Benchmarking the Accuracy of Polygenic Risk Scores and their Generative Methods: Supplementary I

Scott Kulm[1,2,3], **Jason Mezey**[4,5,*], **and Olivier Elemento**[2,3,*]

[1]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY
[2]Caryl and Israel Englander Institute of Precision Medicine, Weill Cornell Medicine, New York, NY
[3]Physiology, Biophysics and Systems Biology Graduate Program, Weill Cornell Medicine, New York, NY
[4]Department of Genetic Medicine, Weill Cornell Medicine, New York, NY
[5]Department of Computational Biology, Cornell University, Ithaca, NY
[*]corresponding authors

Trait	Sex Included	Cancer Self-Report	Non-cancer Self-Report
Lupus	A	NA	1381
Atrial Fibrillation	A	NA	1471
Asthma	A	NA	1111
Celiac Disease	A	NA	1456
OCD	A	NA	1615
Vitiligo	A	NA	1661
Gout	A	NA	1466
Alzheimer's	A	NA	1263
Ulcerative Colitis	A	NA	1463
Crohn's Disease	A	NA	1462
Type 2 Diabetes	A	NA	1223
Stroke	A	NA	1081
Breast Cancer	F	1002	NA
Coronary Artery Disease	A	NA	1075|1076
Rheumatoid Arthritis	A	NA	1464
Type 1 Diabetes	A	NA	1222
Eczema	A	NA	1452
Chronic Kidney Disease	A	NA	NA
Ovarian Cancer	F	1039	NA
Multiple Sclerosis	A	NA	1261
Prostate Cancer	M	1044	NA
Schizophrenia	A	NA	1289
Colorectal Cancer	A	1020	NA
Psoriasis	A	NA	1453
ALS	A	NA	NA

Table 1. The self-reported codes used to define the respective trait. Sex included refers to which sex was included in the analysis, A is all, M is male, and F is female. The "|" symbol refers to a logical or, meaning that any code will indicate a case. The description of each code is provided within the UK Biobank website (https://www.ukbiobank.ac.uk/).

Trait	ICD-9 Code	ICD-10 Code
Lupus	710	M321|M328|M329
Atrial Fibrillation	4273	I48
Asthma	493	J45|J46
Celiac Disease	579	NA
OCD	3003	F42
Vitiligo	7091	L80
Gout	274	M10
Alzheimer's	3310	F00|G30
Ulcerative Colitis	556	K51
Crohn's Disease	555	K50
Type 2 Diabetes	NA	E11
Stroke	431|432|433|434| 435|436|437|438	I60|I61|I62|I633| I64|I65|I66|I67| I68|I69
Breast Cancer	174	C50
Coronary Artery Disease	410|411|412	I21|I22|I23|I241| I252
Rheumatoid Arthritis	714	M05|M06
Type 1 Diabetes	NA	E10
Eczema	69180|6929	L20
Chronic Kidney Disease	585	N18|N19
Ovarian Cancer	183	C56
Multiple Sclerosis	340	G35
Prostate Cancer	185	C61
Schizophrenia	295	F20|F21|F23|F24|F25|F28|F29
Colorectal Cancer	1530|1531|1532|1533| 1534|1536|1537|1538| 1539	C180|C182|C183|C184| C185|C186|C187|C188| C189|C19|C20
Psoriasis	6960|6961	L40
ALS	3352	G122

Table 2. The ICD codes used to define the respective trait. The "|" symbol refers to a logical or, meaning that any code will indicate a case. The description of each code is provided within the UK Biobank website (https://www.ukbiobank.ac.uk/).

Trait	OPCS Code	Medication Code
Lupus	NA	NA
Atrial Fibrilation	K622|K623	1140888482
Asthma	NA	1141168340
Celiac Disease	NA	NA
OCD	NA	1140879544
Vitiligo	NA	NA
Gout	NA	1140875408
Alzheimer's	NA	1141150834| 1141182732| 1141171578
Ulcerative Colitis	NA	1141153242
Crohn's Disease	NA	NA
Type 2 Diabetes	NA	NA
Stroke	U543|Z35	NA
Breast Cancer	B27|B28|B29	1140923018|1141190734
Coronary Artery Disease	K40|K41|K45|K49| K50|K75	NA
Rheumatoid Arthritis	U504	1141145896|1140909702| 1141188588|1141180070| 1140871188
Type 1 Diabetes	NA	NA
Eczema	NA	1141165244|1140854688
Chronic Kidney Disease	M01|M02	NA
Ovarian Cancer	NA	NA
Multiple Sclerosis	NA	1140911642| 1140923792
Prostate Cancer	NA	1141150594|1140870274| 1140921100
Schizophrenia	NA	1140870194
Colorectal Cancer	H04|H05|H06|H07| H08|H09|H10|H11	NA
Psoriasis	NA	NA
ALS	X852	1141195974

Table 3. The OPCS and medication codes used to define the respective trait. The "|" symbol refers to a logical or, meaning that any code will indicate a case. The description of each code is provided within the UK Biobank website (https://www.ukbiobank.ac.uk/).

