frontiers in
GENETICS

REVIEW ARTICLE
published: 13 December 2013
doi: 10.3389/fgene.2013.00276

Two-phase and family-based designs for next-generation sequencing studies

Duncan C. Thomas*, Zhao Yang and Fan Yang

Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

Edited by:
Xuefeng Wang, Harvard University, USA

Reviewed by:
Jigang Zhang, Tulane University, USA
William C. L. Stewart, Columbia University, USA

****Correspondence:***
Duncan C. Thomas, Department of Preventive Medicine, University of Southern California, 1450 Biggy Street, NRT 2502, Los Angeles, CA 90089-9601, USA
e-mail: pdthomas@usc.edu

The cost of next-generation sequencing is now approaching that of early GWAS panels, but is still out of reach for large epidemiologic studies and the millions of rare variants expected poses challenges for distinguishing causal from non-causal variants. We review two types of designs for sequencing studies: two-phase designs for targeted follow-up of genomewide association studies using unrelated individuals; and family-based designs exploiting co-segregation for prioritizing variants and genes. Two-phase designs subsample subjects for sequencing from a larger case-control study jointly on the basis of their disease and carrier status; the discovered variants are then tested for association in the parent study. The analysis combines the full sequence data from the substudy with the more limited SNP data from the main study. We discuss various methods for selecting this subset of variants and describe the expected yield of true positive associations in the context of an on-going study of second breast cancers following radiotherapy. While the sharing of variants within families means that family-based designs are less efficient for discovery than sequencing unrelated individuals, the ability to exploit co-segregation of variants with disease within families helps distinguish causal from non-causal ones. Furthermore, by enriching for family history, the yield of causal variants can be improved and use of identity-by-descent information improves imputation of genotypes for other family members. We compare the relative efficiency of these designs with those using unrelated individuals for discovering and prioritizing variants or genes for testing association in larger studies. While associations can be tested with single variants, power is low for rare ones. Recent generalizations of burden or kernel tests for gene-level associations to family-based data are appealing. These approaches are illustrated in the context of a family-based study of colorectal cancer.

Keywords: sequencing, two-phase sampling design, family-based study, rare variant association, breast neoplasms, colorectal cancer

INTRODUCTION

In the early days of genomewide association studies (GWAS), the cost of commercial high-density genotyping panels was prohibitive for large-scale epidemiologic studies needed to detect the modest relative risks (RRs) now known to be associated with most common variants for complex diseases (Hindorff et al., 2009). Hence, investigators turned to multi-stage designs, in which only a sample of subjects were genotyped on such platforms and a generous selection of the most significant associations were then tested on an independent sample using custom genotyping techniques. The final analysis typically combined the data from both stages, with a final significance level chosen to ensure genomewide significance after allowing for the number of variants tested in the second. The basic principles were based on a series of papers written before the GWAS era (Satagopan et al., 2002, 2004; Satagopan and Elston, 2003), and subsequent work showed how to optimize the allocation of sample size and first-stage critical values in the GWAS context (Wang et al., 2006; Skol et al., 2007). In particular, Skol et al. (2006) showed that this joint analysis was more powerful than treating the design as discovery followed by independent replication, despite various high-profile journals' requirements for an independent replication study (Panagiotou et al., 2012). Although this became the conventional GWAS design throughout the first decade of the 21st century, rapidly declining costs of commercial GWAS chips have made it feasible for many studies to obtain genome-wide coverage on *all* available subjects in a single stage (Thomas et al., 2009a,b). The cost of custom genotyping for large numbers of hand-picked SNPs was often comparable to standard high-density panels, and having more subjects with genome-wide data allowed for more informative analysis of interactions, subgroups, pleiotropic effects, etc. For a general review of multi-stage designs in genetics, see (Elston et al., 2007).

As we entered the "post-GWAS" era, the focus began to shift toward rare variants and the use of next-generation sequencing (NGS) technologies that could in principle (given a large enough sample size and deep enough sequencing) uncover *all* the genetic variation in a region, not just the common SNPs that have been used to tag the unknown causal variants. In part, this interest stemmed from the increasing recognition that common variants were accounting for only a relatively small proportion of the total heritability of most complex diseases (Manolio et al., 2009; Schork et al., 2009). Amongst other possible explanations

for the "missing heritability," rare variants have been proposed, based on an evolutionary argument (Gorlov et al., 2011) or empirical evidence (Bodmer and Bonilla, 2008) that their effect sizes could be larger, although recent whole-exome sequencing studies have cast some doubt on this hypothesis (e.g., Heinzen et al., 2012). Furthermore, since rare variants tend not to be well tagged by common ones (Duan et al., 2013), use of conventional GWAS panels would tend to miss associations with rare variants. Cost currently precludes application of NGS to whole-genome sequencing on a large scale, so clever study design has again become important (Thomas et al., 2009a,b). One of the first uses of NGS was for targeted follow-up of GWAS hits, for which an alternative to two-stage designs, known as two-*phase* designs, is a natural choice. These differ from the two-*stage* designs described above in that the set of subjects chosen for expensive data collection (e.g., NGS) are a *proper subset* of a larger epidemiologic study rather than an independent sample and that this subset is selected on the basis of information already available on the full study (Whittemore and Halpern, 1997; Thomas et al., 2004; Yang and Thomas, 2011). In the case of NGS, this could involve stratification *jointly* on disease status and carrier status of the associated variant(s). While this would tend to induce a spurious association between any variants in LD with the GWAS SNPs and disease even under the null hypothesis that they are not causal, this bias can be avoided by adjusting for the sampling fractions, and additional information available in the full study can also be incorporated. The basic principles were developed in a series of seminal papers by Norman Breslow with various colleagues (see Breslow and Holubkov, 1997b; Breslow and Chatterjee, 1999; Scott et al., 2007; Breslow et al., 2009b, for summaries of this work). Recently, Schaid et al. (2013a) has provided an excellent discussion of the use of this approach for targeted follow-up of GWAS hits by NGS. However, for whole genome or whole exome sequencing studies, there would be no point in selecting individuals based on whether they carried a specific polymorphism, except to eliminate those known to be carrying a known major mutation.

Most GWAS for discovering common variants associated with disease traits have been conducted using a case-control design with unrelated controls. Not only are unrelated individuals easier to identify and enroll than are entire families (particularly multiple-case families), but the statistical efficiency for discovery or association testing per subject genotyped is typically higher using unrelated controls than using unaffected siblings or other relatives (Witte et al., 1999). However, with the growing interest in rare variants and the availability of NGS, there has been a resurgence of interest in using family-based designs (Zhu et al., 2010; Feng et al., 2011; Ionita-Laza and Ottman, 2011; Shi and Rao, 2011). Family-based designs may have other advantages that outweigh their loss of statistical efficiency. By exploiting information about co-segregation, they may be more efficient at prioritizing potentially causal variants from non-causal ones for subsequent testing for association with disease in larger samples. The ability to exploit Mendelian inheritance may also improve the imputation of rare variants in untested samples (Li et al., 2009; Cheung et al., 2013). Finally, family-based designs can exploit both between- and within-family comparisons in a two-*step* analysis for better power while being robust to bias from population stratification (Lange et al., 2003; Van Steen et al., 2005; Feng et al., 2007; Murphy et al., 2008). In this paper, we focus on the first of these advantages, using a design that sequences a subset of family members initially, ranks the discovered variants in terms of their likelihood of being associated with the trait using the phenotype information on the entire family, and then tests for association in an independent sample. In this sense, the design has elements of both two-*phase* and two-*stage* designs, in that the sequencing set is a proper subset of a larger family-based study and that an independent sample is used for replication or combined analysis.

One consequence of the new focus on rare variants is the need for novel analysis strategies, because testing associations individually with every variant would have very little power due to the large multiple comparisons penalty and their rarity. In a sample of, say, size 200, one might identify about 20 million variants. Most of these are likely to be unrelated to disease and genotyping all of them for a large case-control association study would be neither feasible nor statistically efficient, so some means of identifying those most likely to be causal is needed. Furthermore, under some models of disease causation, multiple variants in a causal gene (or pathway) could affect its function, so aggregating variants within genes may also improve power. To address this need, a host of "burden" tests have been developed based on counts of rare variants, weighted in various fashions (see Asimit and Zeggini, 2010; Cirulli and Goldstein, 2010; Basu and Pan, 2011; Bacanu et al., 2012; Thomas, 2012, for reviews). However, these are ill-suited to the situation where a region contains both deleterious and protective variants (Hoffmann et al., 2010). A random effects model that focuses instead on the *variance* of risk across variants rather than their mean might therefore be more powerful. The first of this type was the C_α test (Neyman and Scott, 1966; Neale et al., 2011), which tests for overdispersion of case-control ratios, conditional on the total number of variants. The Sequence Kernel Association Test [SKAT (Wu et al., 2011; Lee et al., 2012)], based on a general linear mixed model, tests for association between similarity of phenotypes and similarity of multi-locus genotypes across all pairs of subjects. See Schaid (2010a,b) for a general review of the basic statistical foundations of such tests and various choices of kernel functions for genetic applications. Recently, this class of methods has been extended to family studies (Huang et al., 2010; Schifano et al., 2012; Chen et al., 2013; Ionita-Laza et al., 2013; Schaid et al., 2013b). Hierarchical modeling approaches offer another approach to the analysis of rare variants, allowing formal incorporation of external information for prioritization.

A variety of methods for incorporating genomic context, functional, or pathway annotation data have been discussed in GWAS contexts (reviewed by Cantor et al., 2010; Thompson et al., 2013). Examples of prior information might include loci previously reported, pathway or genomic annotation, expression QTL or other functional assays, etc. (Rebbeck et al., 2004; Bush et al., 2009; Karchin, 2009; Nicolae et al., 2010; Wang et al., 2010; Freedman et al., 2011; San Lucas et al., 2012; Minelli et al., 2013). Filtering on such variables has become a popular strategy, but risks eliminating many causal variants whose potential significance has not yet been recognized or loading up the list of prioritized variants with too many non-causal ones based

on irrelevant information. The weighted False Discovery Rate (Roeder et al., 2006; Wakefield, 2007; Whittemore, 2007) and Gene Set Enrichment Analysis (Chasman, 2008; Holden et al., 2008) require specification of weights in advance and there is no obvious way to combine multiple filters. The hierarchical modeling approach described below is more flexible, allowing the weights given to various biofeatures to be determined empirically, based on their observed correlation with disease associations across the ensemble of all variants.

An example of a two-phase design is the Women's Environmental Cancer and Radiation Epidemiology (WECARE) Study of the risk of second breast cancers among survivors of a first breast cancer, focusing on radiation dose to the contralateral breast (Stovall et al., 2008; Langholz et al., 2009), various genes involved in DNA damage response pathways (Begg et al., 2008; Concannon et al., 2008; Borg et al., 2010; Malone et al., 2010; Capanu et al., 2011; Quintana et al., 2011; Brooks et al., 2012; Quintana et al., 2012; Reiner et al., 2013) and their interactions (Bernstein et al., 2010, 2013); a GWAS is also currently in progress. The design is a nested case-control study, with two controls matched to each case on age and year of diagnosis of the first cancer and study center, and "counter-matched" on radiotherapy for treatment of the first cancer (Bernstein et al., 2004). As an illustration of the two-phase design, we are currently performing whole genome sequencing on a subsample of 201 subjects and whole exome sequencing on several hundred more, drawn from the 701 cases and 1399 controls, stratified jointly by case-control status and risk predictors—age at first cancer, family history (FH), radiation treatment, and time since exposure.

As an example of a family-based design, we are currently performing deep targeted resequencing of 11 replicated regions identified by previous GWASs as associated with colorectal cancer (CRC), using ∼4200 samples drawn from the Colon Cancer Family Registries (C-CFR). The C-CFR is an international collaboration of registries of families ascertained through CRC in various ways, some population-based, some from high-risk genetic clinics, some including population controls or control families (Newcomb et al., 2007). To date, 10,662 CRC families have been enrolled, totaling 62,353 individuals, with genetic samples available on 5113 cases and 9196 unaffected family members or population controls with epidemiologic risk factor information, and FH data on many more (http://epi.grants.cancer.gov/(CFR/about_colon.html, accessed 3/8/13). For the purpose of comparing different designs, we have selected some samples from multiple-case families and some from unrelated cases or controls in various ways. Ultimately these data would be used to compare designs empirically in terms of the yield of significant findings by subsampling from these real data (e.g., to assess whether a lower depth of sequencing, narrower regions, fewer subjects or subjects targeted in different ways would have sufficed).

The aim of this paper is to review recent developments in methods for the design and analysis of NGS studies, with a particular focus on two-phase and family-based designs, and to illustrate the various issues with simulated data and applications to power calculations and preliminary data from these two studies.

RESULTS

TWO-PHASE SAMPLING FOR TARGETED RESEQUENCING

Suppose one has already completed a large case-control GWAS of unrelated individuals, in which one or more tag SNPs have been found to be strongly associated with a particular disease. It is unlikely that the associated SNPs would themselves be causal—more likely they are simply in LD with the truly causal variant(s). The aim of a targeted resequencing study is therefore to exhaustively re-sequence the region to identify *all* variants in the hopes of discovering these causal ones. Two key decisions are required: how to select the subsample to be sequenced; and how to prioritize the variants found in this subsample.

