Genome Wide Association Study using Espresso Methods

Janis Waser

August 20, 2025

1 Goal

Genome wide association studies (GWAS) are increasingly reliable and thus able to explain biological phenomena. A polygenic risk score is easily computable, but only gives limited insight to the inner mechanisms which are involved. Analyzing all possible combinations by brute-force for large data sets is not feasible and we must find methods which circumvent this complexity and still produce reliable results.

We are using minimal cover algorithms (namely Espresso) to find a small selections of single nucleotide polymorphisms (SNPs) in a causal relationship for a given binary phenotype. With these selections, we would like to explain the phenotypes. Further, we investigate the genome-wide spanning relations between SNPs and potential groupings of the phenotype such as different subtypes of a disease.

Contents

2   Introduction

Genome wide association studies (GWAS) are done to identify significant risk loci from genome data. The goal is to understand the relation between the observed phenotype and the underlying genetic structure.

For this purpose many tools such as PRS already exist. "A polygenic risk score (abbreviated PRS) uses genomic information alone to assess a person's chances of having or developing a particular medical condition. A person's PRS is a statistical calculation based on the presence or absence of multiple genomic variants, without taking environmental or other factors into account."[4] Through stratification different ethnicities can be captured accurately.

Rarely is it possible to assess the relationship between multiple single nucleotide polymorphisms with PRS as a tool due to it being limited to a single genomic variational presence or absence. Two or more independent causes for a disease might greatly affect the capability of this model. On the other hand, PRS is resistant against undetected cases in the control group.

We want to explore a different method using Espresso which from a truth table finds a minimized circuit. We hope that thanks to this circuit, we are able to extract more information especially of the relations between different SNPs. A minimized cover can be understood as simplest possible explanation, therefore we actively search for it.

2.1   Quality control and Data

Quality control (QC) is a crucial step when processing data. In the context of genetic data, specific limits are set to ensure the quality of the data. Data which violates this form are excluded either through the exclusion of an entire SNP (column) or a subject (row). It involves analysis of the Hardy-Weinberg equilibrium, the heterozygosity rate, checking cryptic relatedness (familiar matches), limiting the missingness of data, matching gender data with the sex chromosome. SNPs with a low minor allele frequencies (MAF) are also excluded as no significant statement is possible.

Our data stems from a GWAS/PRS tutorial[1]. We also stick to the instructions for the QC given during this tutorial, after which our data has 109 subjects and 1'073'226 SNPs.

For the selection of approaches the number of permitted unknowns is crucial, which is set to 2% per individual and SNP. In the scope this means that there might still exist above 20'000 unknowns per individual.

We found a second data set on imputed simulated blood lipid data from the 1k Genome Project [3]. It also contains more than one million SNPs for 2'504 individuals before QC. There are no missing genotype data. After QC, with 1% MAF threshold 75'733 SNPs remain, no individuals had to be filtered out.

2.2   Espresso

Espresso is a tool that performs 2-level logic minimization. It uses heuristics to find a satisfying minimal cover. It takes a logic table in form of a truth table as input and outputs a minimized circuit. The output is not guaranteed to be optimal. We consider it to be a

minimal cover, though this might not be true.

"-" (don't cares) tell Espresso that it can assign any value to this place. It is possible to specify special input forms for example *.type fr* to indicate that in case of an underspecified truth table the underspecified cases are treated as if they were don't cares.

One wrong data point especially if the phenotype (outcome of the truth table) is affected, can have a huge impact on the found minimal cover. Therefore, we want to emphasize the importance of QC for this approach as it generally relies on the assumption that the data has no inconsistencies at all and that the phenotype is fully determinable by genetic analysis and no outside influence.

2.3 Translation of genetic data into binary

Most genetic data in particular for human autosomal genes is stored in two pairwise inherited strings. The manifestation can be made up from one of the four nucleobases (A/C/G/T) or an indentation of any length or a deletion (non-existence). The information for any particular SNP might also be partially missing due to errors in the laboratory. For any given position there commonly exist two different manifestations. One is labelled as the no-risk allele and the other allele is referred to as risk allele. There is no general consensus on what allele should be considered the risk allele, generally the variant with lower sampling rates is considered the risk allele. Two different studies might find different might find different risk alleles but they would still agree on the same manifestation which is in correlation with the disease.

We abstract the manifestations to a count of the occurrences of the risk allele. This count we decode into two digits long binary number. To avoid the unnecessary big Hamming distance of 2 between the counts of 2 (10_2) and 1 (01_2), while the smaller distance between 2 and 0 (00_2) would only have a Hamming distance of 1, we encode 2 as 11_2. This seems to slightly improve the performance, which is visible in Figure 5.

Phenotypes for each subject should already exist in an easily binary translatable form. We focus on binary phenotypes, it could be imagined to extend the approach to continuous data, but we have to remember that the entire deviation should be explained by genetic data and no outside factors. Different potentially related diseases with pleiotropy might be a more fitting use-case.

3 Evaluation of results

We use a range of different criteria to approximate the quality of our methods. Each criteria has its own advantages and flaws which should always be taken into consideration. Although, no single criteria is sufficient for proofing a working method, it can give an indication whether this path should be pursued or not.