	ICD	Self-report	Medication	Dual-agreement
ICD-9	X	X	X	X
ICD-10	X	X	X	X
Noncancer Self-Report		X	X	X
Cancer Self-Report		X	X	X
OPCS		X	X	X
Medication			X	X

Table 4. The codes within the UK Biobank which defined each phenotyping method. If any of the coding mediums (ICD-9, ICD-10, Noncancer Self-Report, etc.) reported a disease event, then the individual was treated as a case. Except for the double phenotyping method, in which case two sources had to report a disease event.

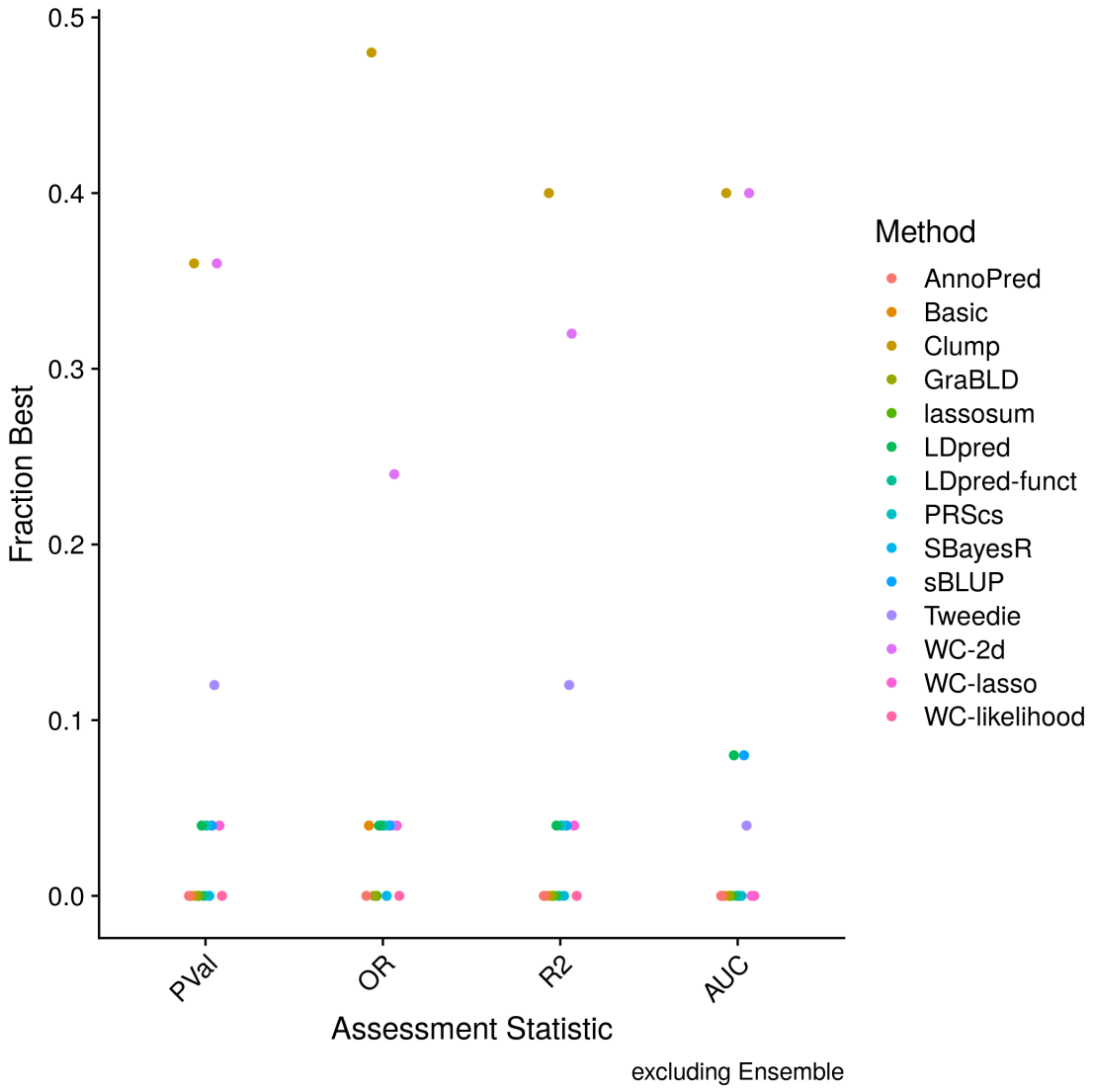

Figure 1. The fraction a method was chosen as best in the comparison phase, not including ensemble. Each trait is ranked according to the assessment statistic and the best method is chosen. The fraction, number of times a method is chosen divided by the total 25 traits, is indicated.

