Published in final edited form as:

Genet Epidemiol. 2012 July ; 36(5): 472–479. doi:10.1002/gepi.21641.

Optimized Selection of Unrelated Subjects for Whole Genome Sequencing Studies of Rare High-Penetrance Alleles

Todd L. Edwards^{1,2} and Chun Li^{2,3}

¹Vanderbilt Epidemiology Center, Division of Epidemiology, Department of Medicine, Vanderbilt University, Nashville, TN 37203

²Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37212

³Department of Biostatistics, Vanderbilt University, Nashville, TN 37212

Abstract

Sequencing studies using whole-genome or exome scans are still more expensive than genome wide association studies (GWAS) on a per-subject basis. As a result, subsets of subjects from a larger study are often selected for sequencing. To perform an agnostic investigation of the entire genome, subjects may be selected that capture independent ancestral lineages, i.e. founder genomes, and thus avoid redundant information from regions that were inherited identical by descent (IBD) from a common ancestor. We present SampleSeq2 which can be used to select a subset of optimally unrelated subjects with minimal IBD sharing. It also can be used to estimate the number, G_T , of founder chromosomes in a sample or select the minimum number of subjects that will carry a target G_T . We evaluated SampleSeq2 compared to a random draw of a small number of subjects both by simulation and using the Anabaptist genealogy. SampleSeq2 provided an increase in G_T relative to a random draw across a range of small sample sizes. This increase in founder chromosomes improves the power of association tests, mitigates the effect of cryptic relatedness on parameter estimates, increases the total yield of alleles from sequencing, and minimizes the average size of regions shared IBD around disease alleles in cases.

Software: <http://biostat.mc.vanderbilt.edu/SampleSeq>

Keywords

resequencing; ancestry; cryptic relatedness; study design; subject selection

INTRODUCTION

Next-generation sequencing technology allows investigators to search the genome at single base resolution to identify genetic determinants for traits. Since it is generally not feasible for an investigator to sequence all subjects in a large cohort, questions arise as to the optimal design for studies aimed at identifying novel variants that explain the genetics of a trait. Common questions and considerations include: finding an optimal balance between numbers of subjects and the depth of sequencing, whether to follow-up by further sequencing of selected variants or imputation, whether to use DNA pooling or family-based designs, and the choice of specific subjects for sequencing (Duncan Thomas and Fan Yang, personal

Correspondence should be addressed to: Chun Li, PhD, Center for Human Genetics Research, 519 Light Hall, Vanderbilt University Medical Center, Nashville, TN 37212-0700, USA, chun.li@vanderbilt.edu.

Conflict of interest: None declared.

communications). We had previously introduced SampleSeq, a method for efficient selection of subjects for disease allele discovery from targeted resequencing studies of regions with previously observed associations [Edwards, et al. 2011]. In this paper, we propose SampleSeq2, an optimized approach for selecting subjects for agnostic sequencing studies of the entire genome or exome. In our previous work, we condition on associated SNP genotypes, disease status, and the prevalence of the trait in the population to select subjects most likely to carry a rare causal allele. In this approach, we use global estimates of kinship from either genotypes or pedigree structures to select subjects who are the best available representation of the founder chromosomes for the population.

Many sequencing studies focus on participants who are believed to be unrelated. Cryptic relatedness reduces the number of lineages among the study participants, who are mosaics of a set of ancestral sequences. This notion of mosaicism is also the basis for genotype imputation methods which condition on reference haplotypes of randomly selected and densely genotyped subjects. These methods are based on the assumption that the reference haplotypes are a good representation of the founder chromosomes, and by extension also represent all other members of the population who share those ancestors [Howie, et al. 2009; Li, et al. 2010; Marchini, et al. 2007].

