}^A_{GSO}$. For example, when only one variant is observed in the sequenced cases for a given site, $\hat{p}^A_{ROPS} = M^G/\left(2N^S + 2N^G - 1\right)$, which is smaller than $\hat{p}^A_{GSO} = M^G/2N^G$. We also proved rigorously using probability theory (Supplementary Appendix SA), that when SEQCHIP or ROPS are used to integrate data, the variant frequencies can be slightly underestimated, i.e. $E\left(\hat{p}^A_{ROPS} \middle| M^S > 0\right) < E\left(\hat{p}^A_{GSO}\right)$. Therefore, although more causal variants may be uncovered by sequencing additional samples and sample size can be increased by integrating sequence samples, the accumulated downward biases may mitigate the variant frequencies between cases and controls, and reduce the power for rare variant association tests.

The GSO method only analyzes genotyped samples, and therefore can be vastly underpowered when a large portion of the cases are sequenced. For example, under the model assuming that variant causality is determined by fitness (i.e. variants with selection coefficients $>10^{-4}$ are causal with effects $\beta_k = \log(3)$, $k \in C$), for a

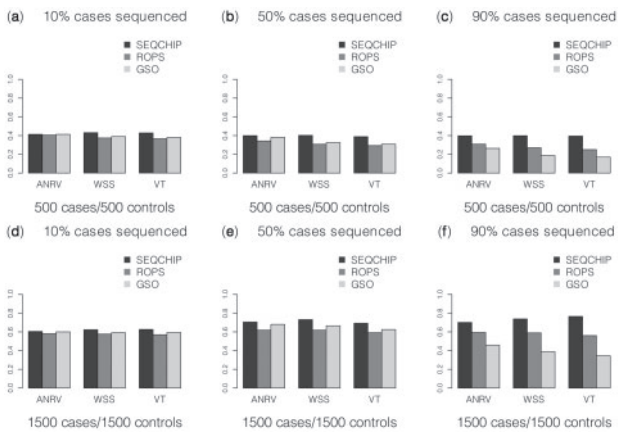

Fig. 1. Comparisons of SEQCHIP, ROPS and GSO. It is assumed that variants with selection coefficients $>10^{-4}$ are causal. ANRV, WSS and VT are used to analyze data. *p*-values for ANRV were obtained analytically, whereas the *p*-values for the other methods were obtained using 5000 permutations. The power was evaluated under a significance level of $\alpha = 0.05$ using 10 000 replicates

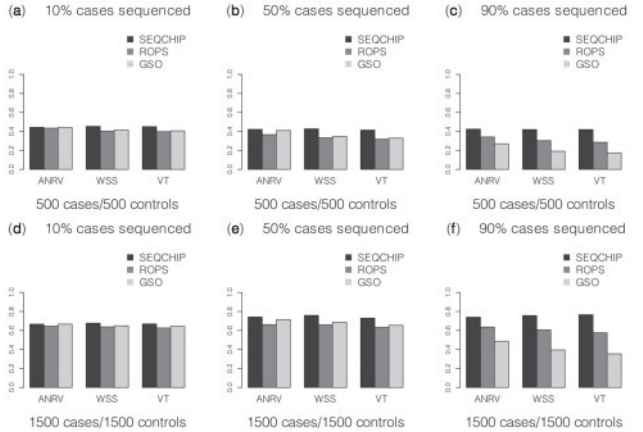

Fig. 2. Comparisons of SEQCHIP, ROPS and GSO. It is assumed that 70% of variants are randomly chosen to be causal. ANRV, WSS and VT are used to analyze data. *p*-values for ANRV were obtained analytically, whereas the *p*-values for the other methods were obtained using 5000 permutations. The power was evaluated under a significance level of $\alpha = 0.05$ using 10 000 replicates

sample of 1500 cases and 1500 controls, if 90% of the cases are sequenced for variant discoveries, the power of WSS is 38.2%. However, when sequence and genotype data are combined using SEQCHIP or ROPS, the power for WSS is, respectively, 73.5 and 58.2% (Fig. 1f). Due to the conservativeness in the estimation of allele frequencies, when a smaller proportion of cases are sequenced, e.g. 10 or 50%, the GSO method can have greater power than the ROPS method. For example, in a sample of 1500 cases and 1500 controls, when 50% of the cases are sequenced, the power for WSS is 61.7% if data are integrated using ROPS, and 66.2% if GSO method is used and only genotype data are analyzed (Fig. 1e).