Approaches to prioritization of variants

One obvious method of prioritization would be on the basis of novelty, i.e., to focus attention on variants that have not been seen previously (or only rarely) among population controls. With the growing catalog of sequence variants in public databases like the 1000 Genomes Project, many causal variants are likely already to have been discovered and those that are novel are likely to be so rare that there would be very little power testing their association individually with disease. Furthermore, most novel variants are likely to be neutral. Nevertheless, the discovery of a novel association with disease is important, irrespective of whether or not the existence of the variant has been previously reported, but a discovery of a novel variant and its association with disease is particularly noteworthy. The same applies to filtering based on differences in allele frequencies between cases and controls within the sequencing subset (Yang and Thomas, 2011). Thus, some investigators have decided not to sequence controls, but this could be ill advised if cases are sequenced in populations not well represented in public databases or on platforms with different discovery characteristics (e.g., depth of coverage, quality control filtering).

Under the hypothesis that a gene may harbor multiple variants any of which could affect function or that a critical pathway could be affected by polymorphism in any of the genes in it, a strategy that aggregates across multiple related variants may be helpful. Methods section Simulation of Gene- and Pathway-level Prioritization in the WECARE STUDY describes a simulation of this strategy based on the WECARE study; results are discussed in the application below.

Hierarchical modeling (Greenland, 2000) entails adding a second-level model for the effect estimates of each variant, allowing the magnitude of the effects, their probability of being non-null, or their covariances to depend upon external information ("prior covariates") (Conti and Witte, 2003; Hung et al., 2004; Chen and Witte, 2007; Hung et al., 2007; Lewinger et al., 2007; Conti et al., 2009; Hoffmann et al., 2010; Capanu and Begg, 2011; Capanu et al., 2011). Hierarchical models involve a first (subject)-level model for individual's phenotypes Y_i as a function of a vector of genotypes $G_{\mathrm{i}} = (G_{\mathrm{iv}})$ at loci v and corresponding regression coefficients $\beta = (\beta_{\mathrm{v}})$, e.g., a general linear model of the form $f[E(Y_i)] = G_i'\beta$, and a second (variant)-level model for the distribution of these regression coefficients as a function of prior covariates Z_v, e.g., a linear regression model of the form $E(\beta_v) = Z_v'\pi$ (or) perhaps a model for their variances or

covariances (Thomas et al., 2009a,b). This approach also has the advantage of allowing for the uncertainty about which effects should be included in the model within a Bayes model averaging framework, e.g., by modeling the *probability* that a variant has no effect as $\text{logit}[\Pr(\beta_v = 0)] = Z_v'\alpha$ (Quintana et al., 2012; Quintana and Conti, 2013). It also allows the data to determine the optimal weights for the various prior covariates rather than having to specify them *a priori*; these papers show how the gain in power from including covariates with high predictive value for classifying causality of variants is offset by very little loss of power from including covariates with low predictive value (since they are usually assigned very little weight), in contrast to filtering approaches, which can lead to substantial loss of power if either sensitivity or specificity is low.

Joint analysis of SNP and sequence data

The subset of subjects from the substudy with the full sequence data would probably not provide adequate power for testing associations with disease directly, either variant-by-variant or by any of the aggregation methods described above. How then might one take advantage of the data from the much larger GWAS from which the substudy subjects were selected? Three basic strategies are possible: (i) by genotyping; (ii) by imputation; or (iii) by joint analysis.

The first of these is the simplest, but most expensive. One simply does custom genotyping in the main study of the prioritized variants. Under the hypothesis that any causal variants are likely to have been discovered by sequencing and that they survived prioritization, then the genotype data for the main study should be sufficient and associations can be tested directly, with appropriate correction for the effective number of independent tests performed (Conneely and Boehnke, 2007). A final analysis can then include a test of whether the novel variants account for the original SNP association (Yang and Thomas, 2011).

Imputation has become a standard approach for GWAS analysis, so that typically several million common variant associations are tested by combining the study data from the SNP panel with population distributions of all common and uncommon variants from such databases as the HapMap and 1000 Genomes projects (Asimit and Zeggini, 2012; Howie et al., 2012). For each variant not on the GWAS panel, one computes the expected allelic dosage and uses this as the covariate in a logistic regression model; this strategy is known to be superior to simply using the most likely genotype, in part because it correctly allows for the uncertainty in the imputation (Stram et al., 2003). Whether this strategy would be viable for rare variants is still unknown, but there are two reasons for concern. First, the strategy relies on linkage disequilibrium, and rare variants tend to have weaker LD than common ones (Duan et al., 2013). Second, it also relies on having sufficiently large reference panels, which would not include newly discovered variants.

Joint analysis of the full sequence data on the subsample and the SNP data on the main study is the most powerful approach and, like imputation does not involve any further genotyping costs. In their series of seminal papers on two-phase studies, Breslow et al. describe three basic analysis approaches: pseudo-likelihood (PL), weighted likelihood (WL), and semi-parametric likelihood (Breslow and Cain, 1988; Breslow and Zhao, 1988; Cain and Breslow, 1988; Breslow and Holubkov, 1997a,b; Breslow and Chatterjee, 1999; Breslow et al., 2003, 2009a,b; Breslow and Wellner, 2007). The simplest of these is the WL approach, so for simplicity, we confine our discussion here to this one (Methods section Likelihoods for Joint Analysis of Two-phase Studies). The basic idea is based on Horvitz-Thomson estimating equations, which use the score function derived from the likelihood for a logistic regression of disease status in the substudy data alone, weighting each subject's contribution inversely by their sampling probabilities, $\Sigma_i\,[Y_i - p_i(\beta)]\;W_i\;G_i = 0$, where $p_i(\beta) = \text{expit}(G_i'\beta)$ and $W_i = N_{si}/n_{si}$, s_i being the sampling stratum to which subject i belongs and N_s and n_s the main study and substudy sample sizes respectively. While simple in concept, the disadvantage is that the only information used from the main study is the stratum sample sizes. A refinement of this approach is to replace the empiric weights based on the realized sample sizes by predicted weights based on a logistic regression of sampling probabilities on additional covariates available for all main study participants. Recent papers show how this basic approach can be stabilized by using "calibrated weights" without requiring assumptions about the validity of an imputation model using influence residuals (Breslow et al., 2009a,b). The utility of this approach for targeted follow-up of GWAS hits is discussed in Schaid et al. (2013a).

Optimization of sampling fractions

As with two-stage designs, it is theoretically possible to optimize the choice of sampling fractions, subject to a constraint on total cost, but in practice this requires knowledge of the true values of various model parameters (causal allele frequencies, RRs, LD with the GWAS SNPs, etc.). Fortunately, the design is often relatively insensitive to these parameters, so that a balanced design in which the various strata are represented by equal numbers n_s in the subsample may be nearly optimal (Reilly and Pepe, 1995; Reilly, 1996). In their article on the application of this design to targeted follow-up of GWAS hits, for example, Schaid et al. (2013a) do not address optimization, but recommend the balanced design.

As with two-stage designs, the basic idea is either to maximize power subject to a constraint on total cost or to minimize the cost required to attain a target power. If the only cost is sequencing the subsample, then it is sufficient to optimize the *proportional* allocation of substudy subjects across strata. If instead one is designing both the main study and substudy *de novo* or if custom follow-up genotyping of the main study is planned, then the *relative* sample sizes of the two phases also need to be considered. In either case, there are likely to be multiple hypotheses being tested, so optimization of power for a specific type of variant may be less helpful than a global optimization. For this purpose, we have previously considered Asymptotic Relative Cost Efficiency, a quantity inversely proportional to the total cost times the variance of the parameter of interest, combining main and substudy data (Thomas, 2007), but more recently in the context of designs for sequencing using DNA pooling, we aimed to optimize power subject to a constraint on total cost (Liang et al., 2012). We adopt a similar approach to optimize designs for testing the 1 degree of freedom Madsen and Browning (2009) rare

variant burden test (Methods section Optimization of Two-phase Studies), but this could be easily extended to maximize power for multi-dimensional hypotheses. Here, we summarize a small simulation study to illustrate the potential of two-phase designs (Methods section Simulation of Two-phase Designs).

Results using the simulated sequence data are shown in **Table 1** for the full cohort (the "ideal" results if the entire sample could have been sequenced) compared with two-phase analyses using (1) imputation, (2) the Horvitz-Thompson WL approach with sample weights (Horvitz and Thompson, 1952); (3) the Breslow-Cain PL (Breslow and Cain, 1988); and (4) the Breslow-Holubkov semi-parametric estimator (Breslow and Holubkov, 1997a,b). The top half of the table provides results for the Madsen-Browning index including all variants present in the parent case-control study, while the bottom half is limited to those seen at least once in the subsample; for the latter, the risk index would include different variants for the different designs, so point estimates are not comparable. The last line gives the average estimate, empirical standard deviation of estimates across replicates, and the estimated non-centrality parameter (NCP) for the Wald test if no subsampling were done. Generally, the imputation approach was the least efficient for all designs, followed closely by WL estimator, while the semi-parametric one the most efficient. Except for the WL, the optimal sampling design was also the most efficient; the inefficiency of the WL in this case seems to be due to some small strata receiving very large weights (for example, 500/2 in the low-risk case stratum compared with 500/214 in the high-risk case stratum, see Footnote b of **Table 1**). In earlier simulations (not shown) we found relatively little inflation in the variance estimates or changes in the point estimates as the number of strata increases (although the number of replicates that failed to converge increased). Further research in the case of many sparse strata would be helpful, as well as on such issues as the size of the region to be sequenced and the depth of sequencing.

Application to the WECARE study of contralateral beast cancers

To illustrate the potential yield from a sequencing substudy, we consider whole genome sequencing of a subset of 200 genetically enriched subjects, with the intent of following up a subset of discovered variants by testing their associations in a larger study of 700 cases and 1400 controls. The sample sizes used for illustration derive from the WECARE study (Bernstein et al., 2004) described above. The 201 subjects in the top part of **Table S1** were selected by prioritizing young age, positive FH, cases over controls, and among cases, those who received radiotherapy (and for these, longer latency). These samples are currently being whole-genome sequenced at an average depth of coverage of 30×. We used simulation to address the following questions:

1. What is the anticipated yield of variants discovered one or more times in this sample, as a function of population MAF and RR?
2. Of those discovered, what proportion would be novel (not in the 1000 Genomes Project), what proportion would be truly causal, and both novel and causal?
3. Among the discovered variants in each category (of MAF, RR, causality, and number of times seen in each series), what

Table 1 | Parameter estimates (SEs) [Wald *Z*-tests] for the simulated two-phase sequencing data using the imputation, weighted likelihood, Breslow-Cain pseudo-likelihood, and Breslow-Holubkov semiparametric maximum likelihood estimators.

Analysis method	Subsample design		
	Case-control	Balanced[a]	Optimal[b]
ALL 1422 RARE VARIANTS IN THE FULL STUDY (47 CAUSAL)			
Imputation	1.69 (0.96) **[1.76]**	1.75 (0.86) **[2.03]**	1.63 (0.79) **[2.06]**
Weighted likelihood	1.88 (0.96) **[1.96]**	1.89 (0.91) **[2.08]**	1.72 (1.13) **[1.52]**
Pseudolikelihood	1.88 (0.96) **[1.96]**	2.03 (0.97) **[2.09]**	2.22 (1.00) **[2.22]**
Semiparametric ML	2.12 (0.98) **[2.16]**	2.22 (0.99) **[2.24]**	2.24 (1.00) **[2.24]**
Full Study	1.80 (0.69) **[2.61]**		
VARIANTS DISCOVERED IN THE SUBSTUDY ONLY			
Average number discovered (causal)	653 (44)	719 (45)	697 (44)
Imputation	1.66 (0.95) **[1.75]**	1.73 (0.86) **[2.01]**	1.64 (0.80) **[2.05]**
Weighted likelihood	2.34 (1.01) **[2.32]**	1.87 (0.93) **[2.01]**	1.73 (1.12) **[1.54]**
Pseudolikelihood	2.35 (1.02) **[2.33]**	2.01 (1.00) **[2.01]**	2.27 (0.97) **[2.34]**
Semiparametric ML	2.56 (1.06) **[2.42]**	2.19 (1.02) **[2.15]**	2.29 (0.97) **[2.36]**

Empirical mean estimates and standard deviations are computed from 1000 replicates with 2000 cases and 2000 controls showing association with at least one GWAS SNP, subsampling 600 subjects, 50 causal rare variants. These results are contrasted across three sampling designs. Coefficients are in units of log RR per Madsen-Browning rare variant summary index divided by 1000; for consistency across designs, all rare variants are included in the index in the top portion of the table; the bottom portion includes just those discovered in the substudy, so point estimates are not comparable across sampling methods. All estimates are adjusted for the risk index.

[a] *100 subjects from each of the 6 strata.*

[b] *Numbers of subjects in the subsample are fixed across replicates at (2, 20, 214) cases and (74, 116, 174) controls, stratified into 3 groups of risk index from low to high, based on overall optimization for all replicates combined.*

is the power for testing association in the main study, after Bonferroni adjustment for the number of markers tested?
4. Putting all these together, what is the expected overall yield of novel, causal discoveries?

