Optimizing strain selection for association studies under hard cost constraints

Christoph D. Rau[1], Patrick H. Bradley[2]

1: Department of Genetics and Computational Medicine Program, University of North Carolina at Chapel Hill

2: Department of Microbiology, The Ohio State University

Abstract

Quantitative genetics methods can be particularly powerful in model organisms and non-human populations, and we now have strain collections such as recombinant inbred lines, etc. that can be phenotyped. Natural diversity is also valuable in non-model systems that do not yet have reverse genetic tools. However, purchasing and phenotyping large collections can be cost-prohibitive. Strain or sample acquisition costs may also vary dramatically for different strains or isolates. Thus, investigators need efficient strategies to optimize experimental power for a given limited budget. In this study, we evaluate several approaches to optimally select subsets of the total cohort to best maintain power when performing genome-wide association studies. Some approaches focus solely on costs, others on genetic diversity, and some on both simultaneously.

Through simulation studies across different minor allele frequencies and SNP effect sizes, we demonstrate that selecting for cost is most beneficial at low-to-moderate budget thresholds, while selecting for diversity is optimal in scenarios involving rare (MAF 5-10%) variants or higher total costs, or when accounting for additional costs per strain studied. We also evaluate these approaches on data from the Hybrid Mouse Diversity Panel (HMDP), and find that an approach that considers both cost and diversity is superior at recovering significant loci and maintaining statistical power under real-world conditions. This approach picks the strains that, for a given budget, minimize the total genetic distance to the strains that were not selected. This approach, which we term “ThriftyMD” (for “Thrifty Minimum Distance”), extends previous distance-based methods to pick a representative panel by explicitly adding a cost constraint.

Overall, our results highlight the trade-offs between cost, diversity, and power in GWAS cohort design, and present the ThriftyMD algorithm as a versatile and robust approach for optimizing study design in resource-limited settings.

INTRODUCTION

The cost of research has grown significantly over the past two decades. One measure is the average size of an R01 grant from the National Institutes of Health, which has increased 87% between 1999 and 2019[1], 64% faster than inflation. These cost increases come from many factors: novel techniques and technologies that supplant other, often cheaper approaches (e.g. PCR to RNA microarrays to RNAseq to single cell RNAseq), more complex study designs which necessitate additional research cohorts and endpoints, and, notably, a drive to increase the size of studies to attempt to both achieve greater statistical power and identify less penetrant drivers of disease conditions[2-4]. For example, over the past two decades, the GWAS community has been driven to search for larger and larger populations to try and recover the effects of polymorphisms with smaller and smaller effect sizes[5,6]. Depending on how they are funded, such large populations may be out of reach for individual investigators. Furthermore, beyond simple financial considerations, researchers who use animal populations are often mandated to follow the three Rs (Replacement, Reduction, Refinement) in their research, which states that if it is possible to achieve the same or similar result with fewer animals, then researchers should make every effort to do so. Taken together, the increasing costs in both money and time of doing research coupled with ethical responsibilities to animal cohorts are powerful incentives to optimize and streamline research projects, as long as this does not come at the expense of statistical power.

One opportunity for this type of streamlining concerns how we select strains to phenotype. Diverse strain collections allow investigators to link genotypes and phenotypes through tools such as GWAS, MPRA assays, and global 'omics analyses[7-12]. However, prices vary widely across strain collections, individual genotypes, and even how a particular strain is propagated and shipped. For example, a mouse from Jackson Labs costs a different amount of money depending on its strain, age, and sex, ranging from less than $30 for a 3-week-old C57BL/6J male, to over $260 for a 9 week old NZO/HILtJ male, to nearly $2,500 a litter for one of the many cryopreserved strains available for purchase. Similar factors apply even to bacterial strains: the type strain of *Lachnospira eligens*, shipped as lyophilized cells, costs $474 from the American Type Culture Collection (ATCC) but 100 EUR (~$113) from the German Collection of Microorganisms and Cell Cultures (DSMZ); a live, active culture from DSMZ would cost 240 EUR ($273). Both ATCC and DSMZ, as well as many other repositories, have unique strains that can only be purchased from them. Being able to optimize the selection of strains to maximize statistical power for a given budget would be a major benefit for the labs which use these populations.

Previous work on the strain selection problem has tended to assume equal costs for all strains. One well-researched approach, related to the more general "Noah's Ark" problem in bioconservation[13], seeks to maximize total diversity, most famously measured by Faith's phylogenetic diversity (PD)[14]. (A related alternative is to maximize the minimum pairwise distance between selected taxa[13].) Assuming equal costs, greedy algorithms (as in Faith 1992), where successive strains are chosen iteratively based on which additional strain leads to the the largest increase in diversity, have been proven to reach global optima[15,16]. Greedy methods to maximize PD have also been applied in genotype imputation studies, where they were shown to improve power relative to random selection[17,18]. However, modifying these methods to account for budgetary constraints is not straightforward. One approach is to weight the diversity metric itself by the cost of a new strain, in which case the greedy algorithm is still guaranteed to be optimal[15], but this requires researchers to pick a cost-weighting scheme, and the effects of different

schemes on power in downstream analyses have not, to our knowledge, been systematically evaluated. In the more intuitive fixed-budget setting, greedy algorithms are only guaranteed to pick the best strains when costs are equal; the impact of unequal costs on its real-world performance are not yet known.

Another potential issue is that maximizing diversity *per se* may not always maximize power in downstream analyses[19], and can lead to the selection of outliers that are not very representative of the most common variants[20]. One alternative is to pick the strains or sequences that are the most genetically representative of the larger panel. To accomplish this, Matsen IV et al[20] developed an approach that minimizes the average phylogenetic distance of query sequences to the closest leaf on a reference tree ("minimum-ADCL"). This problem is conceptually very similar to k-medoids and what Faith refers to as discrete p-median diversity[21], as well as classic 0/1-optimization problems. Kang et al adapted Matsen's min-ADCL approach to pick reference panels for genotype imputation, and showed that this approach outperformed both random selection and maximizing total diversity[22]. Similarly to the maximum-diversity approach, however, the minimum-ADCL method does not consider costs, meaning it would have to be modified for the question we are interested in. Additionally, the minimum-ADCL and maximum-diversity methods could potentially favor common or rare variants, respectively.

In this manuscript, we set out to study different approaches for strain selection under a hard cost constraint, and to directly test how they would affect our power to detect genetic associations. We apply four such methods:

- RandomCost: Selecting strains at random until the budget is exhausted;
- MinCost: Selecting the cheapest remaining strains until the budget is exhausted, to maximize the number of strains chosen;
- MaxDiversity: Selecting strains using a greedy algorithm that maximize diversity at every step until the budget is exhausted; and
- ThriftyMD: Selecting strains using integer linear programming (ILP) that, for a given budget, minimize the genetic distance ("MD") to all non-selected strains.

The MaxDiversity and ThriftyMD approaches extend previous algorithms to add the hard budget constraint, which is novel to this paper. We evaluate these approaches both on entirely simulated trait data, as well as an actual mouse GWAS study of isoproterenol stimulation that we subsample to simulate subset selection. In our simulations, we vary the minor allele frequency (MAF) and effect size (χ^2), and also explore the impact of reducing the variance of prices by adding extra equal, fixed costs per strain (as might be appropriate when the cost of strain maintenance and/or phenotyping is significant).