1. Previous identification of SNPs by other researchers

2. Out of data prediction accuracy of the phenotype

3. PRS

4. Products

5. Literals

6. Time and complexity of the solution

We might also discuss potential issues or perks of a specific method which go beyond this list.

3.1 Previous identification of SNPs by other researchers

If there already exists a genetic analysis on the specific phenotype we are researching, we can compare those findings. In other cases, it might even be the case that datasets or phenotypes are simulated. Knowing the underlying structure, we expect our results to match the simulations SNPs, under the assumption that our method works.

3.2 Out-of-data accuracy

We can apply a k-fold process to our data where we select 90% of the individuals and the rest serves as out-of-sample data. We can then test the individuals not included in our sample to test whether they satisfy the minimal cover or not. From this we determine whether this deviates from their phenotype. Now, a deviation means a wrong prediction. We expect $2p^2 - 2p + 1$ to be the share of correctly guessed phenotypes by random guessing where p is the probability of having the phenotype of the "disease". In practice, an even higher limit should be set at p if $p > 0.5$ else $1 - p$ as simply always guessing the same outcome attains this accuracy.

3.3 PRS

The PRS can be easily determined. It is a single value for each SNP indicating the modeled risk of having the disease when having the risk allele. The distribution of the PRSs is likely to be normal with barycentre around 0. The scores from multiple SNPs can be added to get a (polygenic) risk score.

It is possible to select a couple of SNPs including their direction (risk/non-risk allele) and calculate a score. If the score is extreme it indicates predictive power (in positive or negative way), while scores near 0 indicate low predictive power.

For the purpose of analyzing the combined predictive power of SNPs the PRS is not a particularly effective metric. It only states the individual predictive power, which may be obfuscated by balancing out effects. Generally, a high PRS indicates that some SNPs with high predictive power have been found, but it is not required that they are balanced (unrelated) and therefore low scores should a priori not indicate a failure of the method.
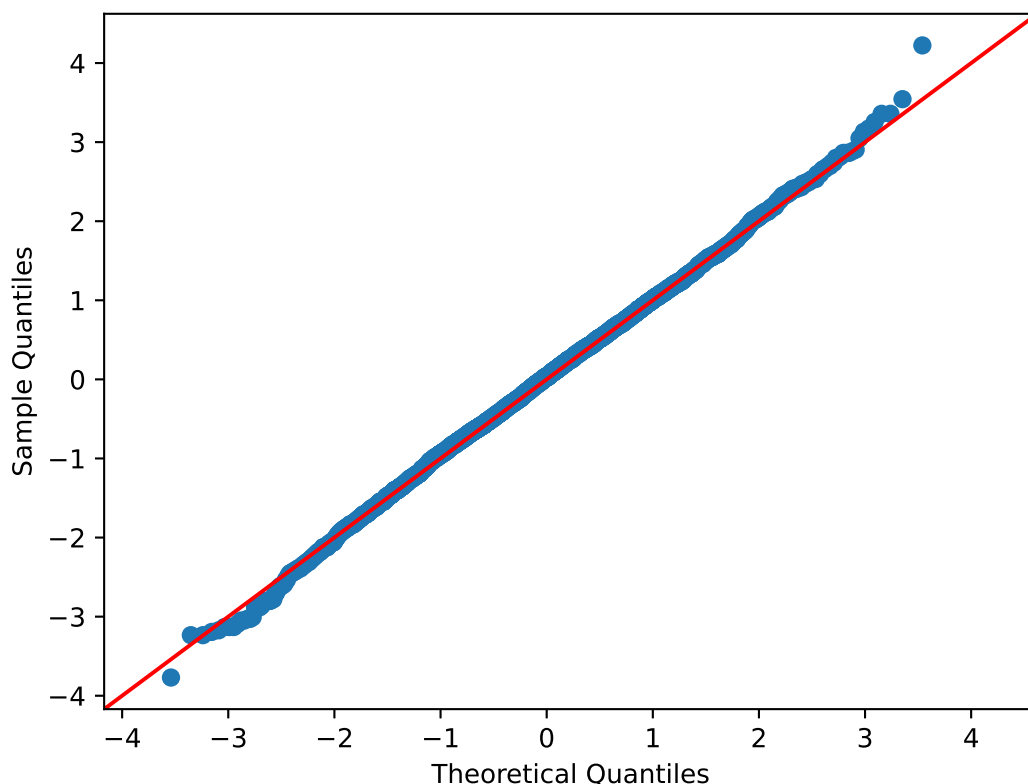
Figure 1: QQ-plot of combined PRS, clearly distributed normally

3.4 Products

The output from Espresso is given as one line per product. The amount of products can be used as an indication of how many subgroups there are needed to find a minimal cover.

Generally, we assume less products in the result to be preferable as this can be seen as simpler relationship, which we deem to be more probable to fit the real cause.

It should be analyzed what impact each of the products have if there are multiple products, eventually it might prove necessary to discard some products which are specifying only to one small subset, while other products are true for many subjects. This is to avoid overfitting by stratifying to fit the given data perfectly but not the out-of-sample data.

If multiple products with considerable sizes (for in- and out-of- sample data) deviate in many SNPs, it can be speculated that two different groups being affected by different subtypes of diseases were classed into one single phenotype. This is especially interesting when multiple causes or outcomes are suspected.