Among closely related subjects the time to coalescence for a sequence is shorter than in less related subjects, and a higher proportion of the genome is inherited identical-by-descent (IBD). As the time to coalescence decreases, chromosomal regions unrelated to the trait of interest may also be shared IBD, and follow the pattern expected for a trait allele. Additionally, the shared regions around influential alleles are less likely to have been broken by recombination and would be longer among more closely related subjects, increasing the cost and effort of subsequent investigations. Furthermore, cryptic relatedness among study subjects causes variance inflation for association test statistics and increased rates of false positive associations [Voight and Pritchard 2005].

An allele that is a strong determinant for a trait will likely lie in an ancestral haplotype clade shared by other subjects who display the trait, which is the logical basis for both homozygosity and admixture mapping [Lander and Botstein 1987; McKeigue 1998]. Several recent examples of this come from studies where observing sequences that are IBD in a few unrelated cases has been an effective strategy for identifying regions containing rare high-penetrance alleles [Hoischen, et al. 2010; Lalonde, et al. 2010; Ng, et al. 2009; Ng, et al. 2010; Simpson, et al. 2011]. Additionally, there has been recent interest in developing methods that use IBD information to augment association tests or identify the intersection of regions that are shared IBD among subjects [Browning and Browning 2011; Browning and Thompson 2012; Moltke, et al. 2011]. Other methods have focused on identifying haplotypes or haplotype clades that tag causal rare variants [Durrant, et al. 2004; Zhu, et al. 2010], or detecting IBD in groups larger than pairs [Leibon, et al. 2008; Thomas, et al. 2008].

The level of relatedness between two individuals can be measured by the fraction of genome shared IBD or by the kinship coefficient [Lange 1997]. Often the subjects from a large cohort where whole genome or exome sequencing is considered already have GWAS data available. GWAS data often have an average density of 6 kb or less, which is sufficient for estimating the IBD distribution [Manichaikul, et al. 2010; Purcell, et al. 2007]. If genotype information is not available, but pedigree structure is, kinship coefficients between pairs of subjects can be calculated and the expected fraction of genome shared IBD can be derived. The amount of sharing between two subjects will vary around this expected value. Thus, if both types of information are available, genotype information is expected to provide a more correct estimate of IBD sharing than solely relying on pedigree structure and the assumption

of independent founders. Furthermore, recent natural selection that influences IBD sharing and errors in pedigree records can be detected using genotypes [Albrechtsen, et al. 2010; Han and Abney 2011].

In SampleSeq2, we seek to select a small number of subjects with minimal IBD sharing. SampleSeq2 also can be used to estimate the number, G_T , of founder genomes in a sample or to select the minimum number of subjects that will carry a target G_T . The algorithm also can be used for simultaneous selection of subjects from multiple groups. By simulation and using the Amish Anabaptist genealogy, we demonstrate that compared to a random selection of subjects, SampleSeq2 lead to an increase in G_T , a narrower region of IBD sharing at disease locus among cases, and a higher ancestral allelic diversity.

METHODS

When whole genome genotype data are available, the fractions of autosomes with IBD = 0, 1, 2 between two subjects can be accurately estimated using programs such as Plink [Purcell, et al. 2007] or KING [Manichaikul, et al. 2010]. Let $p_j = \Pr(\text{IBD} = j)$. If one subject is already selected, the other subject will contribute $d = p_0 + 0.5p_1$ extra founder genomes. The value d can be viewed as the “distance” between the two subjects. The less genome shared by two subjects, the more “distant” they are. The value is one if two subjects are unrelated (i.e., $p_0 = 1$ and $p_1 = p_2 = 0$) and zero if the second subject does not provide any additional founder genomes (i.e., $p_0 = p_1 = 0$ and $p_2 = 1$). We note that $d = 1 - s$, where $s = 0.5p_1 + p_2$ is the fraction of genome shared IBD by the subjects.