Our results show that rather than a single method being clearly superior across all situations, MinCost, MaxDiversity, and ThriftyMD all had advantages and disadvantages. Surprisingly, the simple MinCost approach performed very well, especially for small budgets, where it was sometimes the most powerful of the four. MaxDiversity was favored when the trait depended on rare variants (MAF < 10%) and when the price variance was lower. MaxDiversity also tended to yield solutions with fewer total strains than other methods. ThriftyMD, conversely, was strongest when the trait depended on common variants (MAF > %%) and prices varied sharply across strains. Finally, ThriftyMD and MinCost both decisively outperformed MaxDiversity, which was the next-best method, on the experimental GWAS dataset. MinCost was the

winner at the lowest budgets, while ThriftyMD performed better at medium-to-high budgets, indicating a potential advantage in real-world scenarios. At half the cost, ThriftyMD recovered the highest fraction of SNPs that were significant in the full panel, more than twice as many as MaxDiversity.

METHODS

Data Sources

Mouse genotype data from the Mouse Diversity Array for the Hybrid Mouse Diversity Panel (HMDP) was obtained from the Mouse Genomics Informatics website[11]. Mouse phenotype data was pulled from prior work[12] and available at Mendeley Resources at accession DOI:10.17632/y8tdm4s7nh.1[23]. Mouse costs were pulled from the Jackson Labs website[24] as on 4/05/2024 and represent the cost of 1 male and 2 female mice or one litter of cryopreserved animals, depending on the current status of the animals at the time of accession.

Simulation Studies

Phenotype Creation: We created sets of simulated phenotypes for our simulation studies. For each simulated phenotype, we drew a SNP at random that matched a required Minor Allele Frequency (MAF) criterion (5-10%, 10-30% or 40-50%). We then selected a β value for this SNP based on a desired χ^2 threshold (10, 50 or 100) and added noise based on a multivariate normal (MVN) derived from two variance terms, the first the proportion of variance attributable to genetic effects (σ_g^2), set at 40% consistent with our past work[25] and the second the proportion attributable to all other sources of variation (σ_e^2) as follows:

$$y = \beta X + MVN(0, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})$$

Where **K** is the kinship matrix of the strains of the HMDP calculated using the kinship function of the myFastLMM R package[26].

GWAS Approach: Simulated phenotype-genotype pairs were analyzed using the linear mixed model approach fastLMM[26]. Power was calculated based on the number of associations that met a genome-wide significance threshold of 4.2E-6 as previously determined in the HMDP[11,27-29]. At a χ^2 of 100, there was 100% power in the full panel)(Figure S1A). In the χ^2 of 50 set, power ranged from 82-93% depending on the Minor Allele Frequency (MAF)(Figure S1A). Power in the χ^2 of 10 set of simulated phenotypes was very low (<10%) in the whole panel and was not used going forward.

Strain Subset Selection:

We selected strain subsets for our simulation study as follows:

<u>Random Strain:</u> A random set of strains for each phenotype were selected from all possible strains in intervals of 5 strains between 0-175.

<u>RandomCost:</u> For costs at intervals of $20,000 from $20,000 - $240,000, random strains were selected for each phenotype such that the sum of values for each strain was less than the dollar limit.

MinCost: For costs at intervals of $20,000 from $20,000 - $240,000, strains were selected in order from cheapest to most expensive until the cost limit was reached. As this is fixed based on the cost, all phenotypes had the same strain list.

MaxDiversity: MaxDiversity uses a greedy algorithm where the "best" strain is selected at each step regardless of cost. For costs at intervals of $20,000 - $240,000, each strain with initial cost less than $1,000 was used as a seed. Next, the strain whose genetic distance is the most significantly distinct from the seed mouse was added to the selected strains and costs updated. Finally, the strain whose genetic distance was most distinct from that pair of mice was added and costs updated. This process continued until the cost limit was reached. The results were then upsampled to meet the number of simulations per χ^2/MAF combo.

ThriftyMD: We formulated this problem to be similar to minimum-ADCL[20] and k-medoids, with an additional constraint on total cost and no predetermined number k of representatives. Thus, we were interested in picking k medoids or representative strains with minimum distance to the closest non-chosen strains for each representative, with k allowed to vary between zero and one.

Let *n* be the total number of strains in the collection, $d_{i,j}$ be the genetic distance between strains *i* and *j*, and c_i and c_{tot} the cost of strain *i* and the total budget, respectively. To perform the optimization, we first define two binary variables, x_i and $y_{i,j}$. The variable x_i represents whether or not strain *i* has been chosen as a representative, while the variable $y_{i,j}$ represents whether or not strain *i* has been chosen to represent strain *j*. ThriftyMD then minimizes the following sum:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} d_{i,j}y_{i,j}\,,\ i,j \in \{1..n\}$$

subject to the following constraints:

$$\sum_{i=1}^{n} y_{i,j} = 1, \qquad j \in \{1..n\}$$

$$0 \le x_i - y_{i,j} \le 1, \qquad i,j \in \{1..n\}$$

$$x_i, y_{i,j} \in \{0,1\}, \qquad i,j \in \{1..n\}$$

$$\sum_{i=1}^{n} x_i c_i \le c_{tot}, \qquad i \in \{1..n\}$$

$$0 \le \sum_{i=1}^{n} x_i \le n, \qquad i \in \{1..n\}$$

The first constraint guarantees that exactly one "nearest" representative is assigned to each of the full set of strains. The second guarantees that $y_{i,j}$ is never 1 when x_i is not also 1: this in turn means that only distances from a representative to its assigned strains will be summed in the objective function. The third guarantees that *x* and *y* are binary, the fourth gives the cost constraint, and the fifth allows for any number of strains *k* ($k = \sum x_i$) to be selected. Strains that must be selected (for example, strains that are already

part of a researcher's collection) can also be included as additional constraints on the value of particular x_i.

The method was implemented in R, using the `rcbc` package[30] which provides bindings to the CBC solver[31]. We picked sets at intervals across a total budget range of $20,000 - $240,000.

Validation Approaches

HMDP: The RV phenotype after isoproterenol stimulation was selected as it had the largest number of genome-wide significant loci (Figure 3A)[12]. The mice used in the selected study were not the entire panel, but instead comprised 93 strains for a total cost of $119,241. Costs were set at $20,000 intervals from $0 to $100,000 and 10 subset selections identified per approach. Included SNPs were recalculated for each sample due to changes in MAF across the population due to removal of strains. Phenotypes were then subset and GWAS run through fastLMM followed by analysis of how many significant SNPs were maintained after subsetting as well as how many SNPs which had been present in the original panel were lost due to shifts in MAF.

RESULTS

Selection Algorithms

We sought to determine which of four competing selection algorithms would be best able to maintain power while cutting the costs associated with obtaining a diverse set of samples for a GWAS. These algorithms are as follows:

Random: Strains are selected at random until a defined cost threshold is reached.

MinCost: Strains are selected from cheapest to most expensive until a defined cost threshold is reached.

MaxDiversity: A random strain with low cost (<$200 per mouse) is selected. Select subsequent strains such that they maximize the genetic distance between the already selected strains and the new strain regardless of cost, stopping when you reach a defined cost threshold.

ThriftyMD: Start with a random low-cost (<$200 per mouse) strain. Select the additional number strains that minimize the genetic diversity to non-selected strains, subject to a defined constraint on the total sum of strain costs.