3.5 Literals

The amount of literals like for products indicates how complex the found solution is, again we prefer the simplest solution as we deem if to fit the real causes with a higher chance.
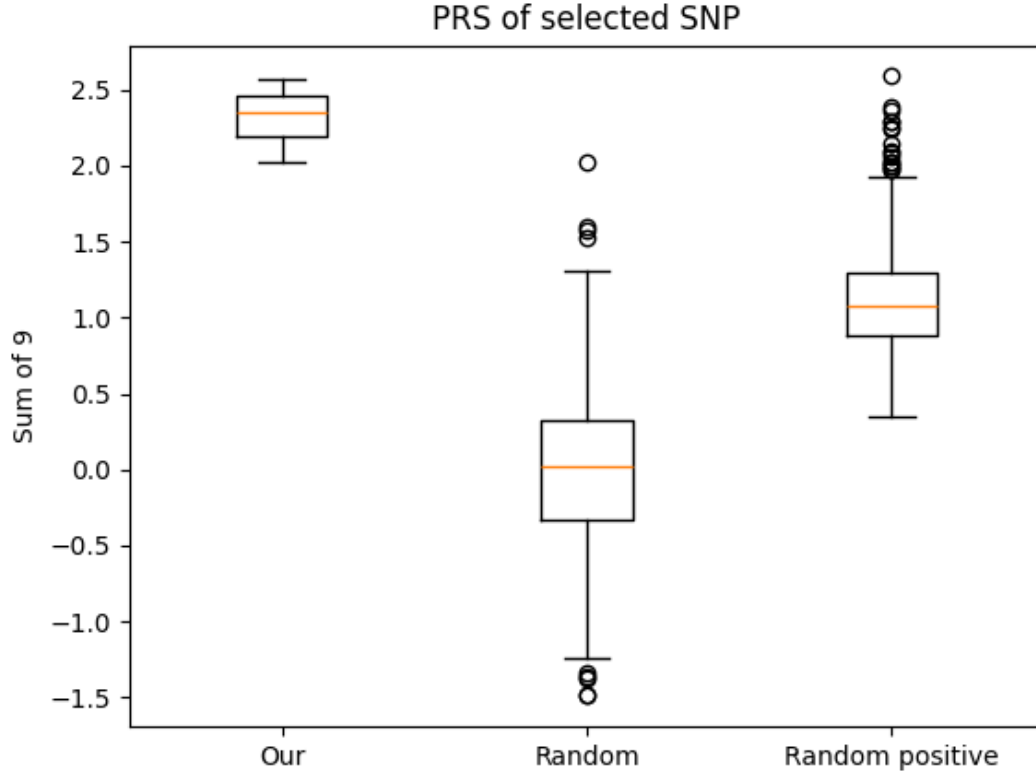
Figure 2: Boxplot of PRS of different selections with same size, our selections stem from the pyramid scheme, they deviate significantly from a random selection.

3.6 Time & Complexity

Since the complexity of our approach is high, it is inevitable to consider it as part of the evaluation. An easy measure of the complexity is the time it took us to run the program. Change of dataset, parameters or computational capabilities all have an influence on the time.

4 Tried methods

Since no literature exists, nor do we have an already working method, we have to build it ourselves. Since most of the algorithms used do not change between the methods, we try to build up an infrastructure relying on the already built methods and modifying them to support additional methods or use the prebuilt block for building new blocks.