If there is not enough genotype information for accurate estimation of the fractions of the autosomes shared IBD between two subjects but their relationship is available, we can calculate their kinship coefficient, Φ , which is the probability that a gene selected randomly from one subject and a gene selected randomly from the same locus of the other subject are IBD [Lange 1997]. The kinship coefficient is a function of nine condensed identity coefficients, which can be calculated for any pair of subjects using programs such as IdCoefs [Abney 2009]. For an outbred population, it has a simple relationship with expected IBD probabilities: $\Phi = 0.25p_1 + 0.5p_2 = s/2$, where $s = 0.5p_1 + p_2$ is the expected fraction of genome shared IBD. Thus we can define the distance between two subjects as $d = 1 - s = 1 - 2\Phi$. Although it is possible to have $\Phi > 0.5$ and thus $d < 0$, this happens only when the two subjects are members of a highly inbred family. Even if this occurs, our method will not be affected as we seek to maximize the distances among the selected subjects.

We notice that if both genetic and pedigree data are available, genetic information tends to provide a more correct estimate of IBD sharing than solely relying on pedigree structure and the assumption of independent founders.

Estimation of Total Number of Founder Genomes

Our goal is to select subjects that have as many founder genomes as possible. The number of subjects, K , may be predetermined based on the budget of a study or a power calculation. For a set of K subjects, their total number of founder genomes is

$$G_T = K - \sum_{i < j} s_{ij} + \sum_{i < j < k} s_{ijk} - \cdots + (-1)^{K-1} s_{12 \dots K}, \quad (1)$$

where s_E is the fraction of the genome shared by all subjects in E . For two subjects i and j , $s_{ij} = 1 - d_{ij}$. For three or more subjects, s_E may be approximately estimated as $(\sum_{i \in E} \prod_{j \in E, j \neq i} s_{ij})/|E|$. As most s_{ij} are close to zero in practice, their products become negligible as the number of items in the product increases. Let $K_1 = K$, $K_2 = K - \sum_{i < j} s_{ij}$, and $K_m = K_{m-1} + (-1)^{m-1} \sum_{E: |E|=m} s_E$ for $m \geq 2$. This series converges to G_T in a zigzag way, that is, $K_2 < G_T$,

$K_3 > G_T$, $K_4 < G_T$, etc. We found stopping at K_8 and averaging K_7 and K_8 gives a good approximation of G_T for $K = 50$.

We seek to select K subjects with maximum G_T . For computational simplicity, our algorithm will be guided by its first-order approximation,

$$K_2 = K - \sum_{i < j} (1 - d_{ij}) = \sum_{i < j} d_{ij} - K(K - 3)/2. \quad (2)$$

This is equivalent to selecting subjects to maximize $\sum_{i < j} d_{ij}$, their total pairwise distance (TPD).

The Algorithm

Our algorithm starts with a random selection of K subjects. In each of the subsequent iterations, a subject is replaced by another one so that the TPD among the selected subjects is increased. This process is carried out until no replacement can further increase the TPD. Specifically, we propose the following algorithm:

1. Select K subjects randomly as a starting set. Set $r = 1$.
2. For each subject in the set, calculate its total distance from the other $K - 1$ subjects.
3. The subject with the r -th smallest total distance, d_r , is the candidate for removal.
4. For each subject not in the set, calculate its total distance with the remaining $K - 1$ subjects in the set. The one with the highest total distance, d_h , is the candidate for replacement.
5.
 - A. If $d_h > d_r$, make replacement, set $r = 1$, and go to Step 2;
 - B. If $d_h = d_r$ and $r < K$, set $r = r + 1$ and go to Step 3;
 - C. If $d_h = d_r$ and $r = K$, record the set and its TPD.
6. Repeat Steps 1–5 multiple times (default 10) and output the set with the highest TPD.

Step 5B ensures that if no replacement can be made then pick the subject with the next smallest total distance from the others in the current set as the candidate for removal.

This algorithm is a greedy algorithm. It guarantees that the TPD of the set always increases, but it may not guarantee the final set has the highest TPD among all possible subsets of K subjects. To ensure we reach the maximum TPD, Steps 1–5 need to be repeated a few times with different starting sets. We show that repeating the algorithm 10 times will be enough, at least for the situations we evaluated, to ensure the selected set has practically achieved the maximum TPD.