These calculations are described in Methods section Calculation of the Expected Yield of Single-variant Tests in the WECARE Study, based on the distribution of simulated allele frequencies and RR shown in **Figure S1**. In a subsample of this size, most variants with MAF >0.1% would be seen at least once, including the most causal variants (**Table 2**). Restricting to those seen more than once considerably reduces the number of variants prioritized, as does eliminating those never or seldom reported in the 1000 Genomes Project sample, but also eliminates most of the truly causal variants. If, however, the goal is to identify *at least some* of the novel causal variants with adequate power to test them for association in the main study, then this design might still discover something in the range of 5–20 causal variants out of the total 1600 simulated, depending on the specific criteria used for prioritization, and of course depending upon the true simulation model parameters.

We also simulated gene- and pathway-level burden tests (Methods section Simulation of Gene- and Pathway-level Prioritization in the WECARE study; **Table 3**). These show a modest improvement in power at the higher levels of aggregation, but power is still low with these sample sizes. The simulated causal variants are predominantly rare, so only about 16% of causal ones are even discovered in this small sample, setting an upper bound for power for single variant tests. Of the 43 discovered causal variants (on average across 100 replicates), 19 are prioritized and 3 of these are found to be significantly associated in the full study sample, for an average power of 1.1%. Of course the corresponding proportions were much smaller for null variants, yielding only 3 false positives in total out of 31 million. (The elevated "false positive" rate for single variant tests compared with the target 0.05 is due to null variants in strong LD with other causal variants.) Similar comparisons yielded 2.9, 5.7, and 4.6% power for gene-regions, genes, and pathways respectively, with type I error rates at or below the target level. The improvement at the region and gene levels probably reflects the increasing benefit from pooling similar variants, while the failure of the pathway burden test to yield even better power may be due to an increasing proportion of truly null genes or variants diluting the effect of the positive ones. These results are based on prioritization at each level at $\alpha_1 = \sqrt{0.01/p}$ where p is the number of tests (variants, genes, etc. discovered in the subsample); while optimization of these values is possible, the results seem to be relatively insensitive across a broad range of choices. The specific results are somewhat more sensitive to the specific model parameters, only one being presented here, but the general patterns remained consistent across all values we considered.

Sequencing is still underway but preliminary results from the first 93 samples from **Table S1** suggest that the majority of subjects (mainly contralateral cases with early onset and/or family-history positive subjects) carry at least one functionally significant, clinically relevant or predicted disease-causing mutation, based on external annotation criteria including Human Genome Mutation Database (Stenson et al., 2009), ClinVar [http://www.ncbi.nlm.nih.gov/clinvar,] MutationTaster (Schwarz et al., 2010) in both known and unknown breast cancer candidate genes and pathways, and >50% carry at least two and 10% carry three or more. The next step is to see whether these variants are differentially distributed between WECARE cases and (unilateral breast cancer) controls or population rates, whether they are associated with radiotherapy (suggesting an interaction effect), and to test these variants in the full WECARE study sample.

Table 2 | Expected total number of discovered variants prioritized and expected number of these that are causal, by minimum number of copies in the sequencing sample and maximum number of copies in 1000 Genomes Project data.

Maximum copies in 1000GP	Minimum copies in sequencing sample		
	c = 1	c = 2	c = 3
NUMBER OF VARIANTS PRIORITIZED			
$c' = 0$	1.5 M	113 K	10 K
$c' = 1$	2.6 M	265 K	30 K
$c' = 2$	3.4 M	418 K	57 K
NUMBER OF PRIORITIZED VARIANTS THAT ARE CAUSAL			
$c' = 0$	41	34	27
$c' = 1$	113	97	79
$c' = 2$	192	168	140
EXPECTED YIELD OF SIGNIFICANTLY ASSOCIATED CAUSAL VARIANTS FROM SECOND STAGE*			
$c' = 0$	0.7	1.0	1.6
$c' = 1$	2.1	2.9	4.2
$c' = 2$	3.8	5.1	7.0

*Bonferroni corrected $\alpha = 0.05$ (i.e., in addition to these causal variants, 0.05 non-causal variants are expected to be declared significant).

FAMILY-BASED DESIGNS FOR PRIORITIZATION

Several investigators have recently reported approaches to efficient selection of individuals for sequencing in family-based designs. Cheung et al. (2013) described an approach for targeted sequencing of regions exploiting already available linkage information to optimize imputation to other family members, but without using phenotype information, whereas Wang et al. (2013) described an approach for whole genome sequencing using phenotype and kinship information. We simulated various family-based designs for whole genome sequencing to address the following questions:

1. What criteria should be used to select families and members for sequencing substudies?
2. What criteria should be used to prioritize variants for subsequent association testing?
3. How do family-based designs for sequencing compare with those using unrelated individuals in terms of probability of discovering novel variants, classifying variants by their

Table 3 | Simulated results of hypotheses tested in the main study for various levels of aggregation in the planned WECARE Study; means over 100 replicate simulations.

Test	True negatives				True positives			
	Total	Discovered	Prioritized	Significant	Total	Discovered	Prioritized	Significant
Pathway	87.4	87.3	0.2	0.00	12.6	12.6	1.1	0.6 (4.6%)
Gene	1016	1006	8.4	0.00	33.7	33.4	5.6	1.9 (5.7%)
Gene-region	2925	2318	32.8	0.05	97.5	79.6	12.9	2.9 (2.9%)
Single variant	31,218	6558	156	3.10	273	43.0	19.4	3.1 (1.1%)

Based on 100 pathways with an average of 10 genes each, each gene having on average 10 exonic variants (r = 1), 20 regulatory variants (r = 2), and 30 other variants surrounding the gene (r = 3) with $\sigma_{P2} = 1.0$, $\pi_P = 0.125$, $\sigma_{G2} = 0.25$, $\pi_G = 0.25$, $\sigma_{V2} = exp[-\eta_{Rv} - 0.25ln(q_v/.01)]$ *where* $\eta_1 = -1.0$, $\eta_2 = -1.5$, $\eta_3 = -2.0$, $logit(\pi_v) = \zeta_{Rv} - 0.25ln(q_v/.01)$ *where* $\zeta_1 = -0.5$, $\zeta_2 = -1.5$, $\zeta_3 = -2.5$.

likelihood of being causal, and power for testing association with disease?

We consider a two-phase design that uses a subset of individuals from a family study chosen on the basis of their phenotypes and relationships to each other for discovery and screening, followed by association testing in the full pedigrees. If enough family-based samples are available, replication using additional family-based samples is preferable to using a different sampling scheme because one would like the spectrum of variants (e.g., MAFs and RRs) being tested in the second stage to be comparable to those discovered in the first. We compare the relative efficiency of this family-based design with a conventional two-stage case-control design with comparable costs.

We evaluated these designs by simulating 4-generation pedigrees with 22 members in each. We sampled haplotypes from the same simulated population described earlier and randomly dropped genes through the pedigrees, generating phenotypes with randomly selected rare variants as causal with the same RR distribution and retaining those pedigrees with some minimum number of cases (Methods section Simulation of Family-Based Designs). To address the first question, families with various numbers of affected individuals were ascertained, and from each of these we selected individuals to sequence in various ways (e.g., two cases of at least second-degree relationship to each other and one unaffected individual). We then tabulated the following statistics for causal and non-causal variants by the relationship of the cases to each other.

- *Rule-based criterion*: the number of families for which all affected members carry the variant and no unaffected ones did among the subset sequenced;
- *Likelihood ratio (LR) criterion*: the ratio of the retrospective likelihoods under the simulated penetrances and allele frequencies vs. the null penetrance (the average rate in the ascertained families); while similar to the lod score used in linkage analysis, here a single-locus likelihood is used to test association with a directly-observed variant, not markers in LD with an unobserved locus;
- *Bayes factor (BF) criterion*: similar to the likelihood ratio, but based on the marginal probabilities under the simulated prior distributions of penetrances and allele frequencies;
- *Score test criterion*: the score test derived from the retrospective likelihood, evaluated under the null hypothesis.

(Methods section Family-based Criteria for Prioritization of Variants). The score test was evaluated both at the single-variant and the regional level, the latter using the family-based SKAT tests (Schifano et al., 2012; Chen et al., 2013; Ionita-Laza et al., 2013; Schaid et al., 2013b). The other tests were used only for ranking variants individually. The score test essentially relates the phenotypes of the entire pedigree to the genotypes of those who have been sequenced using the inverse of the kinship matrix to weight them. If linkage information is available, then a direct test of association with imputed genotypes is possible (Cheung et al., 2013), allowing for residual phenotypic correlations

Figure 1 shows the mean score statistics per family for those with a total of 4 affected individuals in which either an affected sib pair with an affected first cousin or a discordant sib pair with an affected cousin have been sequenced. These were derived for an 11-member sub-pedigree for which exhaustive enumeration of all possible genotypes and phenotypes was feasible. As expected, variants with the largest scores for causal variants (top panel) and the highest probabilities of being causal (bottom panel) were those where both cases were carriers and (if sequenced) the control not. Having the control affected somewhat lowers the average score, but not as much as having an additional case being a carrier increases it, essentially because we are considering a relatively rare disease (population prevalence 5%).

This trade-off is explored further in **Figure 2** for various types of relatives sequenced. Here, we fix the total number of individuals being sequenced at 100 across the designs being compared (e.g., 25 pedigrees with four members each sequenced, 33 with 3, 50 with 2 each, or 100 singletons). Although, as expected, having more families with fewer individuals sequenced increases the *absolute* discovery probabilities (not shown), the *relative* difference comparing causal and null variants goes the other direction, and the relative probability of prioritization also increases for more subjects per pedigree and fewer pedigrees sequenced, for a considerable increase in the overall relative probability of discovery *and* prioritization.

Comparing designs with two individuals sequenced per pedigree shows little difference in the probability of discovery across the relationships among the pairs (**Figure 2**), but shows that the

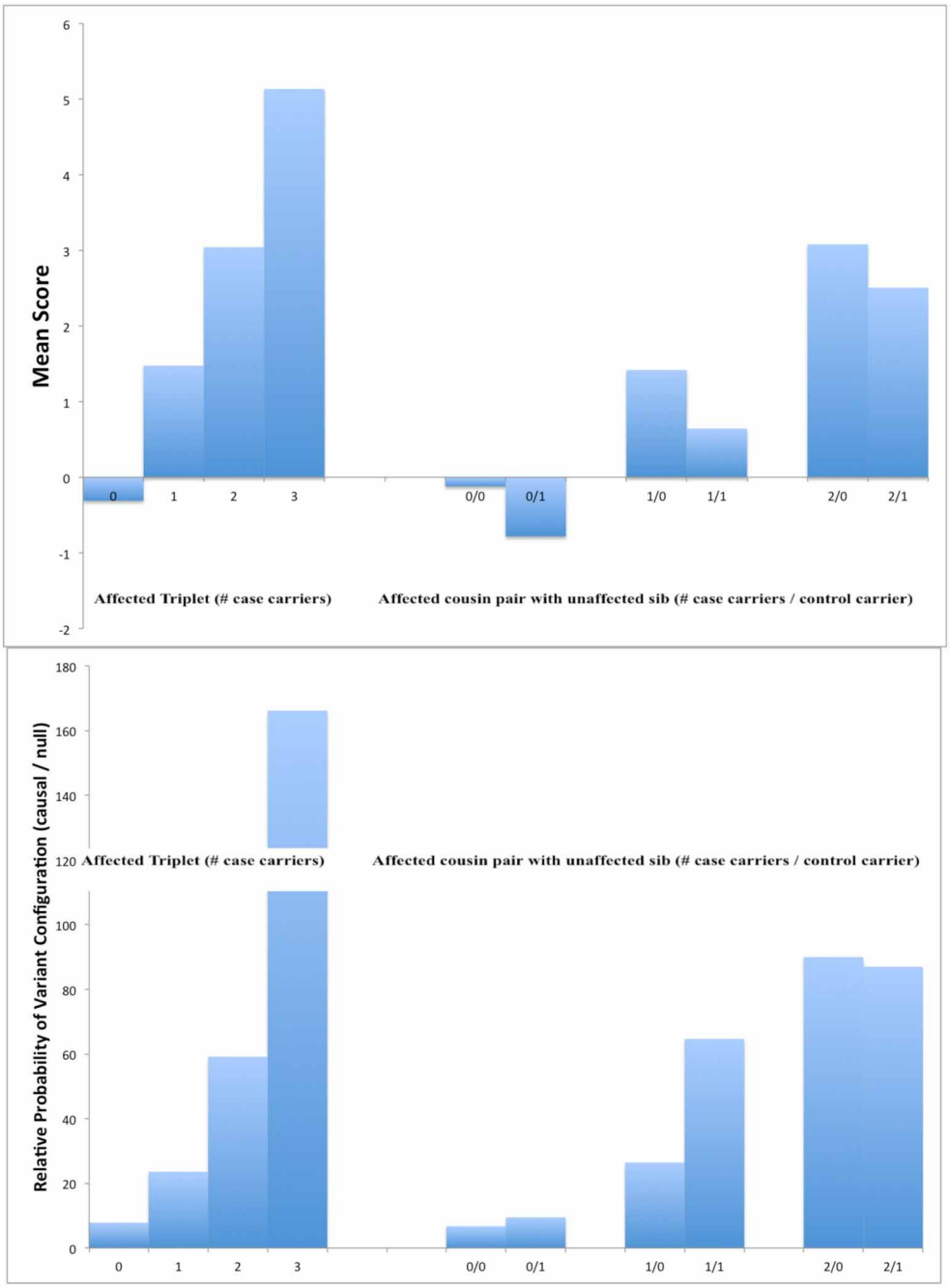

FIGURE 1 | Mean scores for causal variants (top panel) and ratio of frequencies of causal to non-causal variants (bottom panel) in simulated 11-member pedigrees with at least 4 affected members. In each panel, results are shown for a design sequencing an affected sib pair and affected cousin by the number of carriers of the variant allele (left) or an affected first cousin pair and an unaffected sib by the number of carriers among cases and controls (right).

conditional probability of prioritization given discovery and the joint probability of discovery and prioritization is better for more distant relatives.