4.1 Primitive approaches

Some methods seem to be trivial, but they will never work with our dataset. Nevertheless is it a good idea to consider those approaches as they might help us understand the more complex methods.
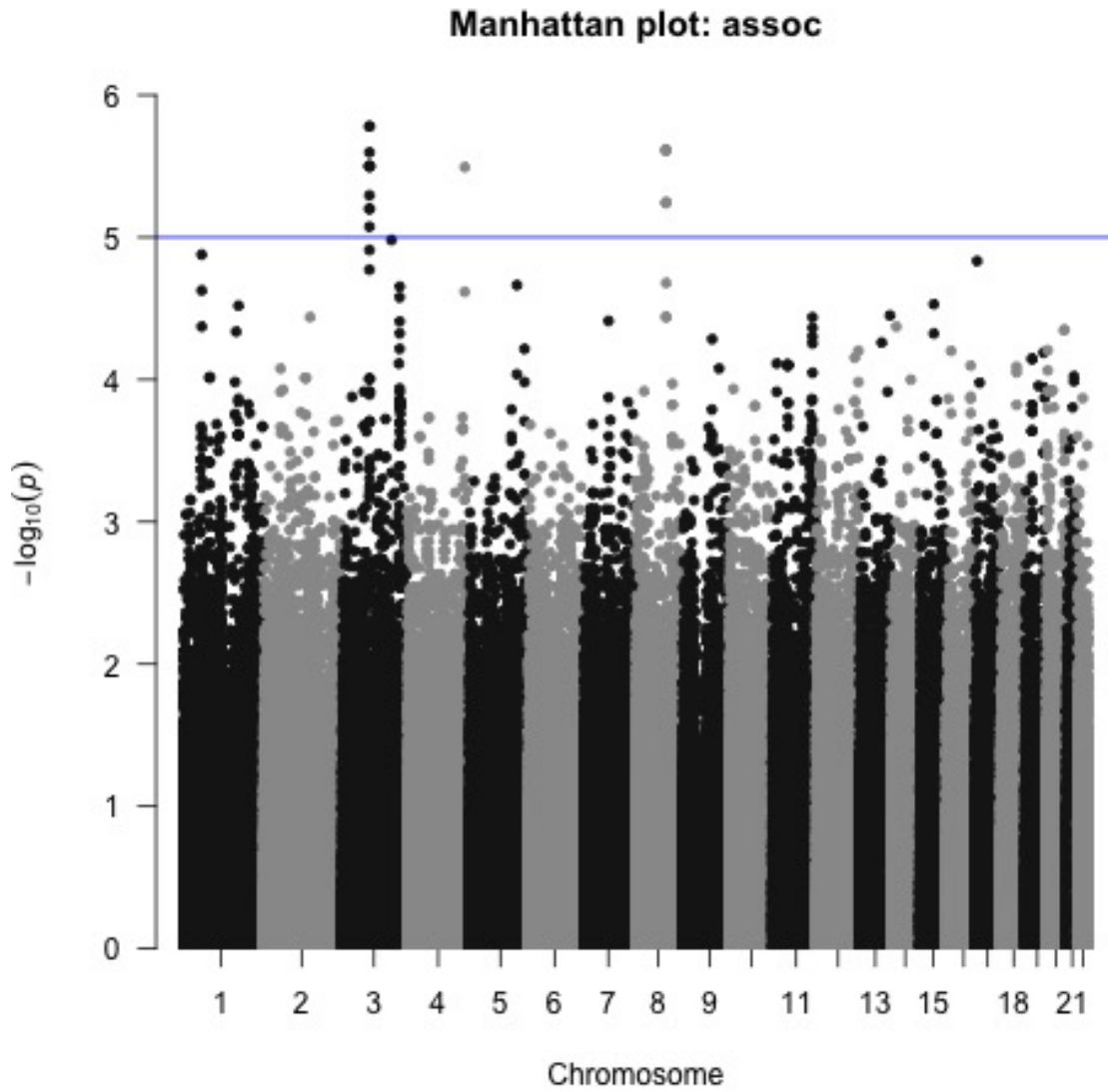
Figure 3: Manhattan plot of our toy data

4.1.1 Entire

Taking the entire dataset and let Espresso try to find a minimal cover is impossible as the time requirements grow exponentially, as we know the problem is not even in P.

For small datasets, it might be possible and useful, though wrong genotyping or incorrect phenotypes will remain hidden. Another approach could be to filter more strictly in an attempt to decrease the size of the dataset.

If previous knowledge exists suggesting that the result is made up by a single product, we could apply a filter and excluded all positions where the subjects having the phenotype differ as this makes it unsatisfiable if it were in the cover.

4.1.2 Omitting parts of the dataset

Taking partial parts of the dataset and never discovering other parts is possible though it is highly doubtful how a satisfying solution can even be found consistently. For discovering the dataset and experimenting with different encodings, we did run some experiments. In the context of this experiment when we refer to n we mean the amount of SNPs.

4. High n yield a smaller product, we explain deviations from this rule by the diverging paths taken in Espresso or in case of different dataset by the difference of possibilities.

5. Same as for products, just more pronounced as in practice this are higher numbers and therefore more possible outcomes (events) exist.

6. This analysis is fast and easy to execute for small n, it scales exponentially with n.

The following table gives an overview of the acquired results with all 109 subjects considered. Each SNP is decoded as two digits, hence the input size is double the amount of selected SNPs. The results are averaged, estimated and depend on the selection of the SNPs. Here the selection was done randomly. We might use those results as baseline assumptions in further experiments:

Input size (2n)	Products	Literals	Time
<120	-	-	< 1s
200	50	3-6	5s
400	40	2-4	10s
1000	30	2-3	40s
2000	25	2	8min

For a random selection with less than 60 SNPs, we can generally not find a cover as the truth table is over-specified.

A random selection of SNPs generally decreases the identified products required for a minimal cover compared to a sequential selection. This trend also affects the literals but is less pronounced. Additionally, it was discovered that in sequential analysis even for a relative high n it might not be possible to find a cover due to over-specification. This is due to missing data not being distributed equally among the data, causing regions with high missingness ratio. This constitutes less of a concern for random selection, as in QC we filtered to have at an average of at most 2% missingness.

The random method will be used exclusively in all future analysis as we deem it to yield the better results for our purposes.

Different encoding methods (see 2.3) also influence the result. The adjusted method is selected for the rest of this document.

