Genome Wide Association Study using Espresso Methods

Janis Waser

July 31, 2025

1 Goal

Genome wide association studies (GWAS) is 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 a minimal cover algorithms 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 Quality control

We want to emphasize the importance of quality control (QC) for this approach as it relies heavily on the assumption that the data has no inconsistencies.

Our data stems from a GWAS/PRS tutorial and the corresponding Git directory which we also use for quality control [1]. It has 109 subjects and 1'073'226 SNPs after the aforementioned QC.

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.

3 Espresso

Espresso is a tool that performs 2-level logic minimisation. It uses heuristics to find a satisfying minimal cover. It takes a logic table as input and outputs a minimized circuit. The output is not guaranteed to be optimal.

"-" can used as don't cares. It is possible to specify special input types such as *.type fr* to indicate that not in case of an underspecified truth tables the underspecified cases are treated as don't cares.

4 Translation of genetic data into binary

Most genetic data is stored in two pairwise inherited strings this is true particular for human autosomal genes. 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. 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 consensus for every SNP 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. For the effect of this choice consider Figure 2.

Phenotypes for each individual should already exist in an easily binary translatable form. We focus on binary phenotypes, if this approach shows promising results it is also possible to extend the approach to continuous data or different potentially related diseases with pleiotropy.

5 Evaluation of results

We use a range of different criteria to approximate the quality of our method. Each criteria has its own advantages and flaws which should always be taken into consideration. No single criteria is sufficient for showing a working method rather does it give an indication.

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.

6 Tried methods

6.1 Primitive approaches

6.1.1 Entire

Taking the entire dataset and let it be solved by Espresso, is impossible as the time requirements grow exponentially, as we know the problem is not even in P.

6.1.2 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 good solution should found consistently. For discovering the dataset and experimenting this we did run some experiments. In the context of this experiment when we refer to n we mean the amount of SNPs.

2. The out-of-data prediction accuracy is as predicted around the 50% mark, potentially slightly above for hign n.
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).
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 and are used as a baseline assumption:

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

If there are less than 60 SNP, it is generally not possible to find an assignment as the truth table would be over-specified.

A random selection of SNP generally decreases the identified products required for a minimal cover, this trends also affects the literals but is less pronounced. Additionally, it was discovered that missing data is not distributed equally among the data and therefore in sequential analysis even for a relative high n it might not be possible to find a cover due to over-specification. For random selections this constitutes less of a concern as we are guaranteed to have at most 2% missingness.

Figure 1: Comparison between random and sequential selection methods

The random method will be used exclusively in all future analysis.

Different encoding methods also influence the result. The adjusted method is selected going forwards (see 4)

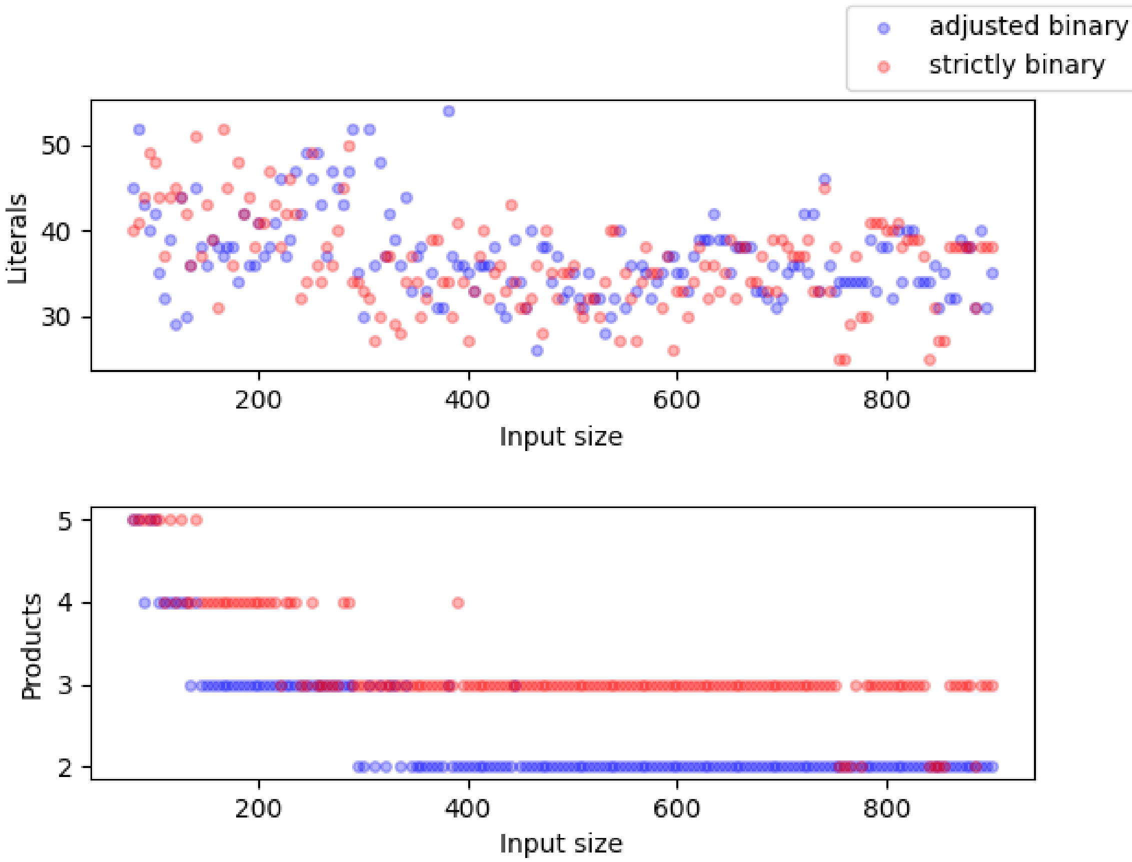

Figure 2: Comparison between different binary encoding methods

The adjusted method will be used exclusively in all future analysis.