Simulation Studies

We began our analysis by examining the effects of taking random sets of strains of mice on the power of the GWAS analysis. The HMDP consists of approximately 175 strains of mice and, at our selected χ^2 and MAF thresholds, we observe approximately 50% power to detect a true positive result at ~60 strains at a χ^2 of 100 and a MAF of 10-30%, or 100-110 strains at a χ^2 of 50 at various MAF thresholds (Fig S1A). At χ^2=100, there is 100% power to detect phenotypes in the whole panel, while at χ^2=50, there is 82-93% power to detect a phenotype.

With our understanding of how random strain selection is likely to affect GWAS results in our panel, we set out to examine the behavior of each of our four selection models (*RandomCost, MinCost, MaxDiversity,* and *ThriftyMD)* at price limits ranging from $20k to $240k (approximately the cost of the

entire panel). We observe marked differences in the behavior of each selection approach depending on the strength of the SNP effect and the MAF. Broadly, *MinCost* and *ThriftyMD* tended to perform better than the other methods at low cost thresholds (e.g. less than $100,000). These methods, as they consider cost when selecting a new strain, reach a higher number of strains at lower cost thresholds (for example, both choose over 100 strains at $60,000 versus under 50 for both *MaxDiversity* and *RandomCost*) (Fig S1B).

At a SNP effect of χ^2=100, we observe at a cost threshold of $60,000 that both *MinCost* and *ThriftyMD* near 75% power, whereas *MaxDiversity* remains at ~45% power and *RandomCost* below 25% power (Fig 1A). As the cost threshold increases, however, all the methods quickly approach 100% power (Fig 1A) . At $100,000, we observe close to 90% power using *ThriftyMD* and *MinCost*, while *MaxDiversity* is ~80% and *RandomCost* Is ~70%. (Fig 1A) At $160,000, the methods are indistinguishable from one another.

At a weaker SNP effect of χ^2=50, we observe that minor allele frequency (MAF) plays a major role in determining the efficacy of different approaches. At MAFs of 40-50%, we observe that *RandomCost* is, strikingly, a more effective choice than *MaxDiversity*, likely driven by a slightly higher number of strains and, consequently, power, compared to the other approach (Fig 1D). Both methods are still significantly less effective than *MinCost* or the *ThriftyMD* algorithms, which reach 70% power at $100,000 vs ~20% power for *MaxDiversity* (Fig 1D). At MAFs of 10-30%, we observe our weakest overall power, with <50% power at $100,000 for *MinCost* and *ThriftyMD*, 26% power for *MaxDiversity* and only 13% power for *RandomCost* (Fig 1C). At low MAF (5-10%), however, we observe that *MaxDiversity* now is the best choice for strain selection above $120,000 and roughly comparable with *MinCost* and *ThriftyMD* at the 80 or 100k thresholds (Fig 1B). Surprisingly, *MinCost* performed essentially as well as *ThriftyMD* in this setting.

Compared to *MinCost* or *ThriftyMD, MaxDiversity* selections always have fewer strains (e.g. 70 vs 120 at $100,000, Fig S1B). Despite this (and the resultant loss of power due to a smaller sample size), at times *MaxDiversity* exhibited comparable power to the other approaches due to its maximized genetic diversity. For example, at $120,000 for low MAF (5-10%) SNPs, *MaxDiversity* had better power (67% vs 54%) than *MinCost* or *ThriftyMD* while still using 37% fewer mice (Fig 1B).

In most cases, downstream experiments would require additional costs per animal/strain, so reducing the number of strains should lead to lower overall downstream costs. To simulate these additional costs (RNAseq, histology, experimental manipulation, etc) per strain used, we added $2,500 to the cost of each mouse strain to account for the increased costs inherent in using more animals. This, in essence, makes the variance of cost between mouse strains lower and results in less of a disparity in cohort size between the models (Fig S1C).

When we repeated our simulation studies with these additional costs, we observe that at χ^2 of 50 and MAFs from 40-50%, *MinCost* and *ThriftyMD* still maintained the highest power, although not to the same extent as previously observed when experimental cost per strain was not considered (Fig 2D). At χ^2=50 or 100 and MAF 10-30%, *MaxDiversity* and *ThriftyMD* are indistinguishable from one another, better than *MinCost* at higher cost thresholds, and consistently better than *RandomCost* (Fig 2C). Without the additional costs of research per strain, *MaxDiversity* was a comparable choice to the other approaches at low MAF (5-10%). With the additional costs, however, it becomes the clear winner in terms of recovering

power (e.g. 75% vs 55% power at 400k total costs) at most timepoints, with the other approaches only comparable at very low (<120k) or very high (>560k) total costs. (Fig 2B).

Validation

Finally, we applied our selection criteria to real data pulled from a prior HMDP study on cardiac dysfunction[12]. In particular, we examined the effects of strain selection on GWAS loci for right ventricle weight after catecholamine challenge, which had the largest number of genome-wide significant SNPs in the original study (Fig 3A). We first examined the effects of removing strains on the number of SNPs tested, as changes to the underlying strains can affect the SNPs selected for GWAS studies due to changing MAFs.

As expected, *MaxDiversity* did the best job of preserving the total number of SNPs tested, losing an average of only 557 SNPs at a 40k cost cutoff (0.24%) compared to 6,541-10,056 (2.9-4.4%) for the other approaches (Fig 3B). Both *MinCost* and *RandomCost* saw a number of significant loci from the original study removed as a result of changing MAF-associated thresholds (27, or 17% of the significant SNPs at a cutoff of 40k), although both *ThriftyMD* and *MaxDiversity*, as they account for genetic diversity to some extent in their selection process, lost very few significant loci (<5 at any threshold)

The *ThriftyMD* algorithm was by far the most successful at recovering the largest number of significant loci from the full study (45.1% at 60k vs 7.5% for the *MaxDiversity* algorithm) (Fig 3C). Surprisingly, the *MinCost* algorithm showed a consistent recovery at every timepoint of between 30-40% of the original significant SNPs (Fig 3C). We binned all SNPs into tiled 500kb windows across the genome to account for possible LD effects for SNPs in close proximity to one another, but observed largely identical results (Fig S2).

DISCUSSION

In this study, we set out to identify selection algorithms which could be used to minimize the cost of performing large experiments on samples with diverse genetic backgrounds while retaining overall power and significant SNP recovery. We discuss the results of each of our selected algorithms in turn.

RandomCost:

As perhaps expected, purely random selection of strains up to a fixed cost threshold proved to be the least effective of our four tested approaches, consistently returning the lowest power and eliminating the most significant SNPs at each threshold, with the exception of our simulation study of SNPs with MAF 40-50%, where it modestly outcompeted the *MaxDiversity* algorithm due to returning a slightly larger cohort on average (as *MaxDiversity* tends to pick up more strains that require recovery from cryopreservation and are, therefore, more expensive) while not losing much diversity since the underlying test SNP is already so common in the population.

MinCost:

Selecting simply the cheapest strains possible appeared to be a surprisingly good solution in many cases, consistently returning well-powered results comparable to the more sophisticated *ThriftyMD* selection algorithm. This is likely due to the power gained by using a larger number of inexpensive

strains. *MinCost* results in approximately twice as many strains at a 100k cutoff, where the approach has 2-3x as much power in most situations compared to *MaxDiversity* or *RandomCost,* other than in low MAF contexts (5-10%) where the benefits of selecting the cheapest strains adversely affects the ability to recover sufficient minor alleles to recover the phenotype. *MinCost*, when applied to real data, is remarkably consistent, recovering between 40-50% of loci regardless of cutoff, and outperforms *ThriftyMD* in retaining significant SNPs at low budgets. It does lose the second largest number of total SNPs due to changes in MAF thresholds after subsetting, however, which includes a number of the loci which were significant in the full panel and which likely explains the difference in recovered power between *MinCost* and *ThriftyMD* in the validation data despite their similar profiles in the simulated data.