4.1.3 Iterative approaches

We can try to start with some small selection and iteratively build up until the entire data set is included. We start with the smallest possible set such that the truth table is not over-specified and add one SNP, we evaluate this table, find a minimal cover and from the number
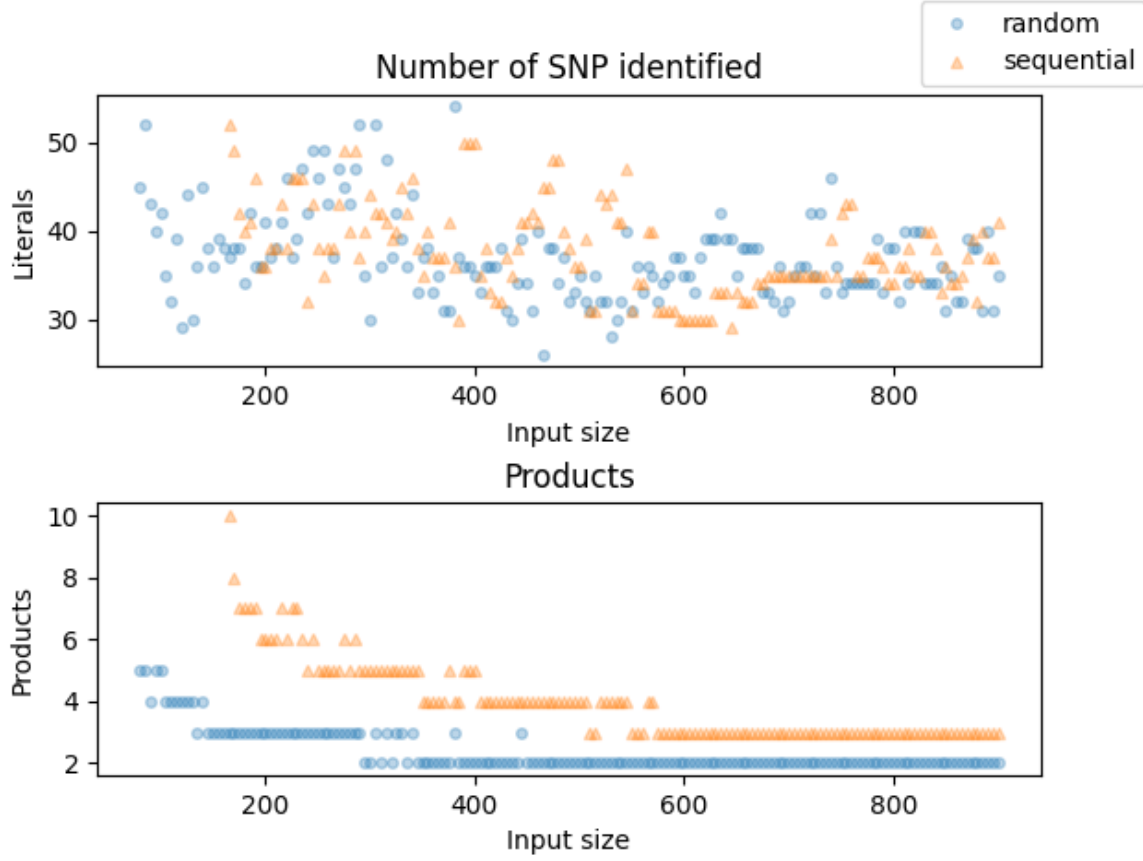
Figure 4: Comparison between random and sequential selection methods

of literals in the minimal cover, we decide whether this SNP stays in the selection depending on some criteria. Then we do same for every SNP until all SNP were included at one point.

The time required to go through the entire dataset is immense, we are not able to tell Espresso what covers we already found and therefore we do not gain anything from the iterations before. From some point onwards there are likely no gains as the best solution is a local minima of only a few SNPs.This approach does not consider all possible relationships between different SNPs. In fact, it only remarks the ones with a strong PRS anyways or a tendency towards the starting set, making this approach not equilibrated. Furthermore, we are not guaranteed by Espresso to have an optimal minimal cover.