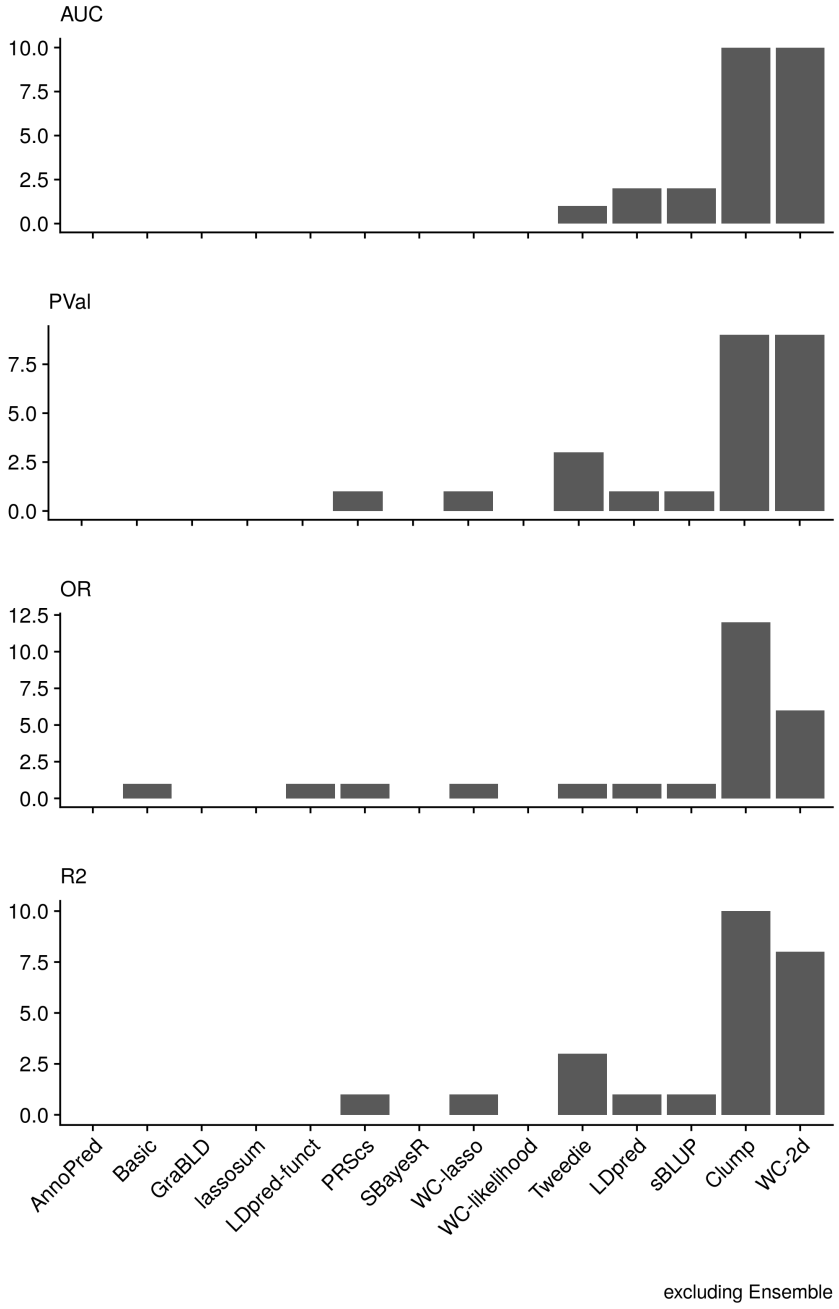

Figure 2. Histogram of the number of times a method was selected as best in the comparison phase, not including ensemble. Each trait is ranked according to the assessment statistic and the best method is chosen. The exact number of times it was chosen is indicated.