Obviously, the more stringent the cutoff for any of these criteria, the fewer the variants that would be prioritized, but non-causal variants tend to be eliminated much faster than causal variants (**Figure S2**), so the challenge is to choose a threshold that minimizes the false positive proportion, subject to the total number of variants that can be tested in subsequent replication efforts. The relative performance of the various prioritization criteria is illustrated in **Figure 3** as Receiver Operating Curves, varying these thresholds. Although the BF criterion is the best overall in

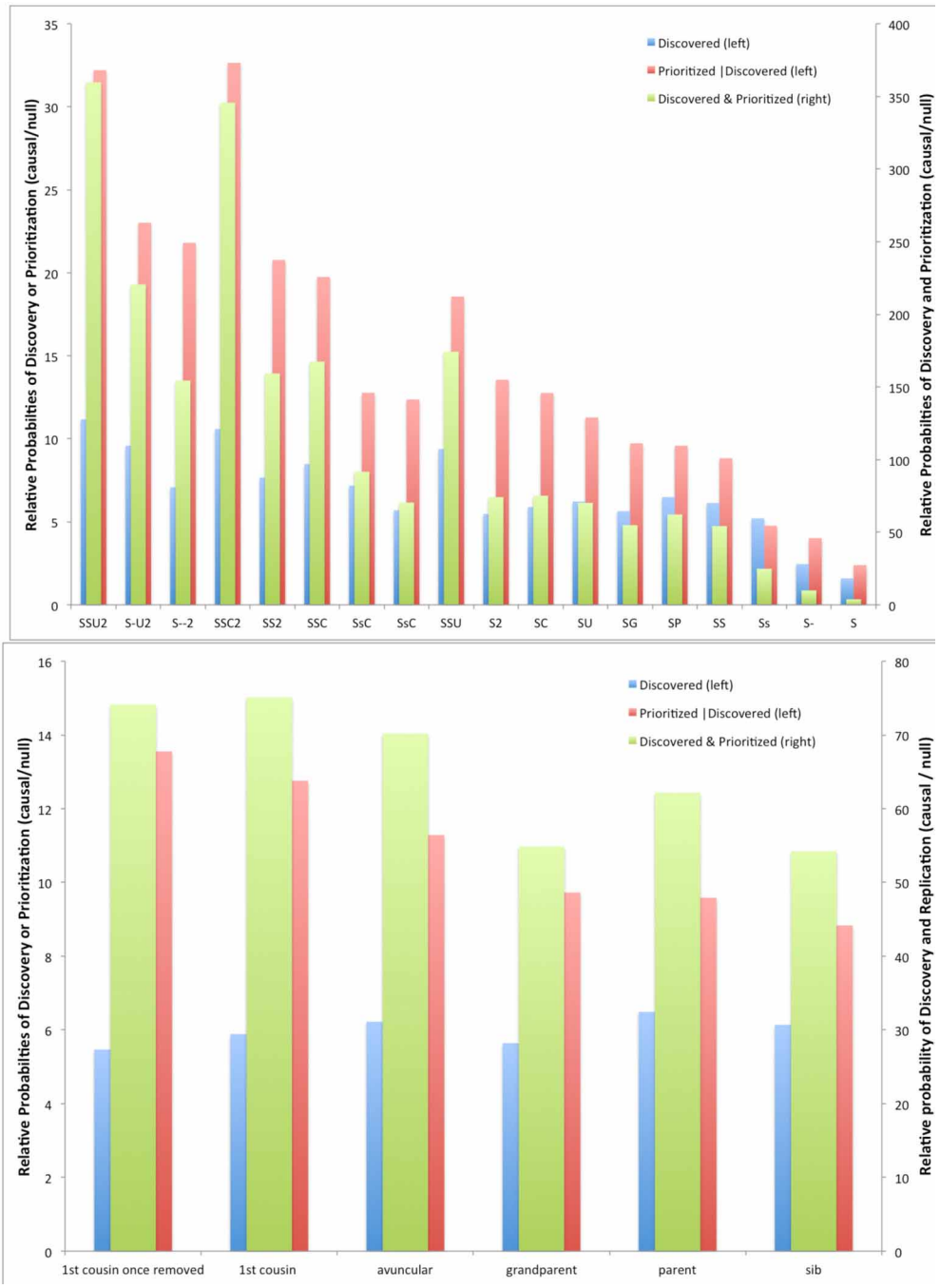

FIGURE 2 | Relative probabilities of discovery, prioritization, and both between causal vs. null variants for different criteria for selecting members for sequencing in simulated 11-member pedigrees with at least 4 affected members. Top panel, all designs; **bottom panel**, detail for designs with only two members sequenced. (Codes for **top panel**: S, sib; C, cousin; 2, first cousin once removed; U, uncle; G, grandparent; P, parent; Upper case, affected, lower case, unaffected; hyphen, affected but not sequenced.)

terms of the area under the curve, it is the most computationally intensive and the score test is nearly as good and much faster to compute.

RELATIVE EFFICIENCY OF FAMILY- vs. POPULATION-BASED DESIGNS. We compared the power of a two-stage family-based design with that of a conventional case-control design. The overall power for any design with independent tests in the two stages is simply the product of the powers for the two stages (Methods section Calculation of Power for Two-stage Designs). The probabilities of discovery, prioritization, and replication are illustrated in **Table 4** for a range of design parameters—total sample sizes, proportions allocated to stage 1, numbers of copies required to be judged a discovery, and the minimum threshold required

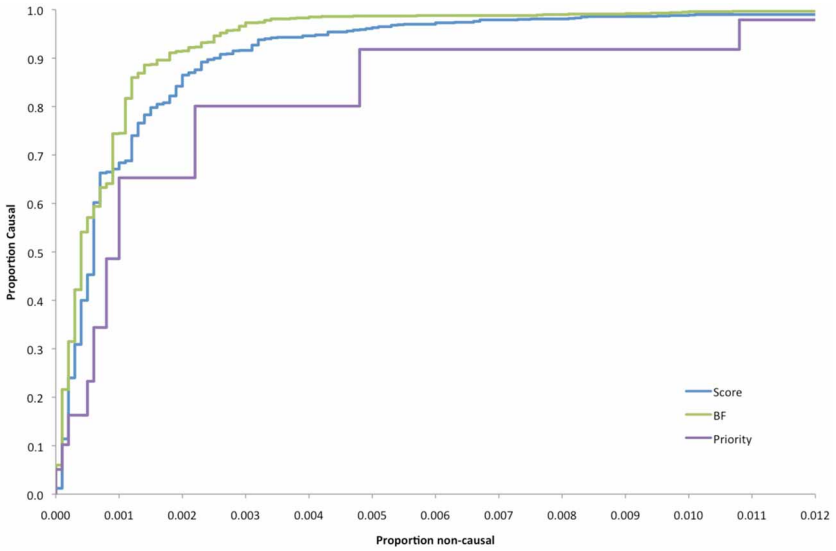

FIGURE 3 | Receiver operating curves comparing different prioritization schemes.

Table 4 | Some near-optimal multi-stage family-based and case-control designs (The first row of each block is the one with the highest ARCE among those investigated; the second is the one with better power among those with similar costs.)

Total sample size[a]	Proportion allocated to stage 1 (%)	Minimum copies for discovery	Criterion for prioritization[b]	Proportion of causal discovered (%)	Total number prioritized[c]	Power for all causal all novel) variants	Total cost (millions)	ARCE[d]) (×1000)
FAMILY-BASED DESIGNS (COSTS: $1000/FAMILY, $5000/SEQUENCE, 5¢/GENOTYPE)								
1800	30	12	3.0	38	3,591	17% (9%)	$12.3	0.252
2400	30	14	3.0	13	3,894	21% (13%)	$16.8	0.241
FAMILY-BASED DESIGNS (COSTS: $5000/FAMILY, $1000/SEQUENCE, 5¢/GENOTYPE)								
2100	50	12	4.0	56	346	19% (12%)	$13.8	0.264
2400	50	12	4.0	59	378	21% (13%)	$15.8	0.260
CASE-CONTROL DESIGNS (COSTS: $100/SUBJECT, $5000/SEQUENCE, 5¢/GENOTYPE)								
7000	20	12	0.001	62	5,502	16% (9%)	$17.8	0.171
9000	20	14	0.001	65	5,862	20% (12%)	$23.1	0.164
CASE-CONTROL DESIGNS (COSTS: $500/SUBJECT, $1000/SEQUENCE, 5¢/GENOTYPE)								
6000	40	16	0.0001	70	823	17% (11%)	$8.1	0.416
7000	40	16	0.0001	74	912	20% (13%)	$9.4	0.407

[a] *Number of 22-member pedigrees for family-based designs; number of cases, number of controls for case-control designs.*

[b] *Minimum score test for family-based designs; minimum p-value for case-control designs.*

[c] *Assuming 1000 causal variants out of a total of 20 million.*

[d] *Total number of true positives, inversely weighted by the square root of MAF, divided by total cost.*

for prioritization—under two different cost structures. (The full range of choices considered is shown in **Figure S3**.) Obviously, as the number required for discovery is lowered or the threshold for prioritization is raised, fewer variants in total would be prioritized, leading to a less stringent multiple comparisons penalty, but at some point the overall power decreases because too many of the truly causal variants are either not discovered or not prioritized. Although the overall power (the proportion of *all* causal variants discovered, prioritized, and replicated) for any of these designs is only about 20% (for a sample size of 1000 families), the majority of those not found are either very rare or have very small effect sizes. Of particular interest are the numbers of *novel* variants (those not in the 1000 Genomes Project database) that are discovered, prioritized, and replicated. Since these too are predominately rare, power for them is even lower, but depending upon the total number that actually exist, they could still represent a substantial yield of true positive findings.

These comparisons are provided for the "optimal" designs of each type and one alternative design that yields better power at modestly larger cost. Assuming costs of $1000 per family for enrollment and obtaining pedigree phenotypes, $100 per subject enrolled in a case-control design, $5000 per whole-genome sequence, and $0.05 per subject-genotype, the optimal family-based design turns out to require 540 pedigrees in stage I and

1260 in stage II at a critical value of the score test for prioritizing variants of 3.0; this yields 167 true-positive replicated associations out of 1000 simulated at a cost of \$12 M. The corresponding optimal case-control design would require 1400 case-control pairs in stage I, 5600 in stage II at a critical value of $\alpha_1 = 0.001$ in stage I, for a yield of 159 true positive replicated associations at a cost of \$18 M (about 2/3 the cost-efficiency of the family-based design). Of course, with different cost ratios, these optimal designs would change, as illustrated in **Table 4** for the case where enrollment costs are 5 times larger and sequencing only \$1000 per whole genome. In this instance, case-control designs turn out to be the more cost-efficient. For this simulated MAF distribution, only 175 of the 1000 causal variants would be novel, but in every situation considered, the power for discovering such rare variants is still more than half that for all causal variants. No striking differences were seen between the spectrum of RRs and MAFs discovered, prioritized, and replicated by the two types of designs with sample sizes chosen to yield similar overall power.

APPLICATION TO THE COLORECTAL CANCER FAMILY REGISTRIES DATA

Included in the Colorectal Cancer Family Registries are a few large families for whom all available family members have either been genotyped for previously replicated GWAS SNPs or whole exome sequence variants. We analyzed one of these—a large Australian pedigree comprising 145 individuals with a total of 7 colorectal, 1 Lynch syndrome, and 9 other cancer cases. Genotypes for 32 GWAS-associated SNPs were available for 49 of the members, including 5 of the CRC and Lynch cases and 4 of the other cancers. These data were used to illustrate the effect of subsampling. We selected individuals under various criteria, calculated LR, BF, and score statistics using only the SNP data for these selected individuals [but all the phenotype information (Visscher and Duffy, 2006)], and compared these results to those from the complete genotype data to see which criteria best distinguish variants that are "truly" associated (based on the complete data) from "false" positives.

Figure 4 shows the correlation of each statistic computed using all available genotype data (as the "gold standard") with those using only the subsample of genotypes (averaging over 10 replicate subsamples). For the CRC and Lynch syndrome phenotype, we compared subsets of 1–4 cases and 0–3 controls out of the 5 available cases and 44 available controls. These showed little improvement in correlation for any of the test statistics from adding more than about 2 cases. Adding the genotypes for one or two unaffected members somewhat improved the correlation for LRs and BFs when there were 2 or more cases, but surprisingly worsened the correlation when only a single case was included; this may simply reflect instability due to the small number of cases in total. Results were somewhat more stable when all 9 cancer cases with available genotypes were considered, allowing comparisons of larger subsamples of cases. Adding more than about 3 cases did not materially improve the correlation and adding 1–3 controls improved the correlations only modestly, again reducing them when a single control was added to a single case. The bottom panel shows that the more distant relative pairs were more informative about distinguishing apparently associated from non-associated SNPs (based on the complete data).