MaxDiversity

Constrained by an algorithm that attempts to maximize overall genetic diversity at every step, and with much of that diversity contained within more expensive strains, the subsets selected by the *MaxDiversity* approach were consistently smaller than for the *MinCost* or *ThriftyMD* algorithms, and comparable to the *RandomCost* cohorts at lower cost thresholds before falling behind at higher cutoffs. These smaller cohorts resulted in reduced power compared to *MinCost* or *ThriftyMD* in most situations, although depending on the MAF, the relationship of *MaxDiversity* to the other approaches in terms of power recovered did change. At high MAF (40-50%), *MaxDiversity* was the weakest approach identified. In this case, the high abundance of the SNP in the population rendered the genetic advantage of the *MaxDiversity* algorithm over simple random selection largely meaningless, while the slight advantage of *RandomCost* over *MaxDiversity* in terms of strain number gave the former a slight edge in terms of power. At middle levels of MAF (10-30%), *MaxDiversity* shows a modest but consistent improvement in power compared to *RandomCost* as the diversity-maximizing algorithm provides a small benefit to the association testing. At low MAF (5-10%), however, *MaxDiversity* returns comparable power to that of *MinCost* or *ThriftyMD*, even surpassing these approaches at higher cost thresholds, and significantly better results than *RandomCost* throughout.

When applied to our validation dataset, the diversity-maximizing approach of *MaxDiversity* meant that we lost the fewest SNPs. However, *MaxDiversity* recovered among the fewest SNPs that were significant in the full panel, and was much less effective than both *ThriftyMD* and *MinCost* at lower cost thresholds. This suggests that in real experimental data, the effects of MAF and the number of selected strains may be more important than what was seen in our simulation studies.

ThriftyMD

While the *MinCost* algorithm concerns itself only with the cheapest set of strains for a given price threshold and the *MaxDiversity* algorithm seeks out the most diverse subset possible (after an initial seeding with an inexpensive strain), the *ThriftyMD* algorithm seeks to optimize diversity and cost at the same time using integer linear programming (ILP). In most cases, this balanced approach was among the best at maintaining power at any given cost threshold, with the notable exception of higher cost, lower MAF simulations, where it was beat out by the *MaxDiversity* algorithm. Intuitively, this makes sense, as ThriftyMD would be expected to prioritize common variants, potentially at the expense of rare outliers that might be more informative at lower minor allele frequency. When examining real data, the *ThriftyMD* algorithm lost the second fewest number of SNPs due to subsetting-induced MAF shifts, and was the

best overall at recovering the SNPs from the validation cohort that were actually significant (although *MinCost* outperformed it at the lowest budgets).

Cost Adjustments Affect Algorithm Efficacy

The cost of ordering a mouse strain can vary widely, from under $66 per breeding trio for the very popular C57BL/6J strain to $2,850 for a litter of one of many cryopreserved lines[24], a 47-fold difference. This is the main reason why the *MinCost* algorithm resulted in so many more strains than the *MaxDiversity* algorithm. However, in a typical application, there are many additional costs that are incurred after strain selection, including sequencing, reagents, labor, and mouse housing expenses. When we modeled this by adding an additional $2,500 to the total cost of each mouse strain, the difference between the cheapest and most expensive strains shrank to 2.1-fold, reducing the importance of minimizing cost at the possible expense of diversity. With these additional costs, the *MaxDiversity* algorithm actually had the strongest power across several MAF thresholds, possibly because this setting is closer to equal pricing, where greedy algorithms are optimal. These results highlight that the optimal choice of subset cohort may depend on the experiments that will be used to phenotype the panel.

Limitations and Other Considerations

In the current study, we modeled the effects of selecting subsets of strains from a larger cohort based purely on cost and/or genetic diversity. Other considerations, however, can also play a role in strain selection. For example, there is a wealth of previously generated data on a number of these strains available through websites such as the Mouse Phenome Database that may guide researchers to select certain strains with divergent phenotypes of interest as a 'seed' around which to grow a larger panel[32]. It would be relatively easy to adapt any of the selection algorithms discussed here to accommodate more than one starting strain, and knowledge of which strains are likely to show phenotypic extremes makes it more likely that the resulting panel will recover as much of the power of the full panel as possible given cost constraints.

Another limitation of this approach is the use of simulated phenotypes for the first part of the study, which were drawn from a multivariate normal distribution with correlated noise. This noise did include population structure (which we drew from previous estimates[25]), a major potential confounder in experimental GWAS data. However, the algorithms also performed differently in simulations vs. on subsampled experimental data. Most notably, the *MaxDiversity* algorithm under-performed on real data, where *ThriftyMD* performed best. Potential explanations include variables that the simulation did not model but that were present in the experimental data, or that a multivariate normal distribution does not model the real-world phenotypic data well. Finally, although the approaches we detail here are applicable to any form of research in which researchers must choose between different genetically diverse organisms who have differing costs to use, some of these observations may be specific to the inbred mouse lines from the HMDP genetic reference population that we used as a test case. Other collections may vary, for example, in their redundancy (many bacterial collections, for instance, have hundreds or thousands of isolates of *E. coli*, potentially interfering with methods that do not consider diversity) or in how evenly the panel samples natural diversity (which may be higher than average in a set of recombinant inbred lines).

Conclusions

In this study, we set out to provide some clarity to a difficult, real-world problem in genetics research – the cost of working on genetically diverse populations has increased faster than the funding available to perform such research[1,2]. With this in mind, we explored the use of four different budget-constrained versions of algorithms to select cohort subsets, with the aim of maintaining as much of the power of a larger, more complete set of samples as possible. Of these approaches, one considered neither cost nor genetic diversity, one each considered either cost or genetic diversity independently, and one considered both. Some of our results were somewhat unexpected, such as the strong performance of simply taking the cheapest strains possible regardless of genetic diversity, and the diversity-maximizing algorithm depending so strongly on the minor allele frequency of the significant SNP in question. Overall, the *ThriftyMD* algorithm, which uses integer linear programming to identify a sub-cohort that best represents a larger cohort given a particular budget, performed strongly across a majority of scenarios, and was the best performer in a validation dataset at only half the budget of the full experiment. Although some power loss is unavoidable when choosing a subset of a larger panel of strains, this study does show that there is room to preserve funding for downstream analyses or personnel costs by optimizing both diversity and costs.

Data and Code Availability

Mouse Diversity Array genotypes are available from the Mouse Genomes Informatics website

HMDP Phenotype data is available through Mendeley Resources at DOI:10.17632/y8tdm4s7nh.1

All code is available at https://github.com/ChristophRau/StrainSelect

Acknowledgements

We wish to thank Wesley Tansey for productive conversations at the start of this project.