4.2 Pyramid scheme

In the pyramid scheme, the algorithm runs sequentially through the entire dataset by separating all SNPs into groups of size k, finding a minimal cover for every group. All the SNPs making part of a found minimal cover are then subsequently taken to the next level. This is repeated until in one level all the identified SNPs from the previous level are less than k SNPs, from where a minimal cover for the entire universe of less than k SNPs can be found.
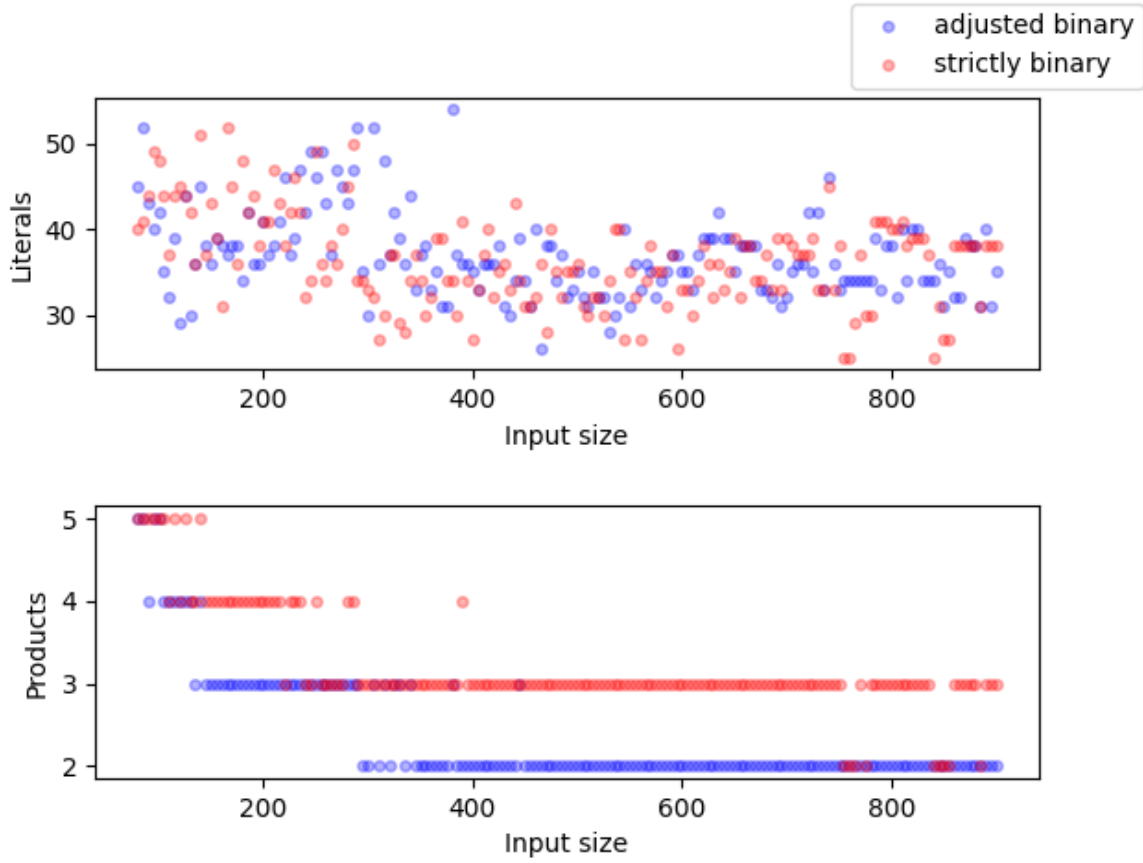
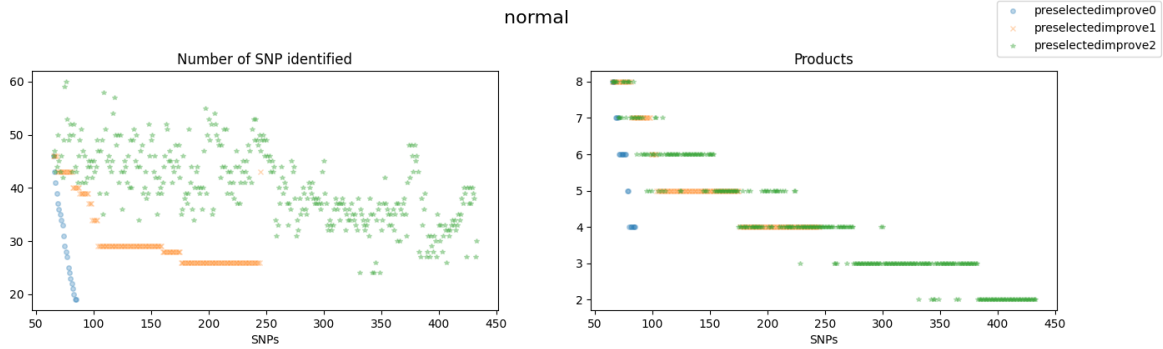Figure 5: Comparison between different binary encoding methods

Figure 6: Iterative approaches with SNP indicating the number of selected SNPs, the limit was set on literals. 0: strictly decreasing, 1: monotonically decreasing, 2: no equivalences (increasing or decreasing)

The selection of the group can be done in many ways, we decided to analyze the random selection approach as well as sequential, which should correspond to the sequences in the DNA as already explained in Chapter 4.1.2. We found random selection to be clearly preferable as

we can see in Figure 4 and also because of the missingness problem with sequential selection.
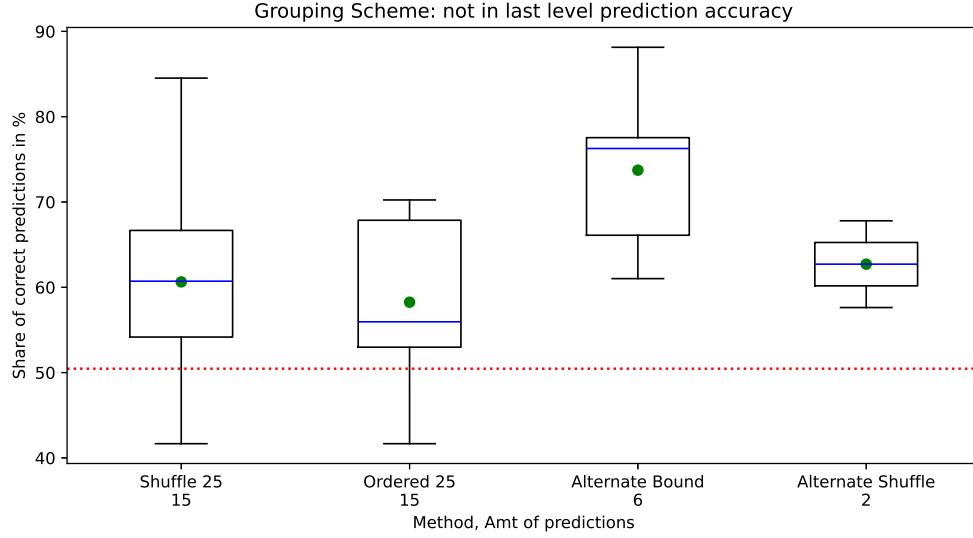
4.3   Grouping scheme

Figure 7: Prediction accuracies of the grouping scheme for the out of sample which means not in last level. Every category originally had 15 sample, because of over-specification some are eliminated, the updated count is displayed below the category name.