6.1.3 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 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.

The time it takes to go through the entire dataset is immense and from some point onwards there are likely no gains. This approach does not consider all possible relationships between different SNPs, in fact in only remarks the ones with a strong PRS anyways or the starting set, making this approach not equilibrated.

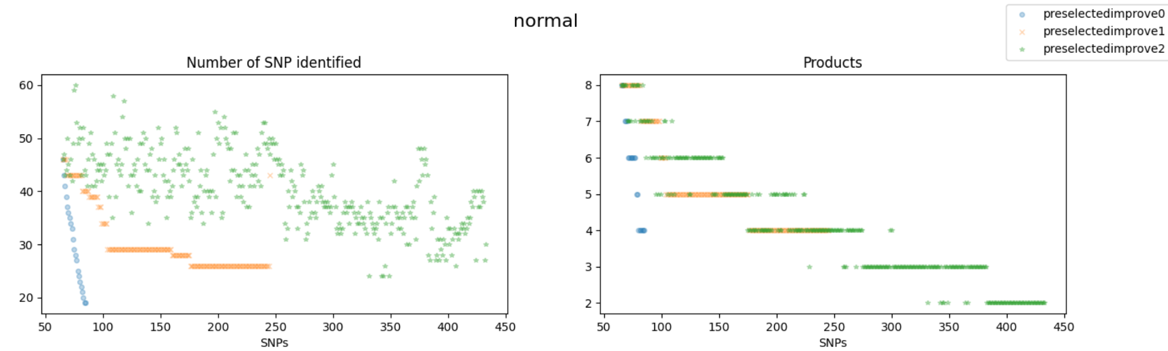

Figure 3: Iterative approaches with SNP indicating the number of selected SNPs, the limit was set on literals. 0: strictly decreasing, 1: monotonically decreasing, 2: no equivalences (increasing or decreasing)

6.2 Pyramid scheme

6.3 Phenotype shuffling

We shuffle the phenotypes of all the individuals, this is achieved through random selection of k persons which have phenotype 2 resp. 1. This altered dataset can be run through the same pyramid scheme as before or also other later scripts. When this is done we expect the resulting minimal cover to be random, meaning that the method we compare to where structure exist should need less literals or products as there is an underlying structure dictating the outcome, where this is obviously not true for shuffled phenotypes. We can also control with the frequencies of the selected SNPs with shuffled phenotypes whether there exists an overlap with the SNPs of our normal method.

6.4 Subgroup scheme

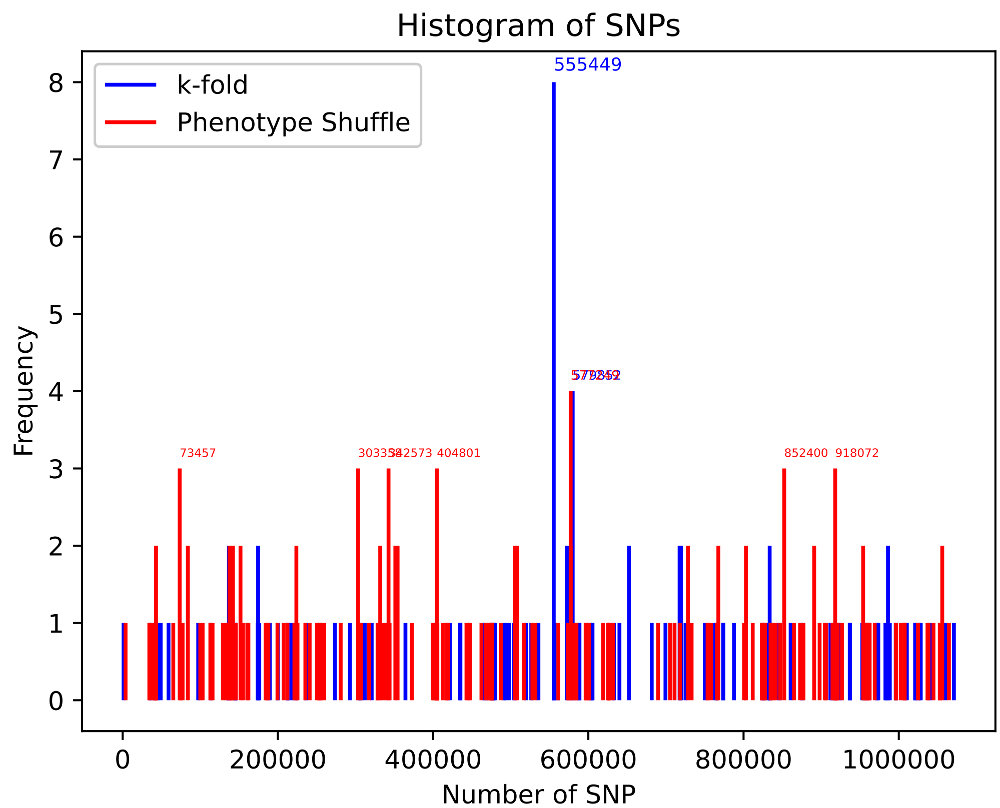

Figure 4: Histogram of mutiple k-folds using the Pyramid scheme and mutiple phenotype shuffling schemes. No overlap is detected, clear trends with the pyramid scheme are visible.)

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/`