CDR was supported by the National Institutes of Health: R01HL162636

References

1. Health NIo. R01-Equivalent Grants: Average Size NIH Data Book2024 [cited 2025 4/19/2025]. Available from: https://report.nih.gov/nihdatabook/report/158.
2. LeBel EP, Campbell L, Loving TJ. Benefits of open and high-powered research outweigh costs. J Pers Soc Psychol. 2017;113(2):230-43. doi: 10.1037/pspi0000049.
3. Gajovic S, Pochet R. The cost of scientific excellence - could it be expensive and out of reach? Croat Med J. 2016;57(5):413-4. doi: 10.3325/cmj.2016.57.413; PMCID: PMC5141463.
4. Wright JR, Levine MN. Researching the cost of research. J Clin Oncol. 2003;21(22):4081-2. doi: 10.1200/JCO.2003.08.972.
5. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, . . . Hirschhorn JN. A saturated map of common genetic variants associated with human height. Nature. 2022;610(7933):704-12. doi: 10.1038/s41586-022-05275-y; PMCID: PMC9605867.

6. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet. 2023;110(2):179-94. doi: 10.1016/j.ajhg.2022.12.011; PMCID: PMC9943775.
7. Chella Krishnan K, El Hachem EJ, Keller MP, Patel SG, Carroll L, Vegas AD, . . . Lusis AJ. Genetic architecture of heart mitochondrial proteome influencing cardiac hypertrophy. Elife. 2023;12. doi: 10.7554/eLife.82619; PMCID: PMC10241513.
8. de Conti A, Tryndyak V, Willett RA, Borowa-Mazgaj B, Watson A, Patton R, . . . Pogribny IP. Characterization of the variability in the extent of nonalcoholic fatty liver induced by a high-fat diet in the genetically diverse Collaborative Cross mouse model. FASEB J. 2020;34(6):7773-85. doi: 10.1096/fj.202000194R; PMCID: PMC7552439.
9. Salimova E, Nowak KJ, Estrada AC, Furtado MB, McNamara E, Nguyen Q, . . . Rosenthal NA. Variable outcomes of human heart attack recapitulated in genetically diverse mice. NPJ Regen Med. 2019;4:5. doi: 10.1038/s41536-019-0067-6; PMCID: PMC6399323.
10. Rau CD, Romay MC, Tuteryan M, Wang JJ, Santolini M, Ren S, . . . Lusis AJ. Systems Genetics Approach Identifies Gene Pathways and Adamts2 as Drivers of Isoproterenol-Induced Cardiac Hypertrophy and Cardiomyopathy in Mice. Cell Syst. 2017;4(1):121-8 e4. doi: 10.1016/j.cels.2016.10.016; PMCID: PMC5338604.
11. Rau CD, Parks B, Wang Y, Eskin E, Simecek P, Churchill GA, . . . Lusis AJ. High-Density Genotypes of Inbred Mouse Strains: Improved Power and Precision of Association Mapping. G3 (Bethesda). 2015;5(10):2021-6. doi: 10.1534/g3.115.020784; PMCID: PMC4592984.
12. Rau CD, Wang J, Avetisyan R, Romay MC, Martin L, Ren S, . . . Lusis AJ. Mapping genetic contributions to cardiac pathology induced by Beta-adrenergic stimulation in mice. Circ Cardiovasc Genet. 2015;8(1):40-9. doi: 10.1161/CIRCGENETICS.113.000732; PMCID: PMC4334708.
13. Hartmann K, Steel M. Maximizing Phylogenetic Diversity in Biodiversity Conservation: Greedy Solutions to the Noah's Ark Problem. Systematic Biology. 2006;55(4):644–51. doi: 10.1080/10635150600873876.
14. Faith DP. Conservation Evaluation and Phylogenetic Diversity. Biological Conservation. 1992;61(1):1–10. doi: 10.1016/0006-3207(92)91201-3.
15. Steel M. Phylogenetic Diversity and the Greedy Algorithm. Systematic Biology. 2005;54(4):527–9. doi: 10.1080/10635150590947023.
16. Pardi F, Goldman N. Species Choice for Comparative Genomics: Being Greedy Works. PLoS genetics. 2005;1(6):e71. doi: 10.1371/journal.pgen.0010071.
17. Zhang P, Zhan X, Rosenberg NA, Zöllner S. Genotype Imputation Reference Panel Selection Using Maximal Phylogenetic Diversity. Genetics. 2013;195(2):319–30. doi: 10.1534/genetics.113.154591.
18. Kang CJ, Marjoram P. A Sample Selection Strategy for Next-Generation Sequencing. Genetic Epidemiology. 2012;36(7):696–709. doi: 10.1002/gepi.21664.
19. McAuliffe JD, Jordan MI, Pachter L. Subtree power analysis and species selection for comparative genomics. Proc Natl Acad Sci U S A. 2005;102(22):7900-5. doi: 10.1073/pnas.0502790102; PMCID: PMC1142384.
20. Matsen FAt, Gallagher A, McCoy CO. Minimizing the average distance to a closest leaf in a phylogenetic tree. Syst Biol. 2013;62(6):824-36. doi: 10.1093/sysbio/syt044; PMCID: PMC3797636.
21. Faith DP. Phylogenetic Pattern and the Quantification of Organismal Biodiversity. Philosophical Transactions: Biological Sciences. 1994;345(1311):45–58.
22. Kang JT, Zhang P, Zollner S, Rosenberg NA. Choosing Subsamples for Sequencing Studies by Minimizing the Average Distance to the Closest Leaf. Genetics. 2015;201(2):499-511. doi: 10.1534/genetics.115.176909; PMCID: PMC4596665.
23. Phenotypic Data for the HMDP Heart Failure Study [Internet]2016.
24. Labs J. Mouse Cost List 2024 [cited 2024 4/05/2024]. Available from: https://www.jax.org/pricelist.
25. Rau CD, Gonzales NM, Bloom JS, Park D, Ayroles J, Palmer AA, . . . Zaitlen N. Modeling epistasis in mice and yeast using the proportion of two or more distinct genetic backgrounds: Evidence for "polygenic

epistasis". PLoS Genet. 2020;16(10):e1009165. doi: 10.1371/journal.pgen.1009165; PMCID: PMC7644088.
26. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8(10):833-5. doi: 10.1038/nmeth.1681.
27. Ghazalpour A, Rau CD, Farber CR, Bennett BJ, Orozco LD, van Nas A, . . . LeBoeuf RC. Hybrid mouse diversity panel: a panel of inbred mouse strains suitable for analysis of complex genetic traits. Mamm Genome. 2012;23(9-10):680-92. doi: 10.1007/s00335-012-9411-5; PMCID: PMC3586763.
28. Lusis AJ, Seldin MM, Allayee H, Bennett BJ, Civelek M, Davis RC, . . . Wang J. The Hybrid Mouse Diversity Panel: a resource for systems genetics analyses of metabolic and cardiovascular traits. J Lipid Res. 2016;57(6):925-42. doi: 10.1194/jlr.R066944; PMCID: PMC4878195.
29. Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, . . . Lusis AJ. A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res. 2010;20(2):281-90. doi: 10.1101/gr.099234.109; PMCID: PMC2813484.
30. Schumacher D, Ooms J, Yapparov B, Hanson J. rcbc: COIN CBC MILP Solver Bindings. 0.1.0.9002 ed2024.
31. Forrest J, Lougee-Heimer R. CBC User Guide. Tutorials in Operations Research. 2014:257-77. doi: 10.1287/educ.1053.0020.
32. Bogue MA, Ball RL, Walton DO, Dunn MH, Kolishovski G, Berger A, . . . Chesler EJ. Mouse phenome database: curated data repository with interactive multi-population and multi-trait analyses. Mamm Genome. 2023;34(4):509-19. doi: 10.1007/s00335-023-10014-3; PMCID: PMC10627943.