The grouping scheme method divides the individuals into different equal-sized groups. For each level one group is chosen and the dataset of only those individuals is considered, the rest is performed similar to the Pyramid scheme.

On each level, only a part of the individuals is considered, this means, that the minimal cover is optimized only for those people. On the next level, different individuals get selected, we can assume their phenotype is determined by the same underlying structure, therefore the previous minimal covers should suffice to find a minimal cover. If this is not the case, we have an over-specified truth table, this might be due to there not being an underlying structure or the missing detection of it or due to errors in genotyping. We tested two methods for passing SNPs down a level: shuffling in every level and preserving original state. Preserving states did generally perform worse in terms of products and literals but also for the out-of-sample accuracy. On the other hand, it might occur that good combinations from the previous level are split when shuffling in every level.

We found group sizes of 25% of the in-sample data to be fitting for the approach with our initial dataset. With alternating groups, such as when splitting the individuals into two groups (group sizes of 50%), the method yields lower prediction accuracies than the Pyramid scheme or 25% group sizes. We suspect it was optimizing for the in-sample data. Additionally, alternating often fails due to over-specification especially if the order from the previous level is not respected (shuffle SNPs after each level).

Later the code was changed to refer to the previous level's selection of individuals if the

final analysis is over-specified. This should resolve most cases where over-specification was still posing a problem. We maintain the stance the randomization between the levels is preferable as this allows for more diverse range of combinations.

5   Creating data with given structure

In this chapter we try to evaluate what is possible to achieve with our method and what has to be expected simply by random occurrences. We simulate or change data such that we have a more direct correspondence, which we can try to predict. This gives us an understanding of the capabilities of our methods.

5.1   Phenotype shuffling

We shuffle the phenotypes of all the individuals, this is achieved through random selection of k persons which have phenotype 2 respectively 1. This preserves p of having the phenotype. This altered dataset can be run to test different methods. When this is done we expect the resulting minimal cover to be random. Our method where structure exists, should need less literals or products as there is an underlying structure dictating the outcome, while this is obviously not true for shuffled phenotypes. We can also check the frequencies of the selected SNPs with shuffled phenotypes to see whether there exists an overlap with the SNPs of our normal method. If this were to be the case repeatedly, this would hint at a problem in our method potentially stemming from the fact that identifying individuals with certain SNPs is easy, which we generally do not want as this would be overfitting.

It was found that not all schemes support phenotype shuffling. For example the grouping scheme relies on the fact that there is some underlying structure determining the phenotype from the genotypes. It was observed that over-specification might occur that since due to the phenotypes being random attributed. This causes the SNP from a previous level to be nearly entirely insignificant.

5.2   Introducing artificial diffusion based on specific criteria

We would like to test how multiple underlying products affect the data. Unfortunately, the available test data only requires one product. If we change the phenotype of all persons based on some specific genotype variation, we hope to see this exact configuration in the results. Furthermore, we expect two products of which one is composed of our chosen diffusion criteria and to be closely related to a valid minimal cover for the original data. Further alterations or even combining of phenotypes might proof to be useful not only for evaluating the results but also for analyzing phenotypes with considerable pleiotropy or suspected subtypes.

We could not replicate the expectations in any of the 5 run experiments. We run the experiment with the grouping scheme with group sizes of 25. We diffused the phenotype by oring the phenotype with the pairwise selection of a specific SNP. This replaced phenotype, we expect to have the same minimal cover plus an additional product containing the specific SNP.

5.3 Simulated data with given phenotype-genotype relations

We create a simulated dataset. For simplicity, we attribute only A (risk) and G (norisk) for all SNP not used for the phenotype generation, while C (risk) and T (norisk) serve for SNP determining the phenotype. This does not have an influence on the results as we encode them independent of the actual letter. Each SNP (with some caveats for the determining SNPs), gets attributed a random MAF uniformly distributed between 5% and 50%.

We mostly tested with the same size as our initial dataset. If less anded 3 SNPs determine the SNP, our methods find this exact cover. Adding more anded SNPs, gives as result only a subset of all determining SNPs with 3 identified SNPs. Apparently they suffice for a minimal cover, repeated tests could unveil more SNPs, identification with the histogram should be possible. Oring the determining SNPs for the phenotype failed to be detected even with only 1 SNP per product. This is due to a limitation in Espresso, which needs to be investigated further.
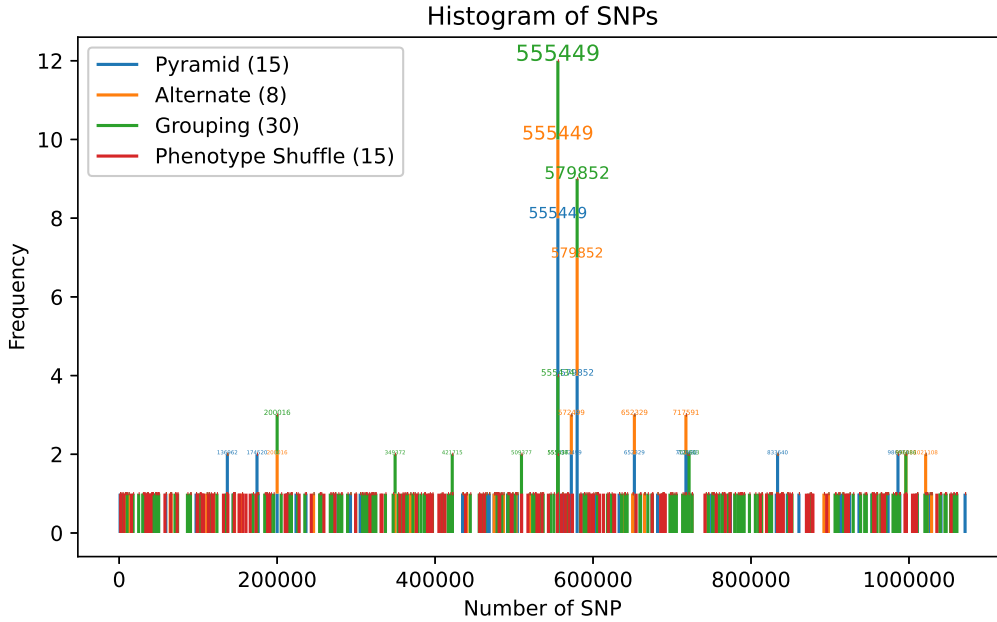
6 Evaluation of the methods