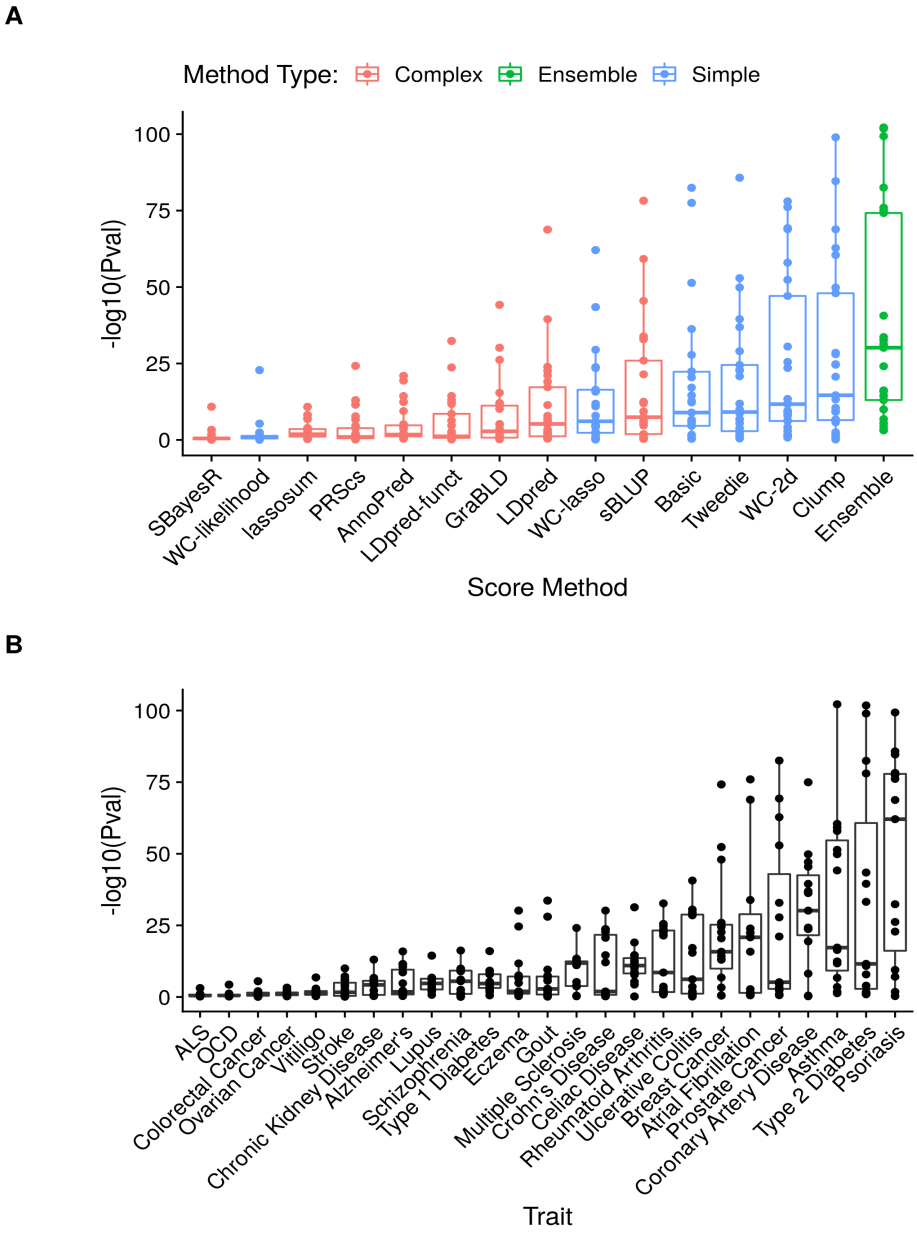

Figure 3. The -log10(pval) for traits and methods. Panel A depicts the -log10(pval) of the methods, where each point represents the value for a specific trait. Panel B depicts the -log10(pval) of the traits, where each point represents the value for a specific method.