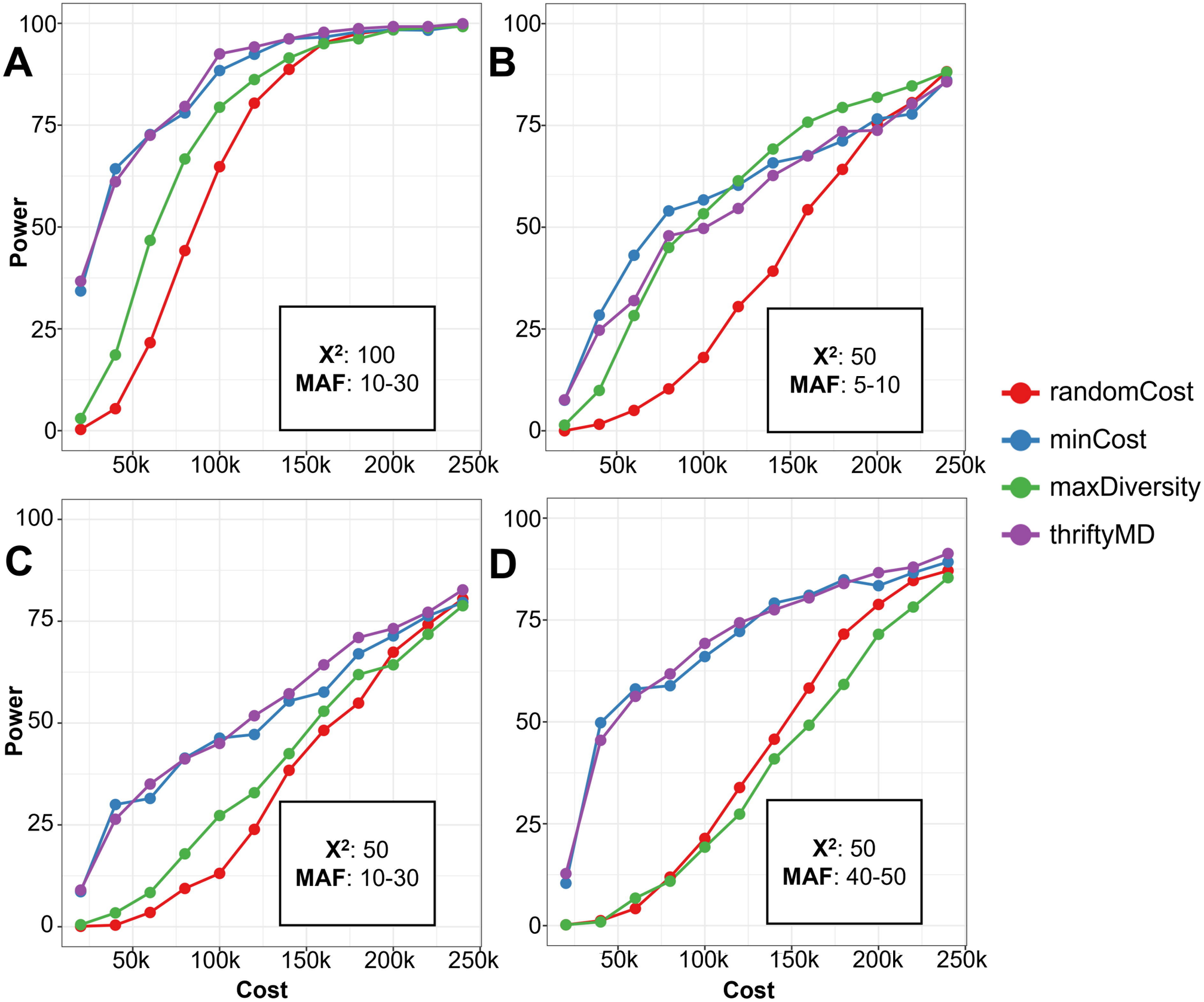
Figure 1: Power Simulations Without Additional Costs Per Strain. Results of simulation studies at various cost thresholds. **A)** χ^2 of 100 and MAF of 10-30%, **B)** χ^2 of 50 and MAF of 5-10%, **C)** χ^2 of 50 and MAF of 10-30%, **D)** χ^2 of 50 and MAF of 40-50%. Points represent the average of 1,000 tests per algorithm/cost combination

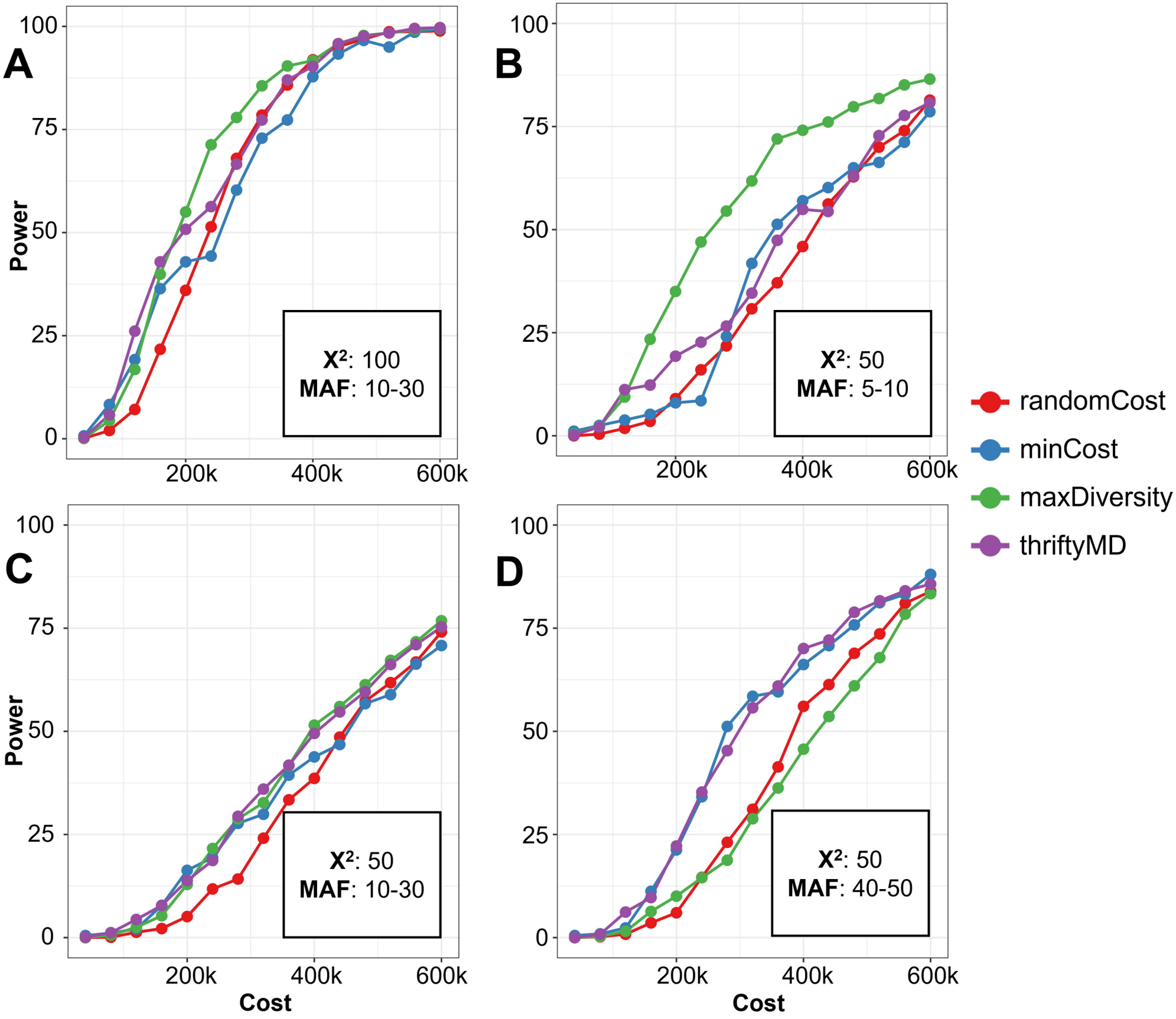
Figure 2: Power Simulations With Additional Cost Per Strain. Results of simulation studies at various cost thresholds after $2,500 was added to the cost of each strain. **A)** χ^2 of 100 and MAF of 10-30%, **B)** χ^2 of 100 and MAF of 5-10%, **C)** χ^2 of 50 and MAF of 10-30%, **D)** χ^2 of 50 and MAF of 40-50%. Points represent the average of 1,000 tests per algorithm/cost combination.