This algorithm can also be used to determine the number of subjects needed to achieve a pre-specified number, G , of founder genomes. This is achieved by running the program for a range of K and calculating the corresponding maximum G_T . The smallest K with $\max(G_T) > G$ is the minimal number of subjects to provide at least G founder genomes.

Simultaneous Selection from Multiple Groups

In some studies, investigators may want to select subjects from multiple groups, hoping to maximize the distance among the selected subjects within each group and simultaneously the distance between groups. For example, one may want to select cases and controls from large case and control pools. In studies with many pedigrees, one may want to select one or more distantly related subjects from each pedigree while maximizing unrelatedness of the

selected subjects across pedigrees. Other examples include selecting subjects matched by covariates or subjects with extreme phenotypes.

The algorithm presented above can be easily extended to allow simultaneous selection of subjects from multiple groups. The starting set will be generated in Step 1 to meet the target number of subjects for each group. When a candidate is nominated for removal in Step 3, only subjects within the same group will be considered as potential replacements in Step 4. This procedure allows for the algorithm to approach the maximum TPD simultaneously within and between groups.

RESULTS

The Amish Anabaptist Genealogy

The Amish pedigree has 4,995 subjects with 827 having DNA samples available for this example [Agarwala, et al. 2003]. We calculated the kinship coefficient using IdCoefs [Abney 2009] and expected fraction of genome shared IBD for every pair among the 827 subjects (Table I). The average pairwise IBD sharing was 2.4%, resulting in average pairwise distance 0.976. Among the total of 341,551 subject pairs, 283,851 (83.1%) pairs have >1% genome expected to be shared IBD, 18,007 (5.3%) pairs have >5% genome shared IBD, and 3702 (1.1%) pairs have >10% genome shared IBD. There are 810 pairs of subjects that are expected to share >50% of the genome.

We compared our approach with the alternative approach of random selection for selecting K subjects ($K = 50, 100$) from the Amish data. We simulated founder chromosomes and let them drop through the pedigree with recombination at each transmission, such that all members of the final generation were mosaics of the founder chromosomes. For each approach, we repeated the selection 20 times and calculated the average G_T and standard deviation (SD).

Using the Amish genealogy, for $K = 50$, the average G_T was 42.2 (SD = 0.02) using our algorithm and 32.1 (SD = 0.83) for random selection. On average, our algorithm resulted in 11.1 extra founder genomes, a 35% increase. While random selection resulted in subsets with 2.38% average pairwise IBD sharing, our method resulted in only 0.79% average pairwise IBD sharing, a 67% reduction. The SampleSeq2 sample would have 80% power to detect one or more copies of an allele with a frequency of 1.9%, and the random sample would have 80% power to detect one or more copies of an allele with a frequency of 2.5%, a 23% decrease in the detectable allele frequency. For $K = 100$, the average G_T was 66.8 (SD = 0.04) using our algorithm and 49.2 (SD = 0.98) for random selection. On average, our algorithm resulted in 17.6 extra founder genomes, a 36% increase. The average pairwise IBD sharing was 2.40% for random selection and 1.11% among subjects for our method (a 54% reduction). The SampleSeq2 sample would have 80% power to detect one or more copies of an allele with a frequency of 1.2%, and the random sample would have 80% power to detect one or more copies of an allele with a frequency of 1.6%, a 25% decrease in the detectable allele frequency.

Simulated Populations

We simulated two island populations and a relatively outbred population. The island populations grew from an outbred founder group of 500 individuals [Bonnen, et al. 2006; Kenny, et al. 2011]. A relatively outbred simulation (S1) began 100 generations ago to a current population of size about 250,000 with a growth rate per generation of $r = 6.4\%$. The other simulation (S2) was more like the Amish data and began 30 generations ago to a current population size of about 10,000, with a growth rate per generation of $r = 10.5\%$ (Table II). In both simulations, we assumed the whole genome was 3,000 cM and the

expected number of recombinations per meiosis followed a Poisson distribution with mean 30. For ease of computation, we assumed no interference and recombination occurred at breakpoints of 1 cM segments. With this strategy, the whole genome consisted of 3,000 segments that could be traced along generations. In each generation, we assumed random mating and for each couple, their number of children followed a Poisson distribution with mean $2(1+r)$.