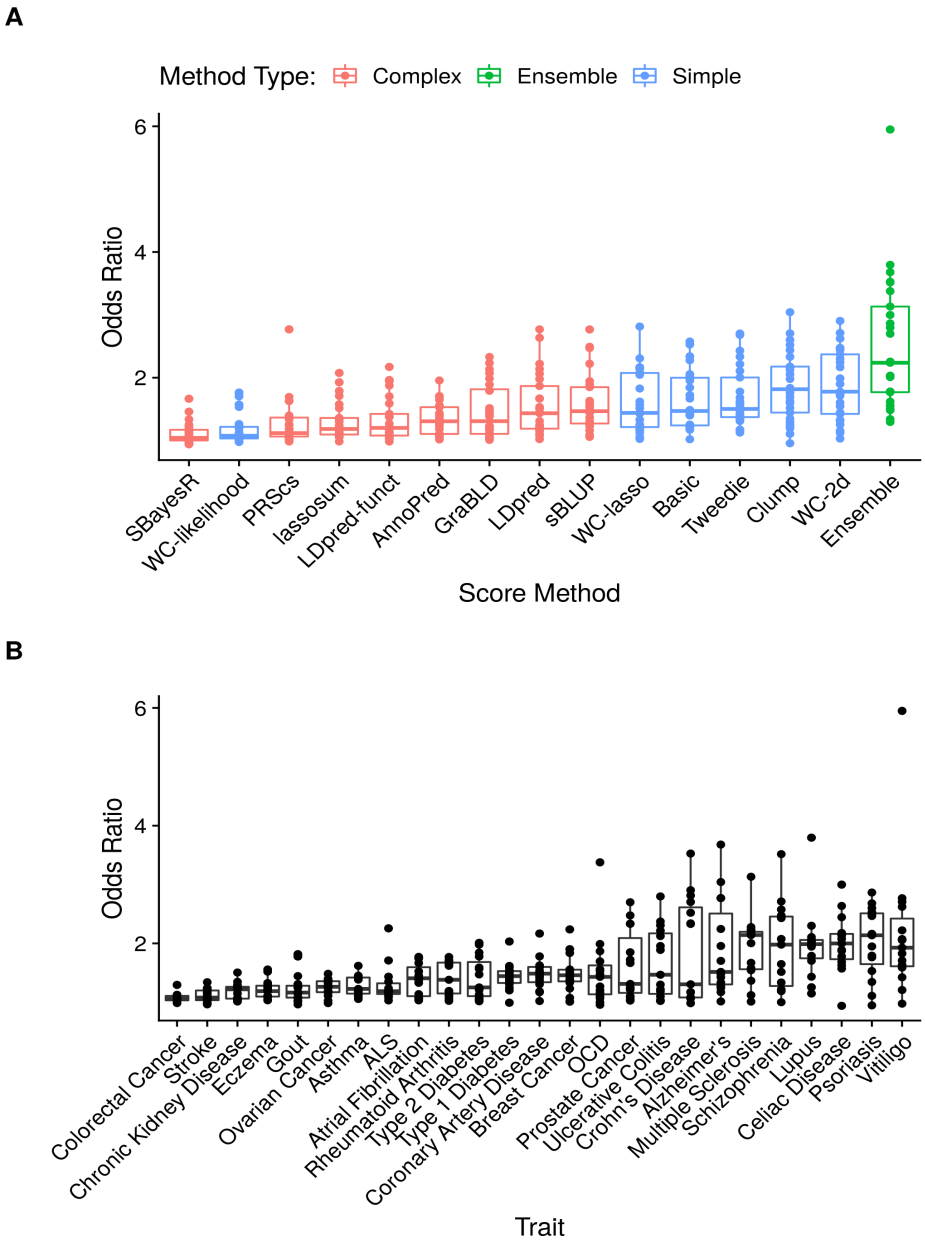

Figure 4. The odds ratios for traits and methods. Panel A depicts the odds ratios of the methods, where each point represents the value for a specific trait. Panel B depicts the odds ratios of the traits, where each point represents the value for a specific method. The odds ratios in each are from the 0.5 cut-off value

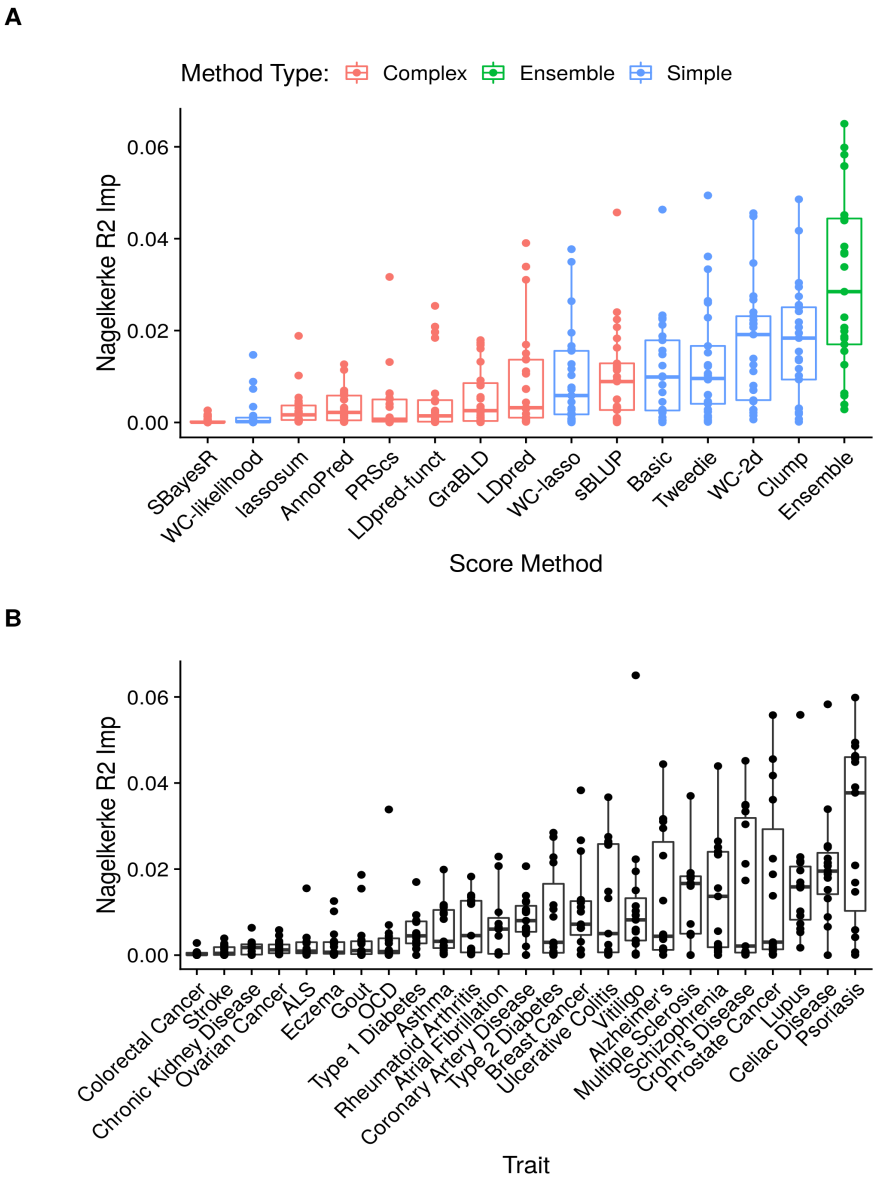

Figure 5. The Nagelkerke R^2 for traits and methods. Panel A depicts the Nagelkerke R^2 of the methods, where each point represents the value for a specific trait. Panel B depicts the Nagelkerke R^2 of the traits, where each point represents the value for a specific method.