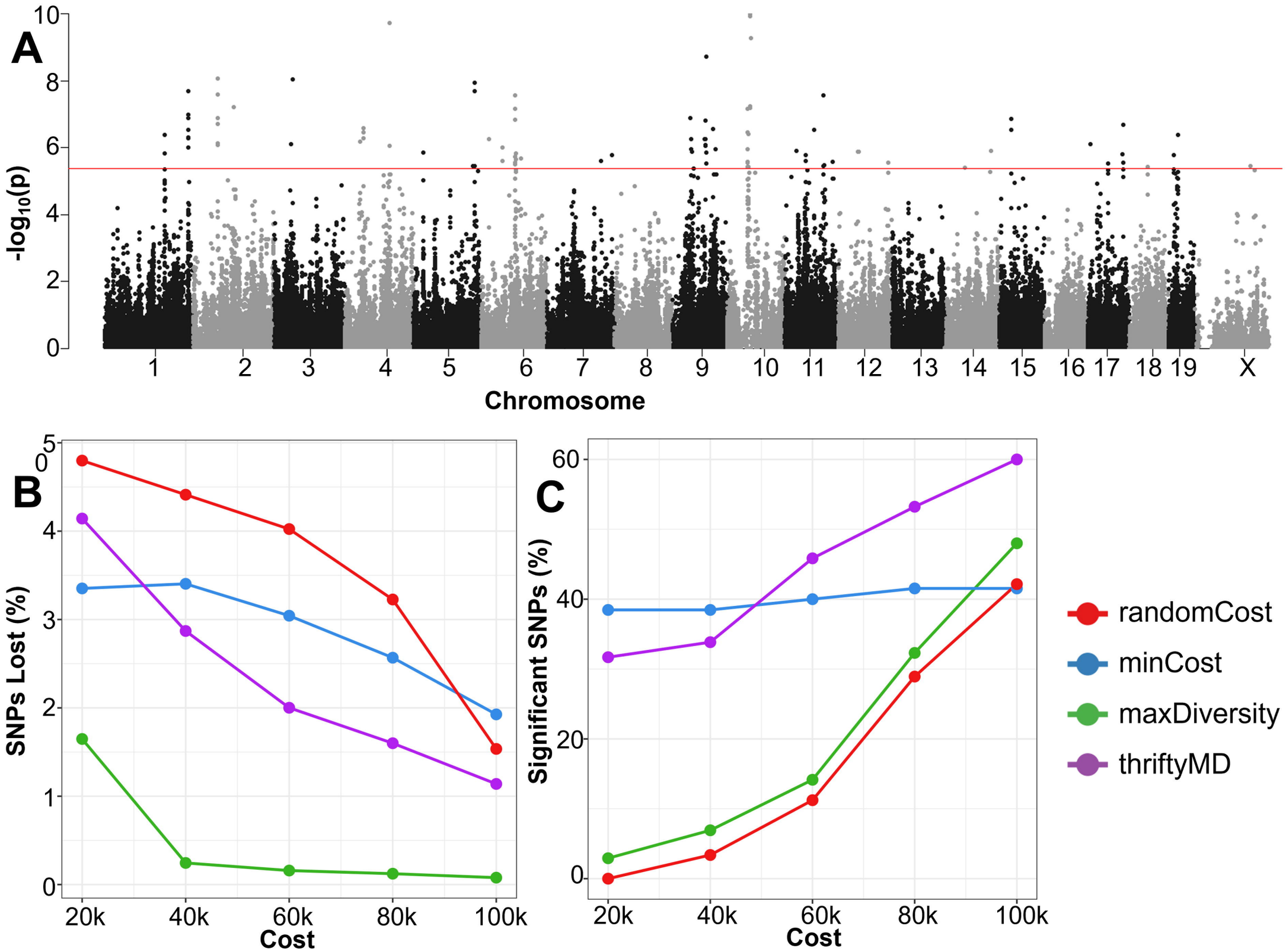
Figure 3: Results of Validation Tests. Results of subsetting studies (N=10 permutations each) at various cost thresholds for real data pulled from the HMDP Heart Failure Study[12]. **A)** Manhattan plot of the results of the whole panel, red line represents the genome-wide significance threshold of a FWER of 5%[29] **B)** Percentage of SNPs untested in the subsets because of changing MAFs. **C)** Percentage of preserved genome-wide significant SNPs after subsetting.