Because only common variants were available in the Australian data, we also performed similar simulations on 15 large pedigrees that had previously been included in a linkage scan (Cicek et al., 2012) and had whole exome data available on 2–3 CRC cases from each. Not surprisingly, in this small dataset, no genomewide significant associations were found by the score test with any of the 359,744 single nucleotide variants (SNVs) called at least once (a third of these were called only 3 or fewer times) or with 100-SNV bins with the regional score SKAT test, nor were the regional tests particularly correlated with the maximum single-SNV tests or with prior annotation. Additional simulations (not shown) based on the real sequence data and simulated phenotypes (conditional on the total number of cases in each pedigree) confirmed that 15 families would be far too few to find any significant causal effects, even if IBD information were used. The design of a larger NGS study is described in the concluding section.

DISCUSSION

One of the advantages of family-based designs is that Mendelian inconsistencies can be used to check for genotyping errors (Pompanon et al., 2005). This has, of course, long been recommended as a routine quality control check in linkage and family-based GWAS studies. This advice becomes even more important when dealing with NGS data because of its inherently higher error rate as a function of depth of sequencing and quality control filters applied (Faye et al., 2013). Further research on approaches to using pedigree information to improve variant calling would be helpful.

In a similar vein, Mendelian inheritance could be exploited for improved imputation of variants in unsequenced family members. On obvious way to proceed would be to first use standard imputation procedures with external reference populations (Howie et al., 2012), treating each subject with GWAS SNP data as independent to obtain preliminary genotype probabilities at the sequenced variants. These could then be combined with the observed genotype calls for the sequenced family members using Mendelian transmission probabilities to obtain refined genotype probabilities (Burdick et al., 2006). While proceeding variant-by-variant in this manner is relatively straightforward, it fails to take LD patterns among the variants into account, but a similar strategy could be applied to haplotypes (Cheung et al., 2013). Such a two-step approach was used in simulations for the Genetic Analysis Workshop 18 (http://www.gaworkshop.org/gaw18/index.html). Ideally, a unified approach that would integrate the two sources of information in a single step would be preferable. In addition to imputation, identity-by-descent information could be used to inform the selection of subjects for sequencing (Cheung and Wijsman, 2013) and directly as a local genetic similarity kernel in family-based SKAT tests.

Homozygosity mapping in families has proven to be a valuable technique for mapping recessive alleles (Kruglyak et al., 1995; Chahrour et al., 2012). Design issues for sequencing studies for such traits are likely to be somewhat different from those considered here and would be useful avenue for further research.

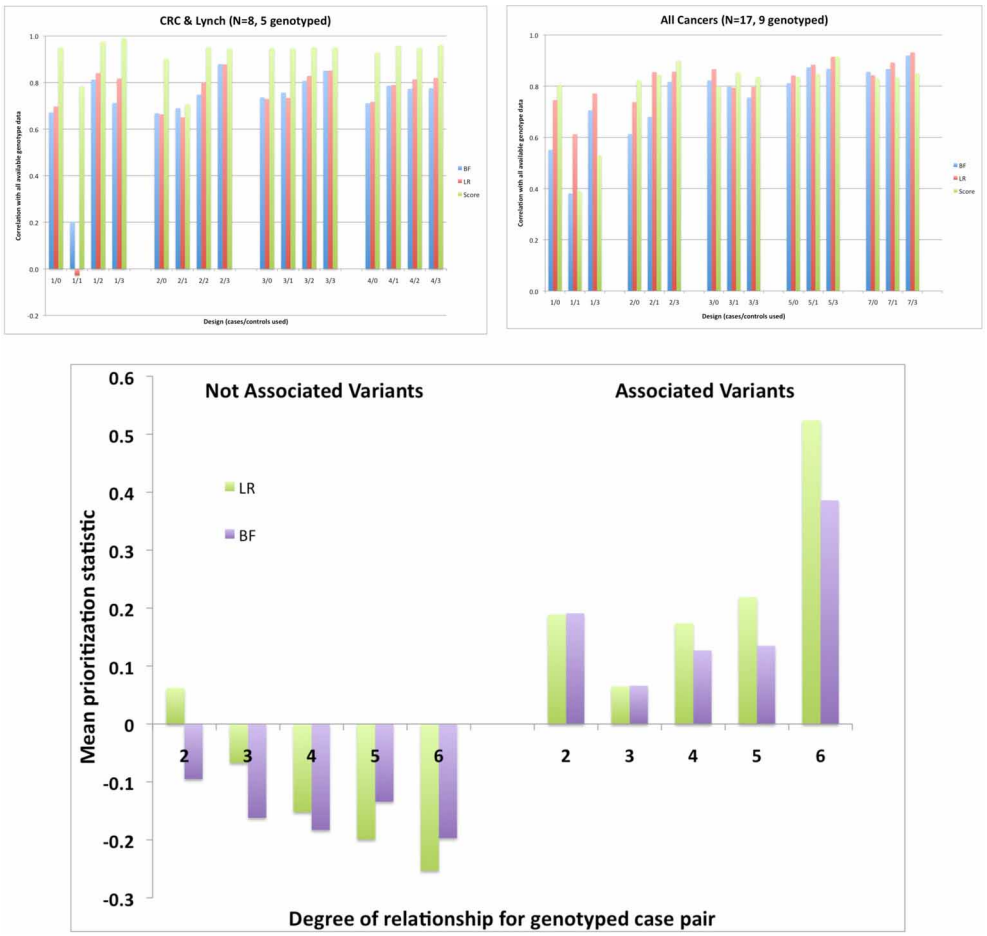

FIGURE 4 | Correlation across 32 GWAS SNPs between the statistics computed from the complete genotype data and those computed using only the genotypes for various subsets of members; top left: 5 genotyped CRC and Lynch syndrome cases; top right: 9 cases of any cancer. Bottom: prioritization statistics by degree of relationship for apparently associated or unassociated variants based on the complete data. Data from a single 145-member Australian pedigree with a total of 8 CRC or Lynch syndrome cases and 15 cases of any cancer and a total of 49 subjects genotyped.

Another possibility is to use a two-step analysis of the *same* data, exploiting between-family comparisons to prioritize variants and within-family comparisons to test the most promising ones, in the spirit of Van Steen et al. (2005). Because these two tests are independent, one then need only correct the significance level for the number of variants passed to the second stage. For quantitative traits, regression of the offspring phenotypes on the mean of the parents' genotypes provides a simple first-step test. For disease traits, one would have to include control trios, nuclear families with varying proportions of affected offspring, or external control individuals to have the variability in phenotypes needed for the first-step test. Practically, however, this approach would require having access to the DNA for case-parent trios (which might not be available for late-onset diseases like cancer) and sequencing of the parents rather than the cases for the first stage; not only would this double the sequencing costs over a more conventional design that sequences only the cases, but it might seem counter-intuitive since the parents may not themselves be affected (even though at least one of each pair must carry any variant that case does). One could of course reverse the two steps, but this would require sequencing the entire trio in the first step and the final inference would not be robust to population stratification. In further simulation studies (not shown), we found that use of external controls tends to be more powerful than between-family comparisons for the first step, but is more susceptible to population stratification bias; this is not a threat to validity if used in the first step, but could reduce power if too many false-positives are passed to the second step, inflating the multiple testing penalty. The two-step analysis approach consistently yielded better power than a two-stage case-control design, however.

Two-stage and two-phase designs are also amenable to considerable cost-efficiency gains by using DNA pooling techniques (Sham et al., 2002) in the first stage, thereby allowing one to sequence many more subjects than would be feasible if one were to sequence individuals. Of course, only aggregate allele frequency

information (Huang et al., 2010), not individual genotypes, are then available [unless one uses molecular bar-coding techniques (Craig et al., 2008)], but these can still be used for discovery of novel variants (Lee et al., 2011) or case-control comparisons of pool allele frequencies (Johnson, 2007; Macgregor et al., 2008; Zhao and Wang, 2009). Further cost-efficiencies are possible by constructing pools of pools, with bar-coding of the sub-pools (Smith et al., 2010). Optimization of designs using DNA pooling has been described by Liang et al. (2012), but extension to family-based studies remains a challenge (Lee, 2005).

We conclude by describing how these considerations influenced the design of a planned whole-exome sequencing study within the colorectal CFR. We are planning a three-stage family-based design, in which the first stage would use already available sequence information for prioritizing about 1000 genes. This would be followed by two stages of replication, each with probands from about 1000 multiple case families and 1000 controls. The first stage would exploit existing control data from the 1000 Genomes Project, while the second and third stages would use individually matched population controls. Because our hypothesis is that causal genes may harbor multiple rare variants—not necessarily the same across families—the two replication stages would perform full resequencing of the entire coding and flanking regions of the prioritized genes, 1000 genes in stage 2, 100 genes in stage 3. For the same reason, we have decided to use a family-based design for all three stages, since variants discovered in multiple-case families may not be well represented in unselected series of population-based cases. After analysis of the sequencing data from each stage, additional genotyping of the prioritized variants would be done on all other available family members for analyses using a conditional segregation analysis (Hopper et al., 1999). Three criteria would be used for prioritization at every stage: a family-based test of co-segregation with disease for each variant separately and for entire genes; a gene-based test of association comparing cases and controls; and filtering based on bioinformatics predictors. The first of these uniquely exploits the information available from a family-based design and can be used to rank genes on the probability they carry at least one causal variant, using an aggregate assessment of the impact of all rare variants in the gene. The three comparisons would be unified through hierarchical modeling, in which both the family-based and case-control comparisons would be incorporated in the likelihood for the first (individual)-level model, and the bioinformatics predictors would be incorporated in the second (variant)-level model. Similar issues are currently being discussed in the design of a large-scale sequencing study for the WECARE project. Since this is not a family-based study, the key decision there is how best to select the subset for sequencing in a two-phase design.

METHODS

All simulations were based on the same population of 10,000 haplotypes of length 250 Kb generated by the COSI program (Schaffner et al., 2005) with the population history parameters provided in their **Table 1**. This population contained 5125 unique variants, of which 4557 had minor allele frequencies (MAF) <0.05, 95% less than 0.01, 79% less than 0.001.

SIMULATION OF TWO-PHASE DESIGNS

We postulated a disease model involving multiple rare variants drawn from the simulated haplotype population with the probability of having any effect and the expected size of the effect depending inversely on the MAF (**Figure S1**). We then sampled pairs of haplotypes at random from this population, computed their risk under this model, and assigned case-control status, continuing in this manner until the target number of 1000 cases and 1000 controls were obtained for the parent GWAS, and tested these data for association with all common SNPs (MAF >5%). If a significant association is found with one or more SNPs, the replicate was retained for the sequencing substudy.

For the subsample, we first constructed a risk index based on a multiple logistic regression of disease state on all non-redundant GWAS SNPs and stratified the phase 1 subjects into three strata of high, medium, and low risk (with cutpoints at the 25th and 75th percentiles). We compared case-control, balanced, and optimal sampling of 600 subjects total out of the available 4000 (Methods Section Optimization of Two-phase Studies). For these subjects, we retained *all* variants, including the causal ones but also many more irrelevant ones.

Finally, we conducted a joint analysis of both phases. We tested association with the Madsen and Browning (2009) index—the number of rare variants weighted inversely by the square root of their allele frequencies—as the rare-variant covariate of interest, treated as continuous, using the WL, PL, and semi-parametric likelihood methods described in Methods Section Likelihoods for Joint Analysis of Two-phase Studies. We found that the risk index used for sampling was a confounder of the Madsen-Browning rare variant index, due to LD among the variants included in each, due to differences between the weights in the Madsen-Browning index and the simulated weights, and due to having used the disease status to construct the risk index, so all results were adjusted for the sampling risk index. This entire process was repeated 1000 times.

LIKELIHOODS FOR JOINT ANALYSIS OF TWO-PHASE STUDIES

Following the general notation used by Breslow and Holubkov (1997a,b), we let V represent a set of GWAS SNPs in a region found to be associated with disease Y, and X represent the causal variant(s) in LD with the GWAS SNPs, to be discovered by sequencing the region on a subsample of subjects. For this purpose, we defined the causal variable X to be the Madsen-Browning index. The imputation strategy entails simply fitting a regression model for $X|V$ to the substudy data, and then using $\hat{X}(V)$ as the covariate for $Y|X$ in the full study. Proper inference would, however, require that the uncertainty in the imputation be taken into account in the analysis of the main study data. We now describe a formal likelihood approach to accomplish this.