For each person in the final generation we knew the mosaic configuration of ancestral segments. We calculated pairwise IBD accordingly. For any subset of subjects, we could tally the number of independent copies for each segment and thus the true G_T . This allowed us to check the accuracy of our approximation of G_T that is solely based on pairwise IBD. The approximation was quite good, although it did underestimate the true value of G_T as the sample size grew larger.

In S1, we established a fully penetrant dominant allele at an arbitrary segment of the genome to identify cases and controls and to observe the average size of the region around the disease allele shared IBD among cases. For a selection of $K = 50$ cases, the average size of the region shared IBD was 2.28 cM (SD = 0.16) for SampleSeq2 and 2.87 cM (SD = 0.24) for random subjects, a 21% decrease (Figure 1). For a selection of $K = 100$ cases, the average size of the region shared IBD was 2.37 cM (SD = 0.08) for SampleSeq2, and 2.93 cM (SD = 0.15) for a random draw, a 19% decrease (Figure 1).

In S2, the average pairwise IBD sharing was 2.0%, resulting in average pairwise distance of 98.0%. Unlike the Amish sample, there were fewer pairs of subjects with high IBD sharing. About 97.2% pairs had 1–3% of the genome shared IBD, and only 2.4% pairs had >3% genome shared IBD, and 0.6% pairs had >4% genome shared IBD. Although this simulated population was more outbred than the Amish sample, they were nonetheless from an isolate population for which SampleSeq2 is expected to help increase the number of founder genomes in the selected subjects. We compared our approach with random selection for selecting K subjects ($K = 50, 100$) from the current generation in S2. For $K = 50$, the average G_T was 35.2 (SD = 0.04) using SampleSeq2 and 33.5 (SD = 0.13) for random selection. On average, our algorithm resulted in 1.7 extra founder genomes, a 5.1% increase. Our algorithm would have 80% power to observe one or more copies of an allele with a frequency of 2.3%, while random draw would provide power to detect an allele with 2.4% frequency. While random selection resulted in subsets with 1.99% average pairwise IBD sharing, our method resulted in only 1.62% average pairwise IBD, an 18% reduction (Figure 2). For $K = 100$, the average G_T was 52.8 (SD = 0.05) using SampleSeq2 and 50.0 (SD = 0.17) for random selection. On average, our algorithm resulted in 2.8 extra founder genomes, a 5.6% increase. The average pairwise IBD sharing was 2.0% for random selection and 1.7% for our method (a 15% reduction). Our algorithm would have 80% power to observe one or more copies of an allele with a frequency of 1.5%, while random draw would provide power to detect an allele with 1.6% frequency.

We also evaluated simultaneous selection of K cases and K controls ($K = 50, 100$) from the subjects in S2 using our algorithm described in Methods, and compared it with separate application of SampleSeq2 to select the same numbers of cases and controls. As expected, simultaneous selection resulted in slightly lower TPD among the selected cases and the selected controls, but relatively higher distance between the two sets.

To evaluate SampleSeq2 on a more outbred population, we simulated population S3 that started with 10,000 founders 100 generations ago and grew to a current population of size about 100,000 with a growth rate per generation of $r = 2.3\%$ (Table II). Due to computational limitations, we could only simulate a 300 cM chromosome instead of the

whole genome. Similar to S1, we established a fully penetrant dominant allele at an arbitrary segment to identify cases. For this population, SampleSeq2 continued to outperform random selection, although the magnitude of improvement was smaller than for an inbred population. For $K = 50$, the average G_T was 43.84 (SD = 0.35) using SampleSeq2 and 43.65 (SD = 0.49) for random selection. For $K = 100$, the average G_T was 78.18 (SD = 0.64) using SampleSeq2 and 77.79 (SD = 0.60) for random selection (Figure 3).