The results of the power analyses for alternative selection coefficient cutoffs (i.e. 10^{-2} and 10^{-3}) can be found in Supplementary Figures S2 and S3. Additionally Supplementary Figures S4 and S5 display the results of the power analysis when 10 and 90% of the variants are randomly chosen to be causal. For these additional power comparisons, the relative performances of different data integration methods remain unchanged. Among the three rare variant association methods that were examined, there is not a single method that is consistently the most powerful. The advantage of different methods over each other is small. This is concordant with other existing reviews on rare variant association analyses methods (Basu and Pan, 2011; Ladouceur *et al.*, 2012).

3.3 Comparison of study designs

Figure 3 displays the power of the two study designs which sequence a portion of the samples to discover variants and genotype the identified variants in the remaining sample. It is demonstrated that when only a small number of samples are sequenced, sequencing only cases is more powerful than sequencing a balanced number of cases and controls. However, as the number of sequenced samples increases, the advantage of sequencing only cases diminishes because the data integration methods may be conservative for estimating rare variant frequencies. It is more powerful to sequence a balanced number of cases and controls to discover variants and then combine the sequence and genotype data by meta-analysis methods. The power comparisons remain similar when the ANRV (Supplementary Fig. S6) and WSS (Supplementary Fig. 7) tests are implemented.

As a comparison, we also calculated the power when all samples are sequenced for the study (Supplementary Tables S2 and S3). The results can be viewed as the 'maximal' achievable power if there are no budgetary constraints and the entire sample can be sequenced.

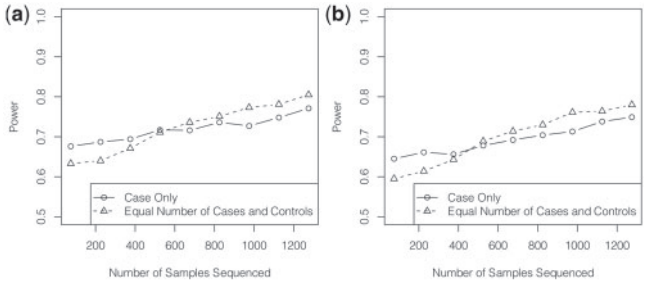

Fig. 3. Power comparison of two-stage study designs. Results are shown when only cases are sequenced for variant discovery (solid lines with circles) and when a balanced number of cases and controls (dashed lines with triangles) are sequenced. (**a**) displays results when 70% of the variants are randomly chosen to be causal, and (**b**) displays results where variants with selection coefficients $>10^{-4}$ are causal. Scenarios are examined when different numbers of samples are sequenced from 1500 cases and 1500 controls. When only cases are sequenced, SEQCHIP is used to correct for the bias. When an equal number of cases and controls are sequenced, sequence and genotype data are analyzed separately and meta-analysis is used to combine the results. For all scenarios VT is used to detect rare variant associations

3.4 Analysis of colorectal adenoma dataset

We jointly analyzed variants in five genes i.e. *APC*, *AXIN1*, *CTNNB1*, *hMLH1* and *hMSH2*. A total of 12 missense mutations were identified through sequencing 124 cases with colorectal cancer, and the identified variants were then genotyped in 483 controls. Three analyses were performed for the dataset.

