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; Ji 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 (Fearon 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 naive 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
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; Imita-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 the 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 that of the genotyped samples. The method can be used with any existing rare variant association tests that can analyze uncertain that of the genotyped samples. The method can be used with any existing rare variant association tests that can analyze uncertain that of the genotyped samples. The method can be used with any existing rare variant association tests that can analyze uncertain that of the genotyped samples.

2 METHODS

2.1 SEQCHIP method

We assume that there are $N^r$ affected individuals and $N^u$ unaffected individuals in the sample. Among the affected individuals, $N^r$ (i.e. individuals 1, 2, …, $N^r$) are sequenced to discover variants. An additional $N^m = N^r - N^u$ cases (i.e. individuals $N^r + 1, \ldots, N^r + N^m$) and $N^m$ 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.

$$X_i = (1, 1, \frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}, \frac{1}{2}),$$

where each entry $k$ represents a site with di-allelic single-nucleotide variations, and $x^s_i$ 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.

$$P_r \left( \sum_{j=1}^{k} x^s_j + 1 \leq n \right) = \sum_{x_k=x_1}^{n} \binom{x_k}{x_1} \left( \frac{1}{2} \right)^{x_k-n},$$

where $p_r$ 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

$$E \left( \sum_{j=1}^{k} x^s_j + 1 \right) = \frac{2N^r p_r^m}{1 - (1 - p_r^m)^{N^r}}.$$}

Therefore, a naive 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 $c_k = 1 - (1 - p_r^m)^{N^r}$.

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.

$$\tilde{X}_i = \left( \tilde{x}_1^s, \tilde{x}_2^s, \ldots, \tilde{x}_K^s \right),$$

where the marginal distribution for each entry $\tilde{x}_k^s$ satisfies

$$P_r \left( \tilde{x}_k^s = 0 \right) = \frac{1 - (1 - p_r^m)^{N^r}}{1 - (1 - p_r^m)^{N^r}} \times \delta \left( x_k^s = 0 \right),$$

$$P_r \left( \tilde{x}_k^s = 1 \right) = \frac{1}{1 - (1 - p_r^m)^{N^r}} \times \delta \left( x_k^s = 1 \right),$$

$$P_r \left( \tilde{x}_k^s = 1 \right) = \frac{1}{1 - (1 - p_r^m)^{N^r}} \times \delta \left( x_k^s = 0 \right).$$

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 Q 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 \((X, Y)\), we considered a corrected dataset \(\{\tilde{X}, \tilde{Y}\}\) with sequence genotypes, where \(\tilde{X} = X + E\) and \(\tilde{Y} = Y\). The corrected genotype dataset is obtained based upon 10 000 replicates. A significance level of \(\alpha = 0.05\) was used.

### 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 \(N_N = N - K\) 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., 2010). Hardy–Weinberg equilibrium was assumed for the general population. Phenotype data were generated according to the following logistic regression model, i.e.

\[
\log \left( \frac{p(Y_i|\tilde{X}_i)}{1 - p(Y_i|\tilde{X}_i)} \right) = \beta_0 + \sum_{c \in C} \sum_{k=1}^{N_N} \rho_k \tilde{X}_i^k
\]

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 \(>10^{-4}, 10^{-5}\) and \(10^{-7}\) 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.

\[
\beta_0 = \log(3), \quad \beta_k = \begin{cases} 
\log(3) & k \in C \\
0 & k \notin C
\end{cases}
\]

as suggested by Bodmer and Bonilla (2008). Under the null hypothesis, all \(\beta_k^C\) 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 calculated 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 sample. 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 [0.03, 0.04] and VT: 0.043 with 95% CI [0.039, 0.047]).
We compared the performance of different data integration methods, portion cases are sequenced, the power for rare variant association genotype samples are analyzed (i.e. GSO method); (ii) when a larger scenario, if data are integrated by ROPS and GSO methods, the 1500 controls, the power for VT test is 76.5% (Fig. 1f). In the same scenarios, if data are integrated by ROPS and GSO methods, the variant frequencies 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 = 30%). For example, when variant = 10^-4 are causal with effects βk = log (k ∈ 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 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 βk = log (k ∈ 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 N5 cases are sequenced to discover variants and N5 cases and N1 controls are genotyped, the variant frequencies in cases are estimated by pROPS = (M ^ S + M ^ G − 1) / (2N ^ S + 2N ^ G − 1), 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 pGSO = M ^ G / N ^ G. It is easy to verify numerically that when variant frequencies are low, pROPS can be smaller in value than pGSO. For example, when only one variant is observed in the sequenced cases for a given site, pROPS = M ^ G / (2N ^ S + 2N ^ G − 1), which is smaller than pGSO. 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 (pROPS | M ^ S > 0) < E (pGSO). 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 βk = log (k ∈ C), for a
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 analyzed separately using SECHIP 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. $10^{-4}$ and $10^{-5}$) 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. Among 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 methods (Basu and Pan, 2011; Lodoe ucz 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, WSS (Supplementary Fig. S0) and VT (Supplementary Fig. S1) tests are implemented.

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

### 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 SECHIP 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 $p_{\text{ANRV}} = 0.034$, $p_{\text{WSS}} = 0.023$ and $p_{\text{VT}} = 0.106$ (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. $p_{\text{ANRV}} = 0.080$, $p_{\text{WSS}} = 0.088$ and $p_{\text{VT}} = 0.144$). Finally, for comparison purposes, the dataset was also analyzed under the naive 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 SECHIP 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). SECHIP can also be used with regression-based methods for detecting primary or secondary traits associations (Lin and Zeng, 2012).

---

**Table 2. The analysis of colorectal adenoma dataset**

<table>
<thead>
<tr>
<th>Corrections</th>
<th>$p$-values for rare variant association tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ANRV$^a$</td>
</tr>
<tr>
<td>SECHIP</td>
<td>0.034</td>
</tr>
<tr>
<td>ROPS</td>
<td>0.080</td>
</tr>
<tr>
<td>Naïve</td>
<td>0.004</td>
</tr>
</tbody>
</table>

$^a$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 SEQCHAMP 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, sequence 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 SEQCHAMP 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. SEQCHAMP 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 Fearhead 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