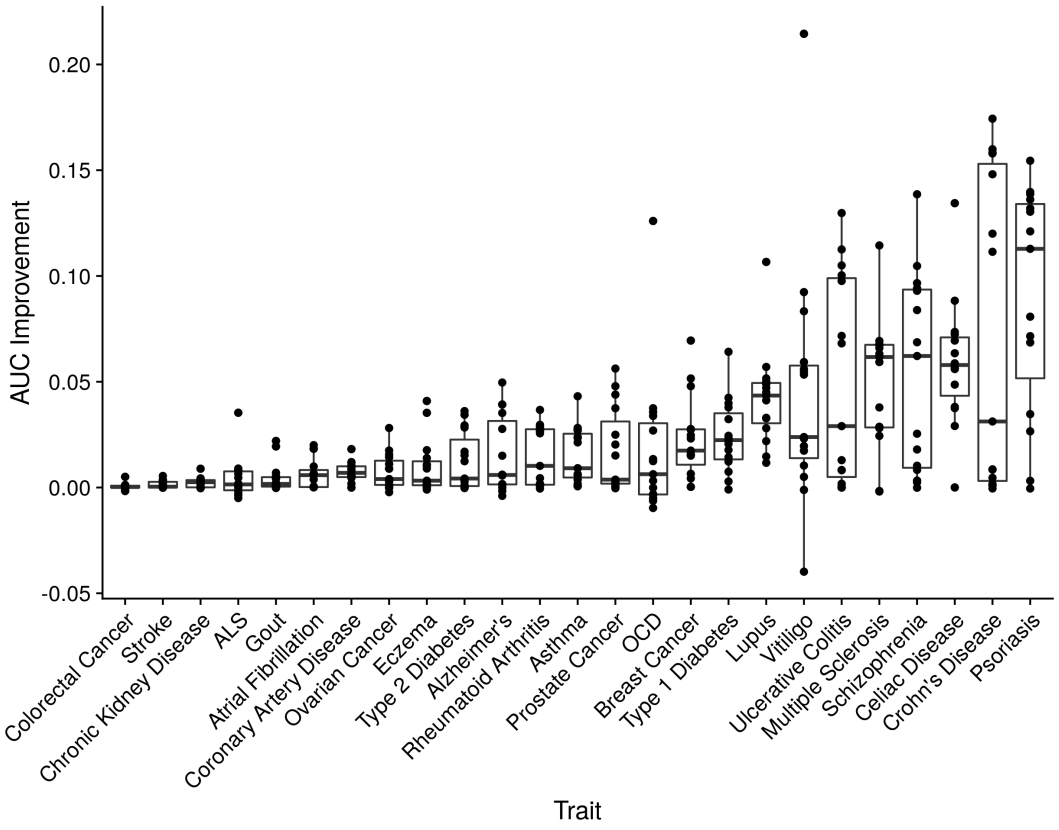

Figure 6. The AUC improvements of the traits, where each point represents the value for a specific method.