First we combined sequence and genotype data with SEQCHIP method, and analyzed the resulting dataset with the ANRV, WSS and VT. One-sided tests were performed, which tests for the enrichment of rare variant alleles in colorectal adenomas patients. The *p*-values are, respectively, given by $p_{\text{ANRV}}^{\text{SEQCHIP}}=0.034$, $p_{\text{WSS}}^{\text{SEQCHIP}}=0.023$ and $p_{\text{VT}}^{\text{SEQCHIP}}=0.106$ (Table 2). The *p*-values for ANRV and WSS are significant. Second, we combined sequence and genotype dataset by the ROPS method. No significant results were observed (i.e. $p_{\text{ANRV}}^{\text{ROPS}}=0.080$, $p_{\text{WSS}}^{\text{ROPS}}=0.088$ and $p_{\text{VT}}^{\text{ROPS}}=0.144$).

Finally, for comparison purposes, the dataset was also analyzed under the naïve strategy, where sequence and genotype data are directly combined without corrections. The *p*-values are clearly biased, for all tests (i.e. $p_{\text{ANRV}}^{\text{naive}}=0.004$, $p_{\text{WSS}}^{\text{naive}}=0.005$ and $p_{\text{VT}}^{\text{naive}}=0.005$). This is concordant with our theoretical expectations and observations from simulated dataset. We also analyzed each gene individually. However, each gene alone contains too few rare variants and the analyses were not significant (data not shown).

4 DISCUSSION

In this article, we developed a data integration method for two-stage case control studies where a portion of cases are sequenced to discover variants, and the identified variants are genotyped in the remaining sample. The SEQCHIP method performs a correction on the variant genotypes observed in sequenced cases, such that the corrected sequence genotypes follow approximately the same distribution as that of the genotyped samples. The integrated dataset can be analyzed by all existing rare variant association tests that can handle genotypes with uncertainties (e.g. imputed genotypes). SEQCHIP can also be used with regression-based methods for detecting primary or secondary traits associations (Lin and Zeng,

Table 2. The analysis of colorectal adenoma dataset

Corrections	*p*-values for rare variant association tests		
	ANRV[a]	WSS[b]	VT[b]
SEQCHIP	0.034	0.023	0.106
ROPS	0.080	0.088	0.144
Naïve	0.004	0.005	0.005

[a]*p*-values for ANRV test were obtained analytically.
[b]*p*-Values for WSS and VT tests were obtained by 10 000 permutations.

2009; Liu and Leal, 2011), where confounders such as population substructures can be controlled. Through extensive simulations, we demonstrate that when SEQCHIP is used to integrate sequence and genotype data, all rare variant association tests have controlled type-I errors. The power can be substantially improved compared with using other data integration strategies, i.e. ROPS and GSO.

The method is mainly developed for combining sequence and genotype data when only cases were sequenced for variant discoveries. A popular alternative two-stage study design is to sequence a combination of selected cases and controls for variant discovery, and genotype the identified variants in the rest of the samples. Under this study design, sequenced and genotyped samples can be separately analyzed and combined using standard meta-analyses methods, which will have controlled type-I error rates. When both cases and controls are sequenced, protective variants for the disease phenotype may be uncovered with higher probability (Rivas *et al.*, 2011). For a two-stage study that combines sequence and genotype data, given a small fixed number of samples that are sequenced, sequencing only cases can be more powerful for detecting causal variant associations than sequencing a balanced number of cases and controls.

In practice, it is of interest to know the optimal fraction of cases to sequence to maximize power. Although sequencing a larger number of samples allows discovering a higher number of variant sites, it does not necessarily lead to improved power. This is because the frequencies of very rare variants identified in sequenced cases can be slightly underestimated by ROPS and SEQCHIP methods, which reduces power. The optimal number depends on the underlying disease model, the size of the cohort, and the proportion of the cases that are sequenced, which will need to be examined on a case by case basis.

Although the cost of sequencing is quickly dropping, genotyping still has a clear cost advantage. The two-stage study design of sequencing cases and genotyping the remaining sample allows extracting genetic information from a much larger number of samples, which can be more powerful than one stage study design given a fixed cost constraint. SEQCHIP is a very useful method for integrating data in a two-stage study design and will greatly accelerate the process of identifying variants involved in complex trait etiologies.