If the first stage is a case-control sample, then the full likelihood would be the retrospective probability

$$L_1(\alpha,\ \beta,\ \gamma) = \Pr(\mathrm{V}|\mathrm{Y}) = \prod_{i=1}^{N} \sum_{x} \Pr(V_i, X_i = x|Y_i)$$

$$= \prod_{i=1}^{N} \frac{p_\gamma(V_i) \sum_x p_\beta(Y_i|x) p_\alpha(x|V_i)}{\sum_v p_\gamma(v) \sum_x p_\beta(Y_i|x) p_\alpha(x|V_i)}$$

and the likelihood for the second stage sample S would be

$$L_2(\alpha,\ \beta) = \Pr(X|Y,\ V,\ S) = \prod_{j\in S} \frac{p_\beta(Y_i|X_i)p_\alpha(X|V_i)}{\sum_x p_\beta(Y_i|x)p_\alpha(x|V_i)}$$

The full likelihood would then be $L(\theta) = L_1(\alpha,\ \beta,\ \gamma)L_2(\alpha,\ \beta)$ where $\theta = (\alpha,\ \beta,\ \gamma)$. In practice, however, both V and X are highly multidimensional and we wish to avoid having to specify their joint LD distribution parametrically. When we do not assume functional forms of $p_\alpha\ (X|V)$ and $p_\gamma(V)$ the likelihood above becomes suitable for semiparametric maximum likelihood (SPML) estimation. Both Breslow and Holubkov (1997a,b) and Scott et al. (2007) have developed profile likelihoods by maximizing out the high-dimensional parameters $p_\alpha(X|V)$ and $p_\gamma(V)$, with a different parameterization. We followed a recent formulation of the problem from Scott et al., in which the estimating equations for β and the constraints for nuisance parameters π are described in a "log-likelihood"

$$l^*(\beta,\ \pi) = \sum_{y,\ v,\ x} n_{yvz} \log p^*_{yv}(x;\ \beta,\ \pi) + \sum_{y,\ v} N_{yv} \log \pi_{yv} - \sum_{y,\ v} n_{yv} + \log(N_{+v}\pi_{yv} - \overline{N}_{yv})$$

where $\pi_{1v} = 1 - \pi_{0v}, \overline{N}_{yv} = N_{yv} - n_{yv}$, and

$$p^*_{1v}(x;\ \beta,\ \pi) = \text{expit}\left[x'\beta + \log\left(N_{+v} - \frac{\overline{N}_{1v}}{\pi_{1v}}\right) - \log\left(N_{+v} - \frac{\overline{N}_{0v}}{\pi_{0v}}\right)\right]$$

By iterating between maximizing a logistic likelihood with fixed offsets containing $_\pi$ and updating $_\pi$ using its constraint equations, we can obtain semiparametric (SPML) efficient estimates of β. The SPML approach has the advantage of being flexible about the distribution of covariates X and V while retaining good efficiency. However, the derivation of the semiparametric estimating equations was complex enough that it appeared only after two approximation methods—the WL and the PL—had been published. The WL approach weights individual score functions from model $p_\beta(Y|X)$ inversely proportional to the sampling probabilities. In our implementation, the original Horvitz-Thompson weights N_{yv}/n_{yv} were used, although an improvement might be to use predicted weights $1/\Pr(S=1|Y,V,Z)$ that could incorporate auxiliary information Z from the full cohort (say, from a logistic model), or better yet, the calibrated weights described in Breslow et al. (2009b). The PL approach was first developed in Breslow and Cain (1988), representing an alternative that uses first phase information. Following their seminal paper, we work with a PL based on $p_{\beta,\ \delta}(Y|X,V,S)$, which incorporates the parameter of interest β and the nuisance log-odds δ for $Y = 1$ in stratum $V = v$. They insert estimates $\hat{\delta}_v = \log(N_{1v}/N_{0v})$ into the PL and then solve for β. Schill et al. (1993) proposed to estimate δ and β simultaneously; this method, although not implemented here, has been reported to yield similar results as the Breslow and Cain (1988) version.

OPTIMIZATION OF TWO-PHASE STUDIES

Our general strategy for optimization of any of these two-phase designs aims to solve the following problem. Suppose we have collected phase I data on N subjects and seek to selectively assemble data on at most n subjects based on available information. The available information from phase I, mainly observations of Y and V, is summarized by the cell sizes N_{yv}. What we wish to optimize are the cell-specific sampling fractions, denoted by $s_{yv} = n_{yv}/N_{yv}$. A natural choice of objective function might be to aim for more precise parameter estimates per unit cost, for example using the Asymptotic Relative Cost Efficiency (Thomas, 2007). However, while this goal could be readily achieved for a scalar parameter as in Reilly (1996), it is less clear when more than one parameter is estimated. In this work, we chose our objective function to be the non-centrality parameter of the likelihood ratio test for $H_0 : \hat{\beta} = 0$ vs. $H_1 : \tilde{\beta} \neq 0$, with $\tilde{\beta}$ being the subset of interest in β. This objective is equivalent to a linear combination of information matrix entries, and is thus a good summary of the standard error estimates. We denote the entire parameter vector including β and other nuisance parameters as θ. It has been shown (Self et al., 1992; Brown et al., 1999) that the non-centrality parameter can be computed as $\lambda = 2E_{\theta\ A}[i(\theta_A) - l(\hat{\theta}_A)]$, where $l(.)$ is the log-likelihood function, θ_A is the true parameter vector under the alternative hypothesis, and $\hat{\theta}_0$ is the parameter vector that maximizes $E_{\theta A}[l(\theta)]$ under the null hypothesis. A slightly different form, $\lambda' = \lambda - \nu$ with ν being the degrees of freedom of the test, is also reasonable. In this particular problem, we used $l^*(.)$ shown in the previous section in place of $l(.)$. It has been shown (Scott and Wild, 1989; Scott et al., 2007) that the profile log likelihood $l^*(.)$ is amenable for standard likelihood ratio tests.

This problem setting requires N_{yv}, $\Pr(X|Y,V)$, β_0 (true value of β) to be either known or pre-specified. To obtain these quantities, we simulated 1000 data sets as described in section Simulation of Two-phase Designs, and consider these data sets as the underlying "super-population." Then we used the estimates of $\Pr(X|Y,V)$, β_0 and the average values of N_{yv} from this super-population as the input to the optimization procedure. Hence we used a fixed optimal design for all simulation replicates, using the SPML. Compared to solving the optimization problem for each replicate, this strategy represents a solution to the "expected" problem.

CALCULATION OF THE EXPECTED YIELD OF SINGLE-VARIANT TESTS IN THE WECARE STUDY

The estimates in **Table 2** were derived using only the MAF distribution from the simulated haplotype population. The expected carrier probabilities in each age, FH, and disease stratum were computed from the assumed distribution of RRs over a grid of MAF and RR values using standard Mendelian inheritance methods. We computed the probability of observing c copies of a variant in the subsample from the Poisson distribution for each MAF and RR bin, summing the expected counts over all sampling strata, and in a similar manner, we computed the probability of observing c' copies among 1000 population controls. The total yield of discovered novel variants is then the sum over these bins of the number of variants in each bin times the probability of seeing $\geq c$ and $\leq c'$ copies.

For association testing in the main study, we computed the NCP for the Mantel-Haenszel test (stratified by age and FH) for each bin and then computed power by reference to the cumulative normal distribution with Bonferroni correction for either the total number of discovered variants or only the number of novel variants. These are again summed over all RR and MAF bins to estimate the expected yield for various values of c and c' given in **Table 2**.

SIMULATION OF GENE- AND PATHWAY-LEVEL PRIORITIZATION IN THE WECARE STUDY

To compare the power of single-variant and burden tests, we selected variants from the haplotypes in a multi-level fashion as follows. We defined 100 pathways p, each comprising 1–20 genes g, further subdivided into three regions r (e.g., "exons," "introns and promoter regions," and "more distant enhancer regions"). Starting and ending locations of each gene and is sub-regions were selected at random and all variants v within these regions were included. Pathways, genes, and variants were selected as causal with probabilities π_P, π_G, and π_v respectively, where π_v for variant v depends upon the type of region and its MAF. Thus, a variant has a causal effect only if all three levels are designated as causal. Each causal variant was assigned a log RR β_v as a sum of pathway, gene, and variant-level effects, each being the absolute value of a normal deviate with zero mean and variance σ_P^2, σ_G^2, and σ_v^2, respectively, σ_v^2 also depending upon the type of region and MAF. We drew two haplotypes at random for each potential subject and computed the genetic log RR as the sum of the β_vs for each variant they carried. Subjects were assigned at random to an age stratum and then to disease (unaffected, unilateral, bilateral) and FH strata with probabilities depending upon their age and genetic RR. This process was continued until the target number of subjects in each age, FH, and disease stratum was obtained and a random subset of these was designated as the sequencing sample. In the real WECARE study, we prioritized the youngest cases, those with a positive FH, and radiotherapy subjects with the longest latency for sequencing. In this way, we selected 201 subjects out of the total of 2199 available for sequencing; the distribution of the entire study sample and the sequencing subsample by age, FH, and laterality is provided in **Table S1**.

For the analysis, we scanned the subsample to identify all variants seen at least twice. All single variants that were seen more frequently than expected (by an amount depending on the number of comparisons) based on the general population MAFs and similarly all pathways, genes, or regions that were seen more frequently than expected were prioritized. These are tested for case-control association in the main study, using a Cochran-Mantel-Haenzsel (CMH) test, stratified by age and FH, with Bonferroni adjustment for the number of comparisons at each level.

SIMULATION OF FAMILY-BASED DESIGNS

Family-based simulations used a fixed pedigree structure of 4 generations with two offspring in each generation for a total of 22 members in 7 nuclear families. We sampled two haplotypes at random for each of the founders from the simulated haplotype population and dropped them at random without recombination through the non-founders. As before, we chose causal variants and their RRs depending upon MAF, computed the genetic log RR as the sum of the β_vs for each causal variant a subject carries, and assigned disease status accordingly, adjusting the intercept to yield a population prevalence of 5%. Families with the required number of cases (set to 4 for most of the results reported here) were retained and the process continued until 1000 such families were ascertained. Various criteria were used to select a subset of family members whose genotypes were to be retained for analysis (the "sequencing subset," e.g., two affected individuals of at least second-degree relationship and one unaffected member), while retaining the phenotype information for the entire pedigree.

For each of the causal and a random sample of the non-causal variants, we computed the likelihood ratio, Bayes factor, and score statistics (described below) and tabulated these values for different configurations of genotypes among the sequenced members and their relationships to each other. For the rule-based prioritization, we also tabulated the number of variants found in at least $f_{\min}$ families and the number of these that were prioritized by having the target genotype configuration (e.g., both cases being carriers and the control not). The distributions of causal and non-causal variants for each criterion are shown in **Figure S2** as a function of the threshold for prioritization; plotting one curve against the other yields the ROCs displayed in **Figure 3**.

FAMILY-BASED CRITERIA FOR PRIORITIZATION OF VARIANTS

Rule-based criterion

Variants were classified on the basis of the number of families in which all sequenced cases carried the variant and any sequenced controls did not.

Likelihood ratio criterion

Following the principles described in Petersen et al. (1998), we estimated the probability that any particular variant is causal under a given genetic model by accumulating likelihood ratio contributions (comparing the likelihoods of the data under the alternative hypothesis that a particular variant is causal to that under the null hypothesis that it is not causal) across families. Letting Y denote the phenotypes of all family members (including those not sequenced), G^{bs} the observed sequence data, β_v the genetic RR and q_v the minor allele frequency for variant v, the likelihood ratio is

$$LR_v = \frac{\Pr(G_v^{obs}|Y;\beta_v, q_v)}{\Pr(G_v^{obs}|q_v)} = \frac{\Pr(G_v^{obs}|Y)}{\Pr(G_v^{obs})\Pr(Y)}$$

where $\Pr(G_v^{obs}, Y) = \sum_{G_v^{unobs}} \Pr(Y|G_v)\Pr(G_v)$. These calculations were done evaluating the likelihood under the simulated RR and minor allele frequency for each variant under the alternative hypothesis and under the induced marginal population risk for the null hypothesis.

Bayes factor criterion

The likelihood ratio criterion requires a maximization of the likelihood under the alternative hypothesis, which can be unstable for rare variants. To avoid this, Petersen et al. (1998) compute a

Bayes factor by averaging over a prior distribution of MAF and RR. Bayes Factors are computed as the ratio of these marginal probabilities of the joint genotypes of the sampled individuals under the true model to that under the null. Of course, we do not know the true values of either β or q, or even their true probability distributions, so we used the simulated probability distributions for $\Pr(q)$ and $\Pr(\beta|q)$, averaging over a random samples of 100 parameter values drawn from their null and alternative distributions.

Score test criterion

We computed the score statistic T_v for a single variant v derived from a multivariate logistic model for the phenotypes of the entire pedigree and the genotypes of the sequenced subset as

$$T_v = \Sigma_f t_{fv} = \Sigma_f (Y_f - p_f \mathbf{1})' K_f^{-1} (G_{fv} - q_{fv} 1)$$

$$= \Sigma_f [\Sigma_{i \in Nf} \Sigma_{j \in Sf} (Y_{fi} - p_f) K_f^{ij} (G_{fjv} - q_{fv})]$$

where Y_f is the vector of phenotypes for family f, p_f is the family-specific disease prevalence, K_f is the kinship matrix, G_{fv} the vector of genotypes for variant v, and q_{fv} is the mean of G_v among the sequenced members. The $G_{fv} - q_{fv}$ deviations are set to zero for untyped individuals, but the inclusion of the kinship terms for typed-untyped pairs allows their phenotypes to contribute. This statistic has mean zero under the null hypothesis and asymptotic variance $\text{var}(T_v) = \Sigma_f {t_{fv}}^2$. For the purpose of prioritizing variants we used the score test ${T_v}^2/\text{var}(T_v)$ for each variant and select the top-ranked ones at some cutoff. This provides a pure within-family comparison, but those families for which all sequenced individuals are no carriers or all are carriers become uninformative. A more powerful test that exploits both between- and within-family information replaces the p_f and q_{fv} by and, the corresponding means over all families. The regional (SKAT) test for all variants in region R is simply $[\Sigma_f (\Sigma_{v \in R} t_{fv}^2)]^2 / \Sigma_f (\Sigma_{v \in R} {t_{fv}}^2)^2$.