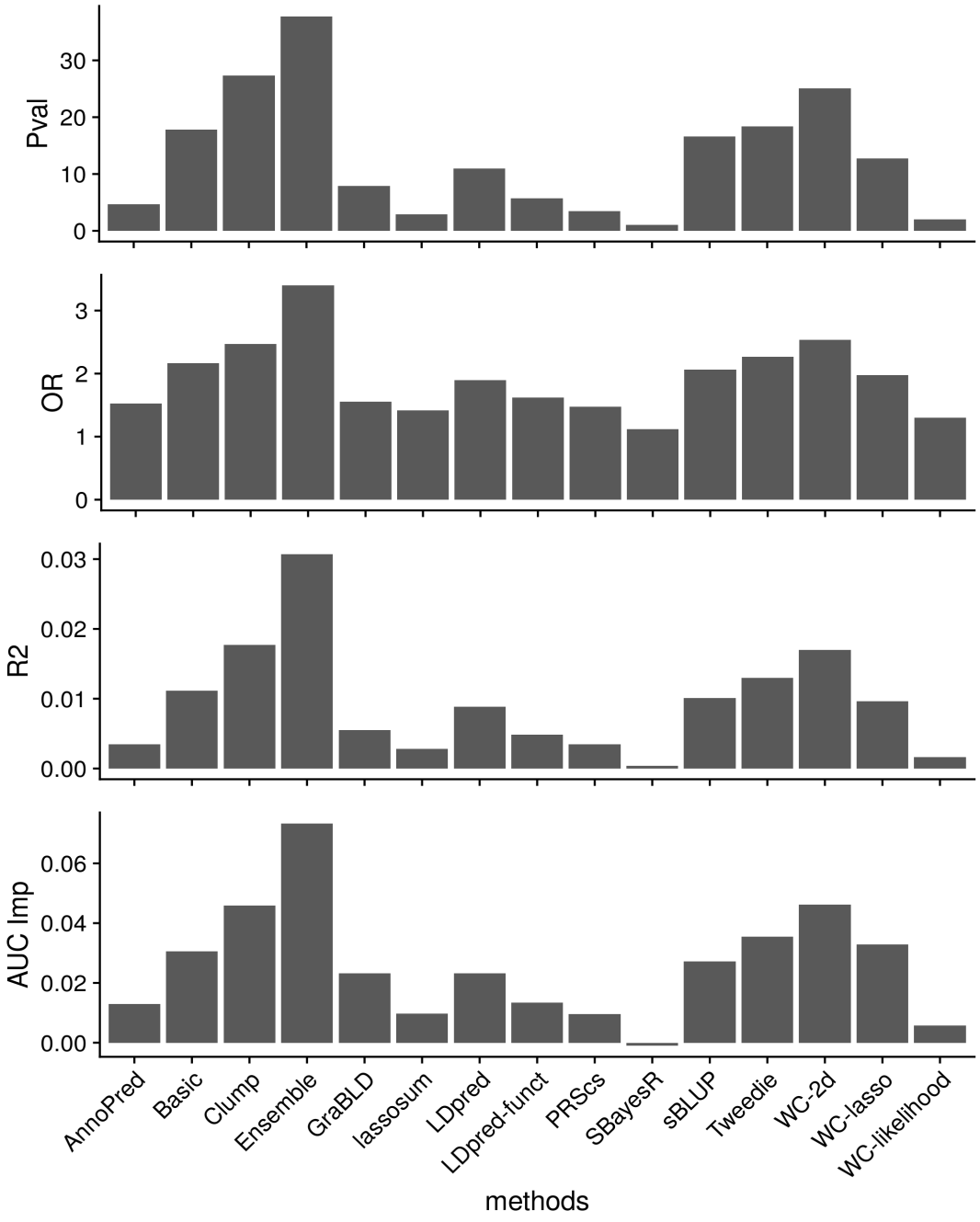

Figure 7. The performance of each method according to various statistics. For each method and statistic indicated, 25 values for each of the traits analyzed were averaged together.