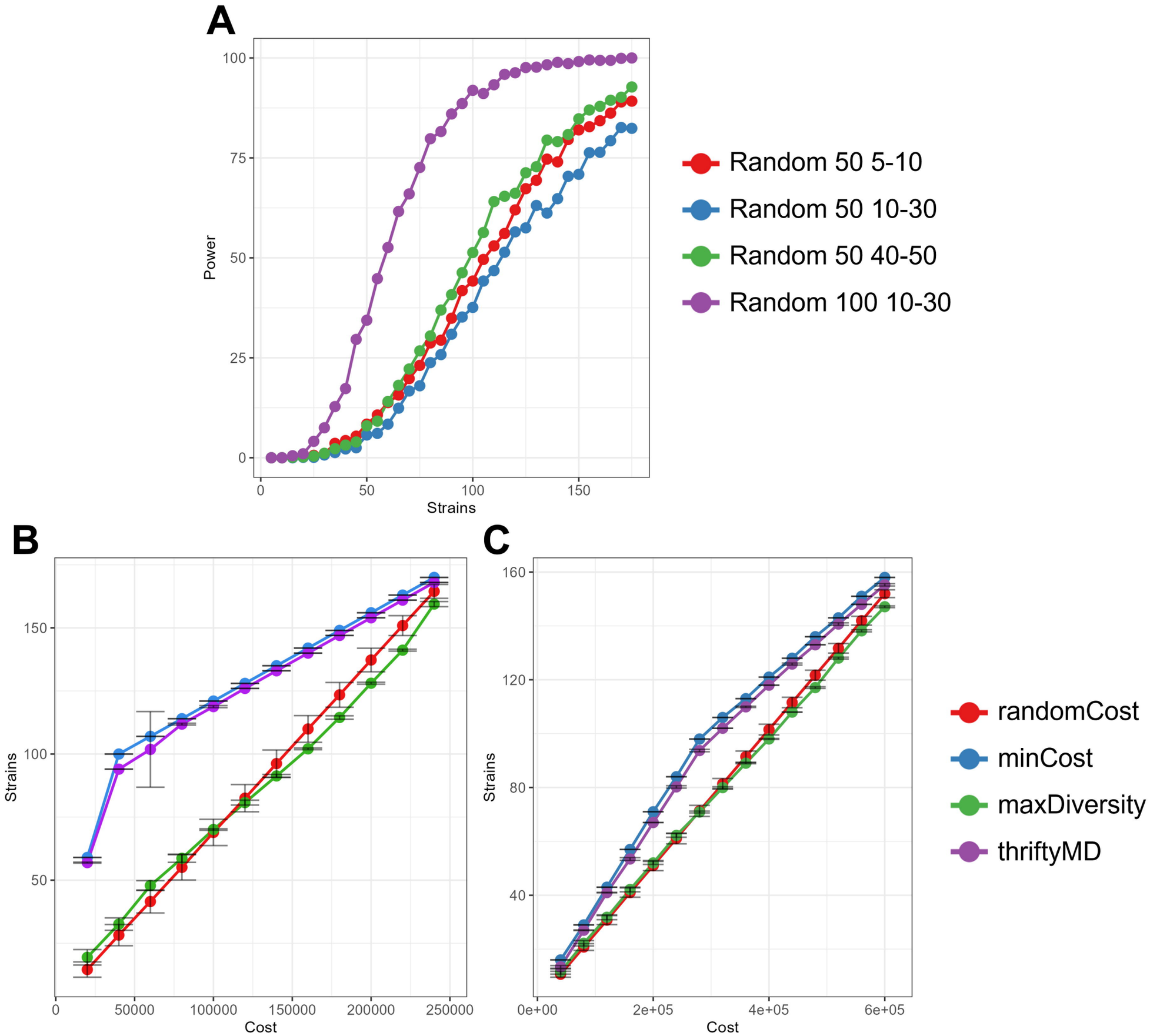
Figure S1. Results of Random Strain Power Testing and Number of Strains Selected. A) Results of 1,000 permutations at 5-strain intervals in the HMDP cohort for various χ^2 and MAF thresholds **B)** Total number of strains selected when no additional costs are added per strain for each approach. **C)** Total number of strains selected when $2,500 is added to the cost of each strain to account for research costs

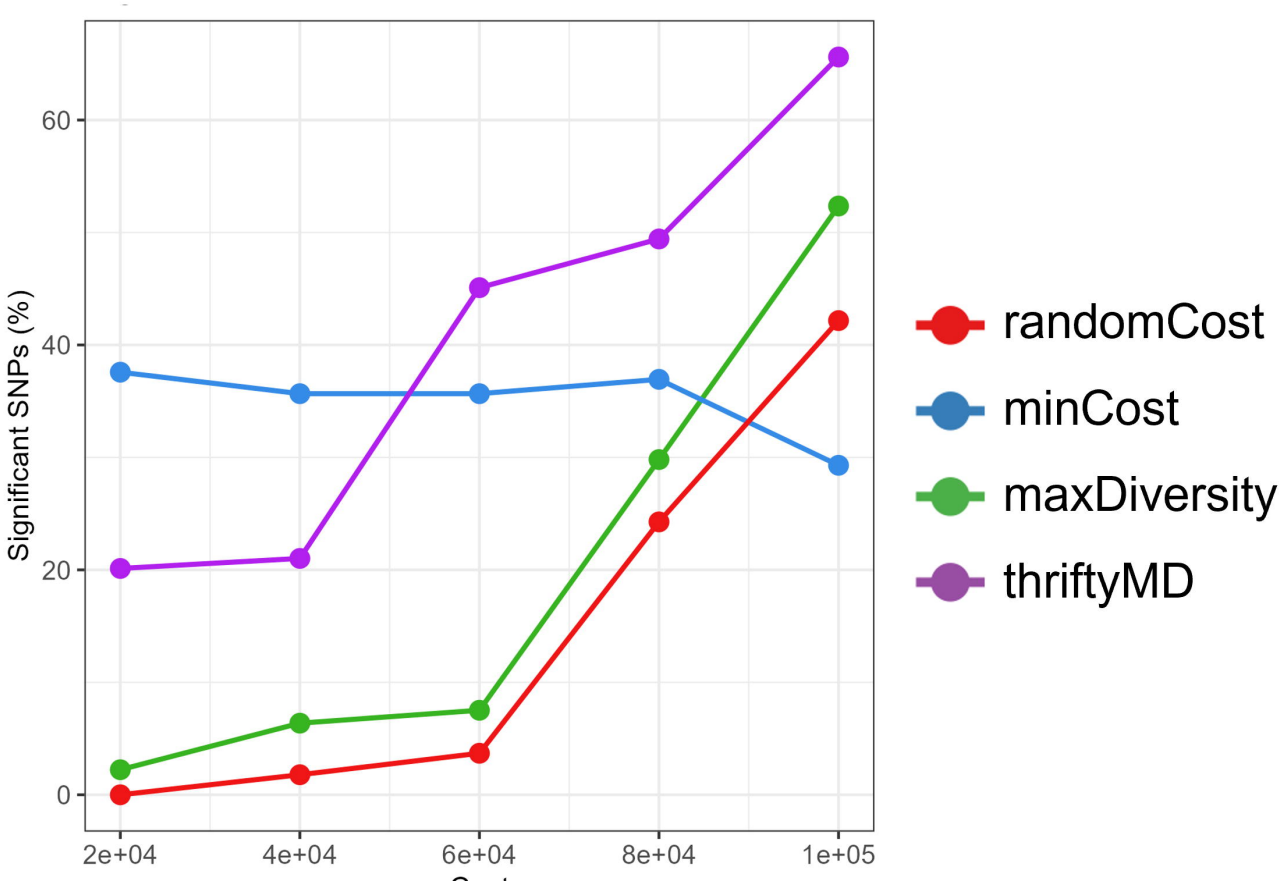
Figure S2. Preserved Genome-Wide Significant Loci in the RV HMDP HF Study After Subsetting and Binning. SNPs were binned into 500kb windows across the genome with the most significant P value per bin representing the entire bin. The percent of significant bins in the original study which are preserved in each subset are displayed


A
B
C
D
Power
Cost
0
25
50
75
100
50k
100k
150k
200k
250k
X²: 100
MAF: 10-30
X²: 50
MAF: 5-10
X²: 50
MAF: 10-30
X²: 50
MAF: 40-50
randomCost
minCost
maxDiversity
thriftyMD


A
B
C
D
Power
Cost
0
25
50
75
100
200k
400k
600k
X²: 100
MAF: 10-30
X²: 50
MAF: 5-10
X²: 50
MAF: 10-30
X²: 50
MAF: 40-50
randomCost
minCost
maxDiversity
thriftyMD


A
-log10(p)
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
B
SNPs Lost (%)
Cost
20k
40k
60k
80k
100k
C
Significant SNPs (%)
Cost
randomCost
minCost
maxDiversity
thriftyMD


A
Power
Strains
Random 50 5-10
Random 50 10-30
Random 50 40-50
Random 100 10-30
B
Strains
Cost
C
Strains
Cost
randomCost
minCost
maxDiversity
thriftyMD


Significant SNPs (%)
60
40
20
0
2e+04
4e+04
6e+04
8e+04
1e+05
Cost
randomCost
minCost
maxDiversity
thriftyMD