Association Study Power

We further evaluated the power of detecting association when SampleSeq2 was used for selecting cases and controls. In our island population S2, we picked a site as the disease locus and certain founder alleles as risk alleles, with risk allele frequency 11.4%. We used an additive disease model with penetrance 0.05 for subjects with no risk allele, 0.10 for those with one risk allele, and 0.15 for those with two risk alleles. We then selected four subsets of size K ($K = 50, 100$): 1) K cases by SampleSeq2, 2) K cases by random selection, 3) K controls by SampleSeq2, and 4) K controls by random selection. We performed four chi-squared association tests for the following four scenarios: i) between 1) and 3), ii) between 1) and 4), iii) between 2) and 3), and iv) between 2) and 4). This was repeated 100 times to evaluate their nominal power at significance level 0.05.

The power was the highest ($K = 50$, power = 0.72; $K = 100$, power = 0.82) when SampleSeq2 was used for selecting both cases and controls, and the lowest ($K = 50$, power = 0.47; $K = 100$, power = 0.47) when random selection was used in both selections. The power was ($K = 50$, power = 0.53; $K = 100$, power = 0.67) for scenario ii) and ($K = 50$, power = 0.63; $K = 100$, power = 0.72) for scenario iii). We further evaluated the reason behind the power improvement of SampleSeq2 over random selection. Figures 4 and 5 show the risk allele frequencies for the four selections. Subsets selected through SampleSeq2 consistently had less sampling variation than those randomly selected. As subsets selected through SampleSeq2 achieve maximum representation of ancestral chromosomes, there probably is less variation from one subset to another than those selected randomly.

Number of Repeats in SampleSeq2

By default our algorithm is repeated 10 times and the subset with the highest TPD is reported. To evaluate whether this is enough, we ran the algorithm at multiple repeat values from 5 to 1000 using the Amish pedigree and our simulated island population S2. In each setting we selected 20 subsets of $K = 50$ subjects and calculated the K_2 for each subset using the approximation (2) (Table III). For the Amish pedigree, we always reached the maximum K_2 even after repeating our algorithm 5 times. For the simulated island population, the average K_2 increased as number of repeats increased, and repeating our algorithm 10 times allowed us to capture 99.6% of the K_2 achieved after repeating 1000 times. Therefore, running the algorithm 10 times practically ensured the maximum TPD was achieved.

DISCUSSION

We have introduced a distance measure that reflects the level of IBD sharing between a pair of subjects. It can be calculated using either GWAS data or pedigree information. Using the distance measure, we proposed an algorithm for selecting a small sample of maximally unrelated subjects from a larger set of available participants. This algorithm can also be used to select a minimal number of subjects to reach a target number of founder chromosomes, minimizing cost of sequencing to obtain the target sample size.

Some circumstances for which our method may be applied are SNP discovery efforts, the aforementioned case-only IBD design, a traditional case-control association design, or

selecting one or more distantly related subjects from each pedigree while maximizing unrelatedness of the selected subjects across pedigrees. The case-only IBD design benefits from SampleSeq2 by reducing or eliminating regions shared IBD that are unrelated to the trait, and minimizing the region around the determinant allele that is shared IBD. This will decrease the cost and scope of fine-mapping those regions for association with traits. In a case-control study, the variance inflation conferred onto test statistics by cryptic relatedness is mitigated, and the effective sample size of independent subjects is increased, both resulting in increased study power. For SNP discovery studies, because SampleSeq2 selects a subset of subjects maximizing the number of sampled ancestries, alleles that are private to a given ancestry are more likely to be represented in the selected sample compared to a random draw, improving the representation of extant alleles in the sample.

The magnitude of the benefits provided by SampleSeq2 depends on the properties of the larger cohort from which a sample will be drawn, and the size of the sample drawn from those participants. More inbred cohorts or populations will be most improved, while outbred cohorts will enjoy fewer advantages. Similarly, the most benefits are observed for smaller sample sizes, with decreasing advantages as the sample gets large. Many epidemiologic cohorts are collected without regard to genetic relationships among subjects within communities where several extended families may reside, and so this approach can assist in avoiding or mitigating effects of cryptic relatedness in such collections.