CALCULATION OF POWER FOR TWO-STAGE DESIGNS

Using family-based simulation described above, we tabulated the average number of families in the simulated sample in which each variant was seen at least once and the NCP as the mean of the simulated score statistics by bins of MAF and RR. To compare designs with different stage 1 sample sizes and discovery thresholds, we rescaled the numbers of families carrying a given variant by the ratio of proposed and simulated sample sizes and recomputed the probability of discovery by reference to the Poisson distribution in each bin. In a similar manner, we computed the probability of prioritization at stage one at threshold $\lambda_{\min}$ in each bin by rescaling the NCPs by the ratio of sample sizes and referred them to the non-central chi square distribution. To extend our simulation results to the whole genome, we multiplied the predicted number of simulated null variants meeting our discovery and prioritization criteria by 20,000,000/1000. The total number of variants carried forward to stage 2 is then simply the sum over all MAF and RR bins of the product of the number of variants in the population times the probabilities of discovery and prioritization. Power for stage two was computed in a similar manner by rescaling the NCPs by the stage 2 sample size and referring it to the non-central chi square distribution with Bonferroni correction for this number of tests carried forward. Thus, the yield of causal variants discovered, prioritized, and replicated is the sum of the number of variants in the population times the probability of discovery, prioritization, and replication over all MAF and RR bins. In each MAF bin, we also computed the Poisson probability that a variant would not have been seen at least twice in 1000 population controls and computed the power for novel variants in a similar manner.

Calculations for case-control designs were similar, except that no simulation was required. The probability of discovery could be computed directly from the Poisson distribution in the combined case and control sample and the NCP s for the chi square test for allelic association computed in the usual way for a 2×2 contingency table of allele counts by case-control status.

ACKNOWLEDGMENTS

Supported in part by NIH grants U01 HG005927, U19 CA148107, R01 ES019876, R01 CA129639, P30 CA014089, P30 ES 07048. The authors are grateful to Drs. Paul Marjoram and David Conti for many helpful discussions and to the investigators from the Colorectal Cancer Family Registries (Dr. Steve Thibodeau, P.I. of the proposed sequencing substudy) and WECARE Study (Dr. Jonine Bernstein, P.I.) for the studies used for illustration, particularly Dr. David Duggan for the summary of preliminary data from the WECARE Study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2013.00276/abstract

Figure S1 | Simulation parameters for models 1 and 2: top: probability of causality and mean RR for causal variants as a function of MAF; bottom: frequency distribution of non-causal and causal variants as a function of MAF.

Figure S2 | Yield of prioritized variants as a function of the number of families required for prioritization, the minimum Bayes factor, and the minimum score.

Figure S3 | Two-stage designs using Bayes factors for prioritization. Top panel, varying number of variants required for discovery (1–4) and minimum BF for prioritization; **bottom panel**, detail of left-most portion, varying the sample size for replication $N = 20(\times 2)10880$. The colors indicate the proportions of simulated causal variants that are discovered, prioritized, and discovered.

Table S1 | Sample sizes used to illustrate power calculations for a two-phase design for the WECARE study (note that the actual sequencing sample is further stratified by radiotherapy and latency for the purpose of studying gene-radiation interactions, factors not considered in these simulations). *UBC, unilateral breast cancer (controls); CBC, contralateral (second asynchronous) breast cancer (cases)

REFERENCES

Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* 44, 293–308. doi: 10.1146/annurev-genet-102209-163421

Asimit, J., and Zeggini, E. (2012). Imputation of rare variants in next generation association studies. *Hum. Hered.* 74, 196–204. doi: 10.1159/000345602

Bacanu, S. A., Nelson, M. R., and Whittaker, J. C. (2012). Comparison of statistical tests for association between rare variants and binary traits. *PLoS ONE* 7:e42530. doi: 10.1371/journal.pone.0042530

Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* 35, 606–619. doi: 10.1002/gepi.20609

Begg, C. B., Haile, R. W., Borg, A., Malone, K. E., Concannon, P., Thomas, D. C., et al. (2008). Variation of breast cancer risk among BRCA1/2 carriers. *JAMA* 299, 194–201. doi: 10.1001/jama.2007.55-a

Bernstein, J. L., Haile, R. W., Stovall, M., Boice, J. D. Jr., Shore, R. E., Langholz, B., et al. (2010). Radiation exposure, the ATM Gene, and contralateral breast cancer in the women's environmental cancer and radiation epidemiology study. *J. Natl. Cancer Inst.* 102, 475–483. doi: 10.1093/jnci/djq055

Bernstein, J. L., Langholz, B., Haile, R. W., Bernstein, L., Thomas, D. C., Stovall, M., et al. (2004). Study design: evaluating gene-environment interactions in the etiology of breast cancer–the WECARE study. *Breast Cancer Res.* 6, R199–214. doi: 10.1186/bcr771

Bernstein, J. L., Thomas, D. C., Shore, R. E., Robson, M., Boice, J. D. Jr., Stovall, M., et al. (2013). Contralateral breast cancer after radiotherapy among BRCA1 and BRCA2 mutation carriers: a WECARE study report. *Eur. J. Cancer* 49, 2979–2985. doi: 10.1016/j.ejca.2013.04.028

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40, 695–701. doi: 10.1038/ng.f.136

Borg, A., Haile, R. W., Malone, K. E., Capanu, M., Diep, A., Torngren, T., et al. (2010). Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum. Mutat.* 31, E1200–E1240. doi: 10.1002/humu.21202

Breslow, N., and Cain, K. (1988). Logistic regression for two-stage case-control data. *Biometrika* 75, 11–20. doi: 10.1093/biomet/75.1.11

Breslow, N., McNeney, B., and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome-dependent sampling. *Ann. Stat.* 31, 1110–1139. doi: 10.1214/aos/1059655907

Breslow, N. E., and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl. Stat.* 48, 457–468. doi: 10.1111/1467-9876.00165

Breslow, N. E., and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Stat. Soc. B* 59, 447–461. doi: 10.1111/1467-9868.00078

Breslow, N. E., and Holubkov, R. (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Stat. Med.* 16, 103–116.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009a). Improved horvitz-thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Stat. Biosci.* 1, 32. doi: 10.1007/s12561-009-9001-6

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* 169, 1398–1405. doi: 10.1093/aje/kwp055

Breslow, N. E., and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Stat.* 34, 86–102. doi: 10.1111/j.1467-9469.2006.00523.x

Breslow, N. E., and Zhao, L. P. (1988). Logistic regression for stratified case-control studies. *Biometrics* 44, 891–899. doi: 10.2307/2531601

Brooks, J. D., Teraoka, S. N., Reiner, A. S., Satagopan, J. M., Bernstein, L., Thomas, D. C., et al. (2012). Variants in activators and downstream targets of ATM, radiation exposure, and contralateral breast cancer risk in the WECARE study. *Hum. Mutat.* 33, 158–164. doi: 10.1002/humu.21604

Brown, B. W., Lovato, J., and Russell, K. (1999). Asymptotic power calculations: description, examples, computer code. *Stat. Med.* 18, 3137–3151. doi: 10.1002/(SICI)1097-0258(19991130)18:22<3137::AID-SIM239>3.0.CO;2-O

Burdick, J. T., Chen, W. M., Abecasis, G. R., and Cheung, V. G. (2006). *In silico* method for inferring genotypes in pedigrees. *Nat. Genet.* 38, 1002–1004. doi: 10.1038/ng1863

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput.* 368–379. doi: 10.1142/9789812836939_0035

Cain, K., and Breslow, N. (1988). Logistic regression analysis and efficient design for two-stage studies. *Am. J. Epidemiol.* 128, 1198–1206.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* 86, 6–22. doi: 10.1016/j.ajhg.2009.11.017

Capanu, M., and Begg, C. B. (2011). Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics* 67, 371–380. doi: 10.1111/j.1541-0420.2010.01469.x

Capanu, M., Concannon, P., Haile, R. W., Bernstein, L., Malone, K. E., Lynch, C. F., et al. (2011). Assessment of rare BRCA1 and BRCA2 variants of unknown significance using hierarchical modeling. *Genet. Epidemiol.* 35, 389–397. doi: 10.1002/gepi.20587

Chahrour, M. H., Yu, T. W., Lim, E. T., Ataman, B., Coulter, M. E., Hill, R. S., et al. (2012). Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genet.* 8:e1002635. doi: 10.1371/journal.pgen.1002635

Chasman, D. I. (2008). On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet. Epidemiol.* 32, 658–668. doi: 10.1002/gepi.20334

Chen, G. K., and Witte, J. S. (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. *Am. J. Hum. Genet.* 81, 397–404. doi: 10.1086/519794

Chen, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* 37, 196–204. doi: 10.1002/gepi.21703

Cheung, C. Y. K., Thompson, E. A., and Wijsman, E. M. (2013). GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *Am. J. Hum. Genet.* 92, 504–516. doi: 10.1016/j.ajhg.2013.02.011

Cheung, C. Y. K., and Wijsman, E. M. (2013). "Design matters! A statistical framework to guide sequencing choices in pedigrees (IGES abstract #9)," in *International Genetic Epidemiology Society* eds C. Greenwood, J. Loreno Bermejo, B. Fridley, J. Houwing-Duistermaat, A. Paterson, S. Shete, et al. (Chicago, IL: Genetic Epidemiology). 3–4.

Cicek, M. S., Cunningham, J. M., Fridley, B. L., Serie, D. J., Bamlet, W. R., Diergaarde, B., et al. (2012). Colorectal cancer linkage on chromosomes 4q21, 8q13, 12q24, and 15q22. *PLoS ONE* 7:e38175. doi: 10.1371/journal.pone.0038175

Cirulli, E. T., and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11, 415–425. doi: 10.1038/nrg2779

Concannon, P., Haile, R. W., Borresen-Dale, A. L., Rosenstein, B. S., Gatti, R. A., Teraoka, S. N., et al. (2008). Variants in the ATM gene associated with a reduced risk of contralateral breast cancer. *Cancer Res.* 68, 6486–6491. doi: 10.1158/0008-5472.CAN-08-0134

Conneely, K. N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of *P* values for multiple correlated tests. *Am. J. Hum. Genet.* 81, 1158–1168. doi: 10.1086/522036

Conti, D. V., and Witte, J. S. (2003). Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am. J. Hum. Genet.* 72, 351–363. doi: 10.1086/346117

Conti, D. V., Lewinger, J. P., Swan, G. E., Tyndale, R. F., Benowitz, N. L., and Thomas, P. D. (2009). "Using ontologies in hierarchical modeling of genes and exposures in biologic pathways," in *Phenotypes and Endophenotypes: Foundations for Genetic Studies of Nicotine Use and Dependence,* ed G. E. Swan (Bethesda, MD: NCI Tobacco Control Monographs), 539–584.

Craig, D. W., Pearson, J. V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J. J., et al. (2008). Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* 5, 887–893. doi: 10.1038/nmeth.1251

Duan, Q., Liu, E. Y., Auer, P. L., Zhang, G., Lange, E. M., Jun, G., et al. (2013). Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. *Bioinformatics* 29, 2744–2749. doi: 10.1093/bioinformatics/btt477

Elston, R. C., Lin, D., and Zheng, G. (2007). Multistage sampling for genetic studies. *Annu. Rev. Genomics Hum. Genet.* 8, 327–342. doi: 10.1146/annurev.genom.8.080706.092357

Faye, L. L., Machiela, M. J., Kraft, P., Bull, S. B., and Sun, L. (2013). Re-Ranking sequencing variants in the Post-GWAS era for accurate causal variant identification. *PLoS Genet.* 9:e1003609. doi: 10.1371/journal.pgen.1003609

Feng, T., Elston, R. C., and Zhu, X. (2011). Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet. Epidemiol.* 35, 398–409. doi: 10.1002/gepi.20588

Feng, T., Zhang, S., and Sha, Q. (2007). Two-stage association tests for genome-wide association studies based on family data with arbitrary family structure. *Eur. J. Hum. Genet.* 15, 1169–1175. doi: 10.1038/sj.ejhg.5201902

Freedman, M. L., Monteiro, A. N., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.* 43, 513–518. doi: 10.1038/ng.840

Gorlov, I. P., Gorlova, O. Y., Frazier, M. L., Spitz, M. R., and Amos, C. I. (2011). Evolutionary evidence of the effect of rare variants on disease etiology. *Clin. Genet.* 79, 199–206. doi: 10.1111/j.1399-0004.2010.01535.x

Greenland, S. (2000). Principles of multilevel modelling. *Int. J. Epidemiol.* 29, 158–167. doi: 10.1093/ije/29.1.158

Heinzen, E. L., Depondt, C., Cavalleri, G. L., Ruzzo, E. K., Walley, N. M., Need, A. C., et al. (2012). Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. *Am. J. Hum. Genet.* 91, 293–302. doi: 10.1016/j.ajhg.2012.06.016

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi: 10.1073/pnas.0903103106

Hoffmann, T. J., Marini, N. J., and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* 5:e13584. doi: 10.1371/journal.pone.0013584

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24, 2784–2785. doi: 10.1093/bioinformatics/btn516

Hopper, J. L., Southey, M. C., Dite, G. S., Jolley, D. J., Giles, G. G., McCredie, M. R. E., et al. (1999). Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. *Cancer Epidemiol. Biomarkers Prev.* 8, 741–747.