Figure 8: Histogram of multiple k-folds using one out of pyramid scheme, grouping scheme or phenotype shuffling schemes. The grouping scheme is superimposed on top of the k-fold. No overlap is detected, clear trends with the pyramid scheme are visible, while for shuffled phenotypes no order is apparent.

From the out-of-sample prediction accuracy, we determine that some approaches perform better than random, though we would hope for higher accuracy.
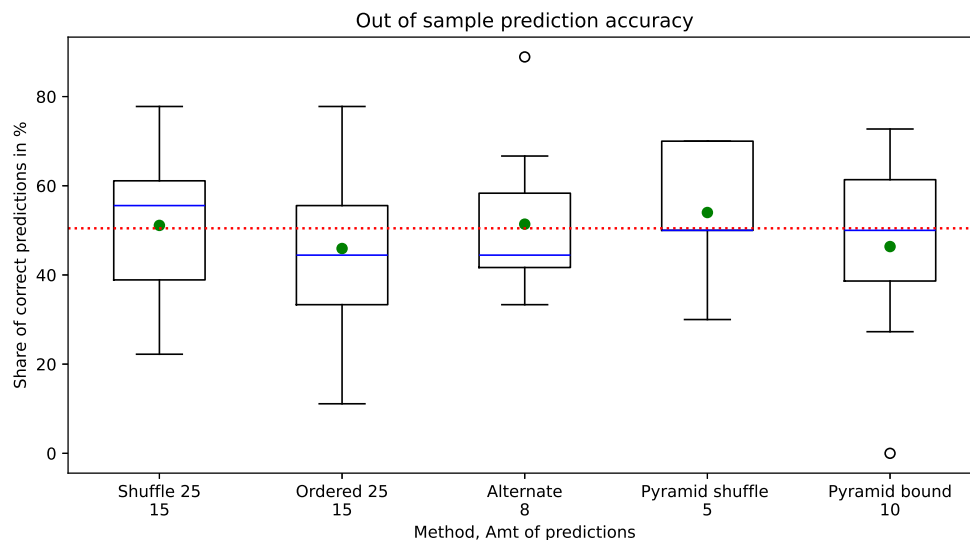
Figure 9: Prediction accuracies of predicted outcomes in data out of sample data, never included in the data before. The red lines indicates the expected guessing accuracy for this data with a random approach. The data is theoretically continuous, here we observe the out-of-sample size to be too small (about 10) , explaining the occurrences of overlapping quartiles and medians.

We consistently see some of the SNPs reappearing in different covers, which might signal the importance of that specific SNP. For shuffled phenotypes, this phenomena was not observed, hence there is no selection which is generally good for identifying persons. This result is expected and supports the claim of significance of the found SNPs.

The histogram seems less crowded than the Manhattan plot of the PRS of Figure 3. Also most of the SNP have a high PRS value indicating correlation with the phenotype. Repeated results are a sign of confidence, but it also needs to be mentioned, that we are far away from having equal results in all iterations. This also indicates that potentially multiple global optima exist or it hasn't even been found yet. Except for the pyramid scheme there does not exist clear "favorite" SNPs but rather a small collection of SNPs appearing more than once.

Another question is whether the global optima or any found cover for this purposes actually has biological credibility in explaining the surveyed phenotype.

References

[1] *Marees AT et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. `https://pmc.ncbi.nlm.nih.gov/articles/PMC6001694/`

[2] *Hajiaghabozorgi, M.* GWAS dataset with simulated binary phenotypes for 1000 Genome Project. March 1, 2023. `https://doi.org/10.5281/zenodo.7683384`

[3] *The 1000 Genomes Project Consortium.* A global reference for human genetic variation. Nature 526, p. 68–74. 2015.`https://doi.org/10.1038/nature15393`

[4] *Teri Manolio* National Human Genome Research Institute (NIH). August 19 2025. `https://www.genome.gov/genetics-glossary/Polygenic-Risk-Score-PRS`

[5] *Peter M. Visscher et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation, The American Journal of Human Genetics . July 6 2017, p 5-22.