The estimator we derived for the number of founder genomes is effective for small numbers of subjects, but loses efficiency as the sample size grows, systematically underestimating G_T . This property does not affect the ability of the method to select an optimally unrelated subset of arbitrary size, but does affect the ability of the method to identify the smallest number of subjects that satisfy the target sample size. Improvements to this estimator are necessary; however, in the current form it can still offer substantial advantages to studies with small samples.

Our method is an effective approach for optimizing the selection of subjects for sequencing studies, whether they are small studies intending to identify regions shared IBD among cases, or larger studies designed to perform association analysis in the sequencing data. The benefits for larger association studies are smaller than for smaller SNP discovery designs; however, there is no cost associated with these small advantages. This method can be applied to either inbred or outbred populations, and works synergistically with existing analysis methods and study designs to improve study performance. While the benefits of using SampleSeq2 in outbred samples of subjects are modest, even those modest benefits come without costs to analytic flexibility, power, statistical validity, or economic resources. Thereby, we offer the investigator a “free lunch”, where use of our method does not preclude the use of other methods or designs, and may improve the likelihood of discoveries and study success. The SampleSeq2 software is available at <http://biostat.mc.vanderbilt.edu/SampleSeq>

Acknowledgments

Funding: This work was funded in part by NIH grant R01HG004517 (to TLE and CL) and Vanderbilt Clinical and Translational Research Scholar award 5KL2RR024975 (to TLE).

Reference List

- Abney M. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*. 2009; 25:1561–1563. [PubMed: 19359355]
- Agarwala R, Biesecker LG, Schaffer AA. Anabaptist genealogy database. *Am J Med Genet C Semin Med Genet*. 2003; 121C:32–37. [PubMed: 12888984]

- Albrechtsen A, Moltke I, Nielsen R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*. 2010; 186:295–308. [PubMed: 20592267]
- Bonnen PE, Pe'er I, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly MJ, Reich DE, Jones KW, Stoffel M, Altshuler D, Friedman JM. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet*. 2006; 38:214–217. [PubMed: 16429162]
- Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. *Am J Hum Genet*. 2011; 88:173–182. [PubMed: 21310274]
- Browning SR, Thompson EA. Detecting Rare Variant Associations by Identity by Descent Mapping in Case-control Studies. *Genetics*. 2012
- Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet*. 2004; 75:35–43. [PubMed: 15148658]
- Edwards TL, Song Z, Li C. Enriching Targeted Sequencing Experiments for Rare Disease Alleles. *Bioinformatics*. 2011
- Han L, Abney M. Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol*. 2011; 35:557–567. [PubMed: 21769932]
- Hoischen A, van Bon BW, Gilissen C, Arts P, van LB, Steehouwer M, de VP, de RR, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, Veltman JA. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet*. 2010; 42:483–485. [PubMed: 20436468]
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009; 5:e1000529. [PubMed: 19543373]
- Kenny EE, Kim M, Gusev A, Lowe JK, Salit J, Smith JG, Kovvali S, Kang HM, Newton-Cheh C, Daly MJ, Stoffel M, Altshuler DM, Friedman JM, Eskin E, Breslow JL, Pe'er I. Increased power of mixed models facilitates association mapping of 10 loci for metabolic traits in an isolated population. *Hum Mol Genet*. 2011; 20:827–839. [PubMed: 21118897]
- Lalonde E, Albrecht S, Ha KC, Jacob K, Bolduc N, Polychronakos C, Dechelotte P, Majewski J, Jabado N. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum Mutat*. 2010; 31:918–923. [PubMed: 20518025]
- Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*. 1987; 236:1567–1570. [PubMed: 2884728]
- Lange K. *Mathematical and Statistical Methods for Genetic Analysis*. 1997
- Leibon G, Rockmore DN, Pollak MR. A SNP streak model for the identification of genetic regions identical-by-descent. *Stat Appl Genet Mol Biol*. 2008; 7:Article16. [PubMed: 18518857]
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010; 34:816–834. [PubMed: 21058334]
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010; 26:2867–2873. [PubMed: 20926424]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007; 39:906–913. [PubMed: 17572673]
- McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am J Hum Genet*. 1998; 63:241–251. [PubMed: 9634509]
- Moltke I, Albrechtsen A, Hansen TV, Nielsen FC, Nielsen R. A method for detecting IBD regions simultaneously in multiple individuals--with applications to disease genetics. *Genome Res*. 2011; 21:1168–1180. [PubMed: 21493780]
- Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. Exome sequencing identifies

- MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet.* 2010; 42:790–793. [PubMed: 20711175]
- Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009; 461:272–276. [PubMed: 19684571]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–575. [PubMed: 17701901]
- Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie FV, Mansour S, Holder SE, Brain CE, Burton BK, Kim KH, Pauli RM, Aftimos S, Stewart H, Kim CA, Holder-Espinasse M, Robertson SP, Drake WM, Trembath RC. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. *Nat Genet.* 2011; 43:303–305. [PubMed: 21378985]
- Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann Hum Genet.* 2008; 72:279–287. [PubMed: 18093282]
- Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.* 2005; 1:e32. [PubMed: 16151517]
- Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol.* 2010; 34:171–187. [PubMed: 19847924]

Fig. 1. Boxplots for the length of region shared IBD around a rare causal allele among cases obtained using SampleSeq2 and by random selection for sample sizes of 50 and 100.

Fig. 2. Boxplots for the number of founder genomes selected using SampleSeq2 and by random selection for sample sizes of 50 and 100 from an island population (S2).

Fig. 3. Boxplots for the number of founder genomes selected using SampleSeq2 and by random selection for sample sizes of 50 and 100 from an outbred population (S3).

Fig 4. Distribution of risk allele counts for: 1) 50 cases through SampleSeq2, 2) 50 cases through random selection, 3) 50 controls through SampleSeq2, and 4) 50 controls through random selection.

Fig 5. Distribution of risk allele counts for: 1) 100 cases through SampleSeq2, 2) 100 cases through random selection, 3) 100 controls through SampleSeq2, and 4) 100 controls through random selection.

Table I

Expected pairwise IBD sharing from the Amish based on pedigree information from the Anabaptist Genealogy Database.

Expected fraction of genome shared IBD	Number of pairs	Percent of total
>0%	341,551	100.0
>1%	283,851	83.1
>5%	18,007	5.3
>10%	3,702	1.1
>50%	810	0.2

Table II

Design and observations from simulation experiments.

	Simulation		
	S1	S2	S3
Starting population size	500	500	10000
Final population size	250,000	10,500	101,511
Growth rate	6.2%	10.1%	2.3%
Number of generations	100	30	100
Average pairwise IBD sharing	2.9%	2.0%	0.47%
Percent of pairs sharing:			
<1% IBD	0%	0.35%	92.89%
1–2% IBD	0.02%	55.39%	6.52%
2–3% IBD	65.62%	41.72%	0.43%
3–4% IBD	34.30%	1.95%	0.08%
>4% IBD	0.05%	0.59%	0.08%

Table III

Average K_2 and standard deviation (SD) for subsets of 50 subjects selected after repeating our algorithm at various times.

Number of repeats	Amish pedigree		Simulated population S2	
	Average K_2 (SD)	Percent of highest K_2 (%)	Average K_2 (SD)	Percent of highest K_2 (%)
5	40.339 (0)	100	30.072 (0.054)	99.58
10	40.339 (0)	100	30.075 (0.045)	99.59
50	40.339 (0)	100	30.128 (0.029)	99.76
100	40.339 (0)	100	30.157 (0.023)	99.86
500	40.339 (0)	100	30.196 (0.025)	99.98
1000	40.339 (0)	100	30.200 (0.018)	100