Horvitz, D., and Thompson, D. (1952). A generalization of sampling without replacement from a finite population. *J. Am. Stat. Assoc.* 47, 663–685. doi: 10.1080/01621459.1952.10483446

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959. doi: 10.1038/ng.2354

Huang, Y., Hinds, D. A., Qi, L., and Prentice, R. L. (2010). Pooled versus individual genotyping in a breast cancer genome-wide association study. *Genet. Epidemiol.* 34, 603–612. doi: 10.1002/gepi.20517

Hung, R. J., Baragatti, M., Thomas, D., McKay, J., Szeszenia-Dabrowska, N., Zaridze, D., et al. (2007). Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiol. Biomarkers Prev.* 16, 2736–2744. doi: 10.1158/1055-9965.EPI-07-0494

Hung, R. J., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P., et al. (2004). Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol. Biomarkers Prev.* 13, 1013–1021.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet.* 21, 1158–1162. doi: 10.1038/ejhg.2012.308

Ionita-Laza, I., and Ottman, R. (2011). Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 189, 1061–1068. doi: 10.1534/genetics.111.131813

Johnson, T. (2007). Bayesian method for gene detection and mapping, using a case and control design and DNA pooling. *Biostatistics* 8, 546–565. doi: 10.1093/biostatistics/kxl028

Karchin, R. (2009). Next generation tools for the annotation of human SNPs. *Brief. Bioinform.* 10, 35–52. doi: 10.1093/bib/bbn047

Kruglyak, L., Daly, M., and Lander, E. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am. J. Hum. Genet.* 56, 519–527.

Lange, C., Demeo, D., Silverman, E. K., Weiss, S. T., and Laird, N. M. (2003). Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am. J. Hum. Genet.* 73, 801–811. doi: 10.1086/378591

Langholz, B., Thomas, D. C., Stovall, M., Smith, S. A., Boice, J. D. Jr., Shore, R. E., et al. (2009). Statistical methods for analysis of radiation effects with tumor and dose location-specific information with application to the WECARE study of asynchronous contralateral breast cancer. *Biometrics* 65, 599–608. doi: 10.1111/j.1541-0420.2008.01096.x

Lee, J. S., Choi, M., Yan, X., Lifton, R. P., and Zhao, H. (2011). On optimal pooling designs to identify rare variants through massive resequencing. *Genet. Epidemiol.* 35, 139–147. doi: 10.1002/gepi.20561

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237. doi: 10.1016/j.ajhg.2012.06.007

Lee, W. C. (2005). A DNA pooling strategy for family-based association studies. *Cancer Epidemiol. Biomarkers Prev.* 14, 958–962. doi: 10.1158/1055-9965.EPI-04-0503

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., and Thomas, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* 31, 871–882. doi: 10.1002/gepi.20248

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 10, 387–406. doi: 10.1146/annurev.genom.9.081307.164242

Liang, W. E., Thomas, D. C., and Conti, D. V. (2012). Analysis and optimal design for association studies using next-generation sequencing with case-control pools. *Genet. Epidemiol.* 36, 870–881. doi: 10.1002/gepi.21681

Macgregor, S., Zhao, Z. Z., Henders, A., Nicholas, M. G., Montgomery, G. W., and Visscher, P. M. (2008). Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Res.* 36, e35. doi: 10.1093/nar/gkm1060

Madsen, B. E., and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5:e1000384. doi: 10.1371/journal.pgen.1000384

Malone, K. E., Begg, C. B., Haile, R. W., Borg, A., Concannon, P., Tellhed, L., et al. (2010). Population-based study of the risk of second primary contralateral breast cancer associated with carrying a mutation in BRCA1 or BRCA2. *J. Clin. Oncol.* 28, 2404–2410. doi: 10.1200/JCO.2009.24.2495

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Minelli, C., De Grandi, A., Weichenberger, C. X., Gögele, M., Modenese, M., Attia, J., et al. (2013). Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genet. Epidemiol.* 37, 205–213. doi: 10.1002/gepi.21705

Murphy, A., Weiss, S. T., and Lange, C. (2008). Screening and replication using the same data set: testing strategies for family-based studies in which all probands are affected. *PLoS Genet.* 4:e1000197. doi: 10.1371/journal.pgen.1000197

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., et al. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.* 7:e1001322. doi: 10.1371/journal.pgen.1001322

Newcomb, P. A., Baron, J., Cotterchio, M., Gallinger, S., Grove, J., Haile, R., et al. (2007). Colon cancer family registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomarkers Prev.* 16, 2331–2343. doi: 10.1158/1055-9965.EPI-07-0648

Neyman, J., and Scott, E. (1966). On the use of c(alpha) optimal tests of composite hypotheses. *Bull. Int. Stat. Inst.* 41, 477–497.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Panagiotou, O. A., Ioannidis, J. P. A., and Project, F. T. G.-W. S. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* 41, 273–286. doi: 10.1093/ije/dyr178

Petersen, G. M., Parmigiani, G., and Thomas, D. (1998). Missense mutations in disease genes: a Bayesian approach to evaluate causality. *Am. J. Hum. Genet.* 62, 1516–1524. doi: 10.1086/301871

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.* 6, 847–859. doi: 10.1038/nrg1707

Quintana, M. A., Berstein, J. L., Thomas, D. C., and Conti, D. V. (2011). Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genet. Epidemiol.* 35, 638–649. doi: 10.1002/gepi.20613

Quintana, M. A., and Conti, D. V. (2013). *Integrative variable selection via Bayesian model uncertainty*. *Stat. Med.* 32, 4928–4953. doi: 10.1002/sim.5888

Quintana, M. A., Schumacher, F. R., Casey, G., Bernstein, J. L., Li, L., and Conti, D. V. (2012). Incorporating prior biologic information for high-dimensional rare variant association studies. *Hum. Hered.* 74, 184–195. doi: 10.1159/000346021

Rebbeck, T. R., Spitz, M., and Wu, X. (2004). Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.* 5, 589–597. doi: 10.1038/nrg1403

Reilly, M. (1996). Optimal sampling strategies for two-stage studies. *Am. J. Epidemiol.* 143, 92–100. doi: 10.1093/oxfordjournals.aje.a008662

Reilly, M., and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 82, 299–314. doi: 10.1093/biomet/82.2.299

Reiner, A. S., John, E. M., Brooks, J. D., Lynch, C. F., Bernstein, L., Mellemkjaer, L., et al. (2013). Risk of asynchronous contralateral breast cancer in noncarriers of BRCA1 and BRCA2 mutations with a family history of breast cancer: a report from the Women's environmental cancer and radiation epidemiology study. *J. Clin. Oncol.* 31, 433–439. doi: 10.1200/JCO.2012.43.2013

Roeder, K., Bacanu, S. A., Wasserman, L., and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* 78, 243–252. doi: 10.1086/500026

San Lucas, F. A., Wang, G., Scheet, P., and Peng, B. (2012). Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* 28, 421–422. doi: 10.1093/bioinformatics/btr667

Satagopan, J. M., and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* 25, 149–157. doi: 10.1002/gepi.10260

Satagopan, J. M., Venkatraman, E. S., and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60, 589–597. doi: 10.1111/j.0006-341X.2004.00207.x

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E., and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* 58, 163–170. doi: 10.1111/j.0006-341X.2002.00163.x

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15, 1576–1583. doi: 10.1101/gr.3709305

Schaid, D. J. (2010a). Genomic similarity and kernel methods i: advancements by building on mathematical and statistical foundations. *Hum. Hered.* 70, 109–131. doi: 10.1159/000312641

Schaid, D. J. (2010b). Genomic similarity and kernel methods ii: methods for genomic information. *Hum. Hered.* 70, 132–140. doi: 10.1159/000312643

Schaid, D. J., Jenkins, G. D., Ingle, J. N., and Weinshilboum, R. M. (2013a). Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. *Genet. Epidemiol.* 37, 229–238. doi: 10.1002/gepi.21708

Schaid, D. J., McDonnell, S. K., Sinnwell, J. P., and Thibodeau, S. N. (2013b). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet. Epidemiol.* 37, 409–418. doi: 10.1002/gepi.21727

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L. R., Peyser, P. A., et al. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* 36, 797–810. doi: 10.1002/gepi.21676

Schill, W., Jockel, K. H., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* 80, 339–352. doi: 10.1093/biomet/80.2.339

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 19, 212–219. doi: 10.1016/j.gde.2009.04.010

Schwarz, J. M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576. doi: 10.1038/nmeth0810-575

Scott, A. J., Lee, A. J., and Wild, C. J. (2007). On the Breslow-Holubkov estimator. *Lifetime Data Anal.* 13, 545–563. doi: 10.1007/s10985-007-9054-0

Scott, A. J., and Wild, C. J. (1989). Likelihood ratio tests in case/control studies. *Biometrika* 76, 806–809. doi: 10.1093/biomet/76.4.806

Self, S. G., Mauritsen, R. H., and Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* 48, 31–39. doi: 10.2307/2532736

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA Pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* 3, 862–871. doi: 10.1038/nrg930

Shi, G., and Rao, D. C. (2011). Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genet. Epidemiol.* 35, 572–579. doi: 10.1002/gepi.20597

Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* 38, 209–213. doi: 10.1038/ng1706

Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2007). Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* 31, 776–788. doi: 10.1002/gepi.20240

Smith, A. M., Heisler, L. E., St Onge, R. P., Farias-Hesson, E., Wallace, I. M., Bodeau, J., et al. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* 38, e142. doi: 10.1093/nar/gkq368

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., et al. (2009). The human gene mutation database: 2008 update. *Genome Med.* 1, 13. doi: 10.1186/gm13

Stovall, M., Smith, S. A., Langholz, B. M., Boice, J. D. Jr., Shore, R. E., Andersson, M., et al. (2008). Dose to the contralateral breast from radiotherapy and risk of second primary breast cancer in the WECARE study. *Int. J. Radiat. Oncol. Biol. Phys.* 72, 1021–1030. doi: 10.1016/j.ijrobp.2008.02.040

Stram, D. O., Pearce, C. L., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., et al. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum. Hered.* 55, 179–190. doi: 10.1159/000073202

Thomas, D. C. (2007). Multistage sampling for latent variable models. *Lifetime Data Anal.* 13, 565–581. doi: 10.1007/s10985-007-9061-1

Thomas, D. C. (2012). Some surprising twists on the road to discovering the contribution of rare variants to complex diseases. *Hum. Hered.* 74, 113–117. doi: 10.1159/000347020

Thomas, D. C., Casey, G., Conti, D. V., Haile, R. W., Lewinger, J. P., and Stram, D. O. (2009a). Methodological issues in multistage genome-wide association studies. *Stat. Sci.* 24, 414–429. doi: 10.1214/09-STS288

Thomas, D. C., Conti, D. V., Baurley, J., Nijhout, F., Reed, M., and Ulrich, C. M. (2009b). Use of pathway information in molecular epidemiology. *Hum. Genomics* 4, 21–42. doi: 10.1186/1479-7364-4-1-21

Thomas, D., Xie, R., and Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. *Genet. Epidemiol.* 27, 401–414. doi: 10.1002/gepi.20047

Thompson, J. R., Gögele, M., Weichenberger, C. X., Modenese, M., Attia, J., Barrett, J. H., et al. (2013). SNP prioritization using a bayesian probability of association. *Genet. Epidemiol.* 37, 214–221. doi: 10.1002/gepi.21704

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B., Lyon, H., Demeo, D. L., et al. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37, 683–691. doi: 10.1038/ng1582

Visscher, P. M., and Duffy, D. L. (2006). The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. *Genet. Epidemiol.* 30, 30–36. doi: 10.1002/gepi.20124

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* 81, 208–227. doi: 10.1086/519024

Wang, H., Thomas, D. C., Pe'er, I., and Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* 30, 356–368. doi: 10.1002/gepi.20150

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Wang, M., Jakobsdottir, J., Smith, A. V., Gudnason, V., and McPeek, M. S. (2013). "Optimal selection of individuals for genotyping in genetic association studies with related individuals (IGES abstract #5)," in *International Genetic Epidemiology Society* (Chicago, IL: Genetic Epidemiology).

Whittemore, A. S. (2007). A Bayesian false discovery rate for multiple testing. *J. Appl. Stat.* 34, 1–9. doi: 10.1080/02664760600994745

Whittemore, A. S., and Halpern, J. (1997). Multi-stage sampling in genetic epidemiology. *Stat. Med.* 16, 153–167.

Witte, J. S., Gauderman, W. J., and Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am. J. Epidemiol.* 149, 693–705. doi: 10.1093/oxfordjournals.aje.a009877

Wu, M. C., Lee, S., Cai, T., 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. doi: 10.1016/j.ajhg.2011.05.029

Yang, F., and Thomas, D. C. (2011). Two-stage design of sequencing studies for testing association with rare variants. *Hum. Hered.* 71, 209–220. doi: 10.1159/000328193

Zhao, Y., and Wang, S. (2009). Optimal DNA pooling-based two-stage designs in case-control association studies. *Hum. Hered.* 67, 46–56. doi: 10.1159/000164398

Zhu, X., Feng, T., Li, Y., Lu, Q., and Elston, R. C. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genet. Epidemiol.* 34, 171–187. doi: 10.1002/gepi.20449

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 September 2013; paper pending published: 29 October 2013; accepted: 19 November 2013; published online: 13 December 2013.

*Citation: Thomas DC, Yang Z and Yang F (2013) Two-phase and family-based designs for next-generation sequencing studies. Front. Genet. **4**:276. doi: 10.3389/fgene.2013.00276*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.