5 ACKNOWLEDGEMENTS

The authors thank Dr Shamil Sunyaev for sharing the simulated genetic datasets; and also thank Dr Walter Bodmer and Ms. Nicola Fearnhead for sharing the colorectal adenomas dataset.

Funding: National Institutes of Health [HL102926, MD005964, HG006493 to S.M.L. and MH084676 to Shamil Sunyaev]; Cancer Research, UK.

Conflict of Interest: none declared.

REFERENCES

Adams,A.M. and Hudson,R.R. (2004) Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, **168**, 1699–1712.

Ahituv,N. *et al.* (2007) Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.*, **80**, 779–791.

Basu,S. and Pan,W. (2011) Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.*, **35**, 606–619.

Bhatia,G. *et al.* (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.*, **6**, e1000954.

Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.

Cohen,J.C. *et al.* (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.

Cohen,J.C. *et al.* (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl Acad. Sci. USA*, **103**, 1810–1815.

Edwards,T.L. *et al.* (2010) Enriching targeted sequencing experiments for rare disease alleles. *Bioinformatics*, **27**, 2112–2118.

Fearnhead,N.S. *et al.* (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA*, **101**, 15992–15997.

Frayling,I.M. *et al.* (1998) The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc. Natl Acad. Sci. USA*, **95**, 10722–10727.

Guan,Y. and Stephens,M. (2008) Practical issues in imputation-based association mapping. *PLoS Genet.*, **4**, e1000279.

Han,F. and Pan,W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.*, **70**, 42–54.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Ji,W. *et al.* (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.

Kim,H.C. *et al.* (2000) The E-cadherin gene (CDH1) variants T340A and L599V in gastric and colorectal cancer patients in Korea. *Gut*, **47**, 262–267.

Kryukov,G.V. *et al.* (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727–739.

Kryukov,G.V. *et al.* (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl Acad. Sci. USA*, **106**, 3871–3876.

Ladouceur,M. *et al.* (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet.*, **8**, e1002496.

Lamlum,H. *et al.* (2000) Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. *Hum. Mol. Genet.*, **9**, 2215–2221.

Li,B. and Leal,S.M. (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.*, **5**, e1000481.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Lin,D.Y. and Zeng,D. (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genet. Epidemiol.*, **33**, 256–265.

Lipton,L. *et al.* (2003) Germline mutations in the TGF-beta and Wnt signalling pathways are a rare cause of the "multiple" adenoma phenotype. *J. Med. Genet.*, **40**, e35.

Liu,D.J. and Leal,S.M. (2010a) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, e1001156.

Liu,D.J. and Leal,S.M. (2010b) Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am. J. Hum. Genet.*, **87**, 790–801.

Liu,D.J. and Leal,S.M. (2011) A flexible likelihood framework for detecting associations with secondary phenotypes in genetic studies using selected samples: application to sequence data. *Eur. J. Hum. Genet.*, **20**, 449–456.

Longmate,J.A. *et al.* (2010) Three ways of combining genotyping and resequencing in case-control association studies. *PLoS ONE*, **5**, e14318.

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

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.

Munafo,M.R. and Flint,J. (2004) Meta-analysis of genetic association studies. *Trends Genet.*, **20**, 439–444.

Neale,B.M. *et al.* (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.

Price,A.L. *et al.* (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Raychaudhuri,S. *et al.* (2011) A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat. Genet.*, **43**, 1232–1236.

Rivas,M.A. *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066–1073.

Romeo,S. *et al.* (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.

Romeo,S. *et al.* (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.*, **119**, 70–79.

Sanna,S. *et al.* (2011) Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.*, **7**, e1002198.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Zawistowski,M. *et al.* (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.*, **87**, 604–617.

Zheng,J. *et al.* (2011) A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet. Epidemiol.*, **35**, 102–110.