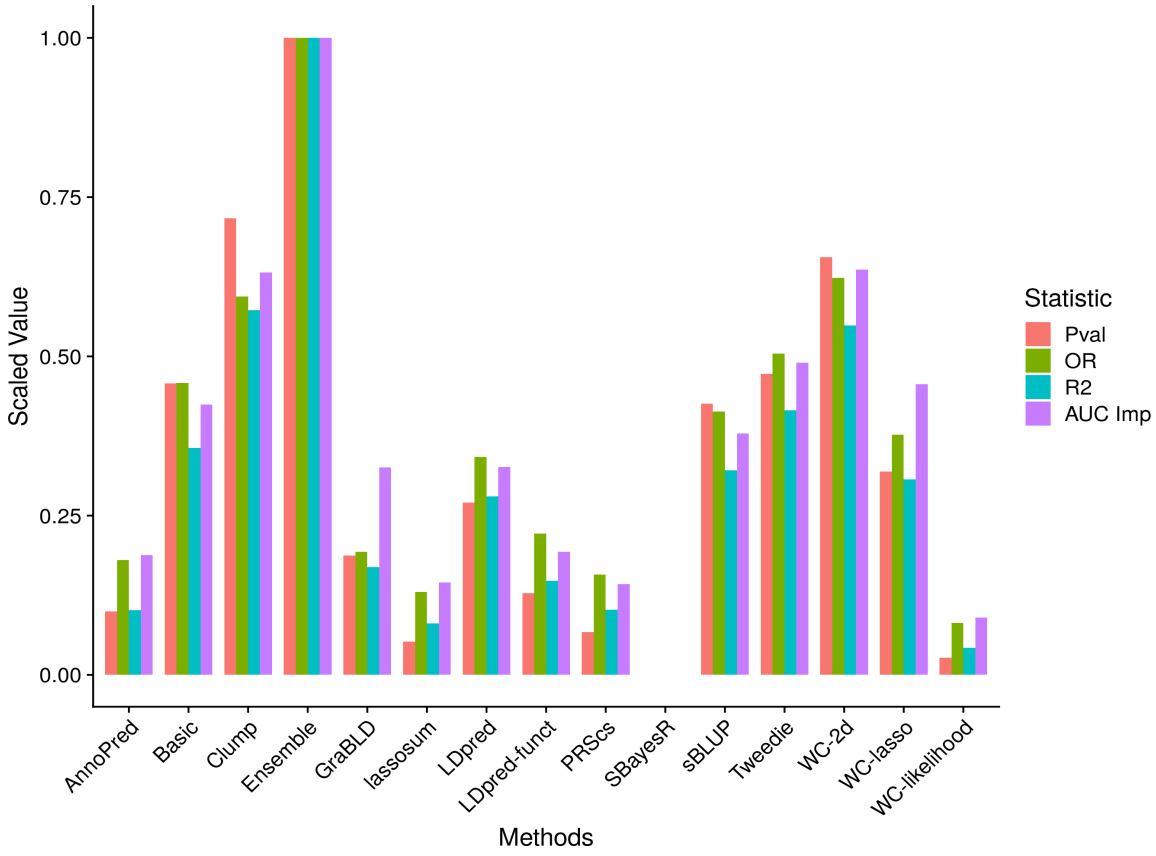

Figure 8. Scaled value of statistics across all methods. The data as described in the previous figure were scaled such that the maximum value of each statistic across all methods had a maximum of 1 and a minimum of 0.

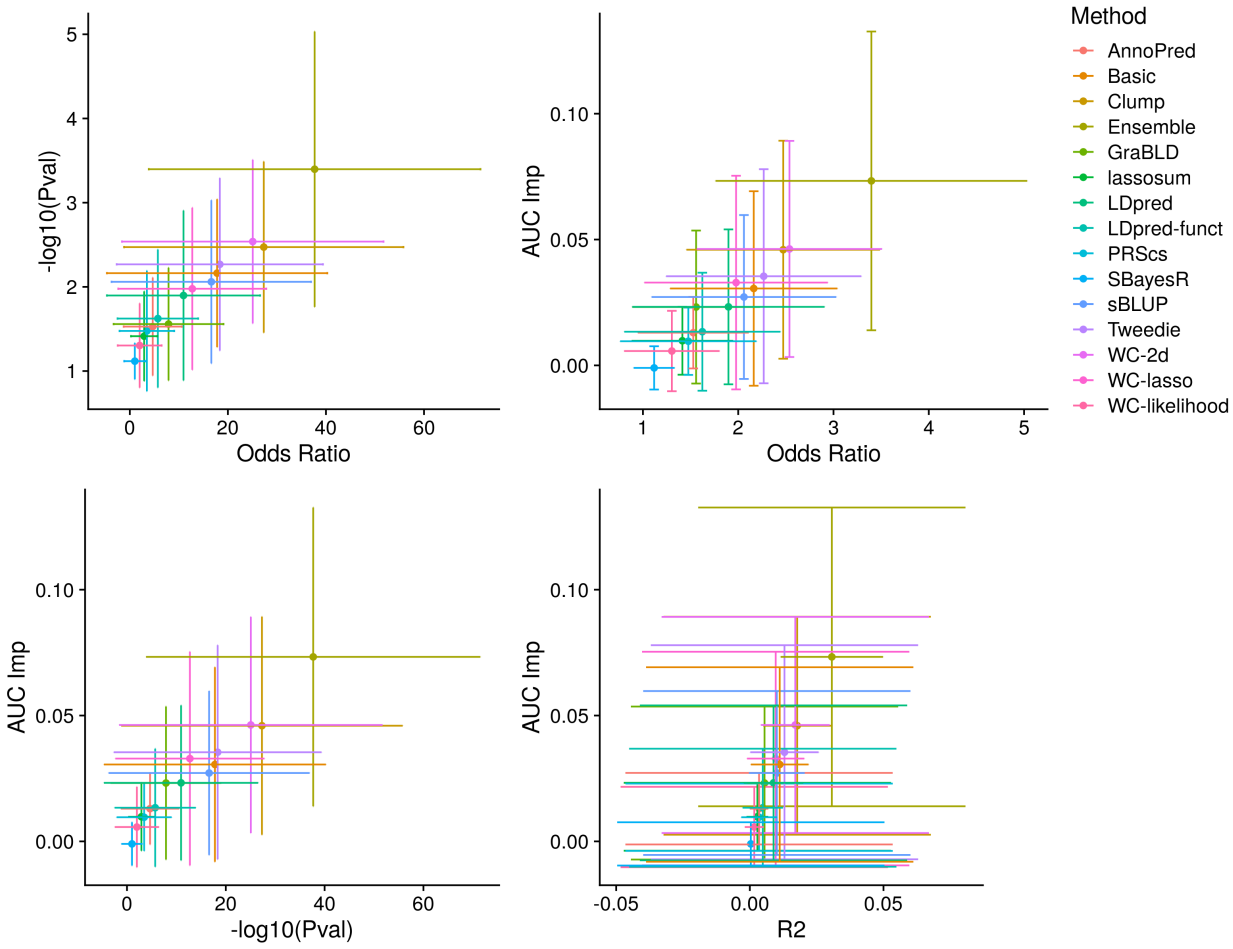

Figure 9. Pairwise comparison of assessment statistics across methods. For each method, assessment statistics for 25 traits were generated. The mean value for each statistic across this sample, with error bars denoting the standard deviation, are indicated.

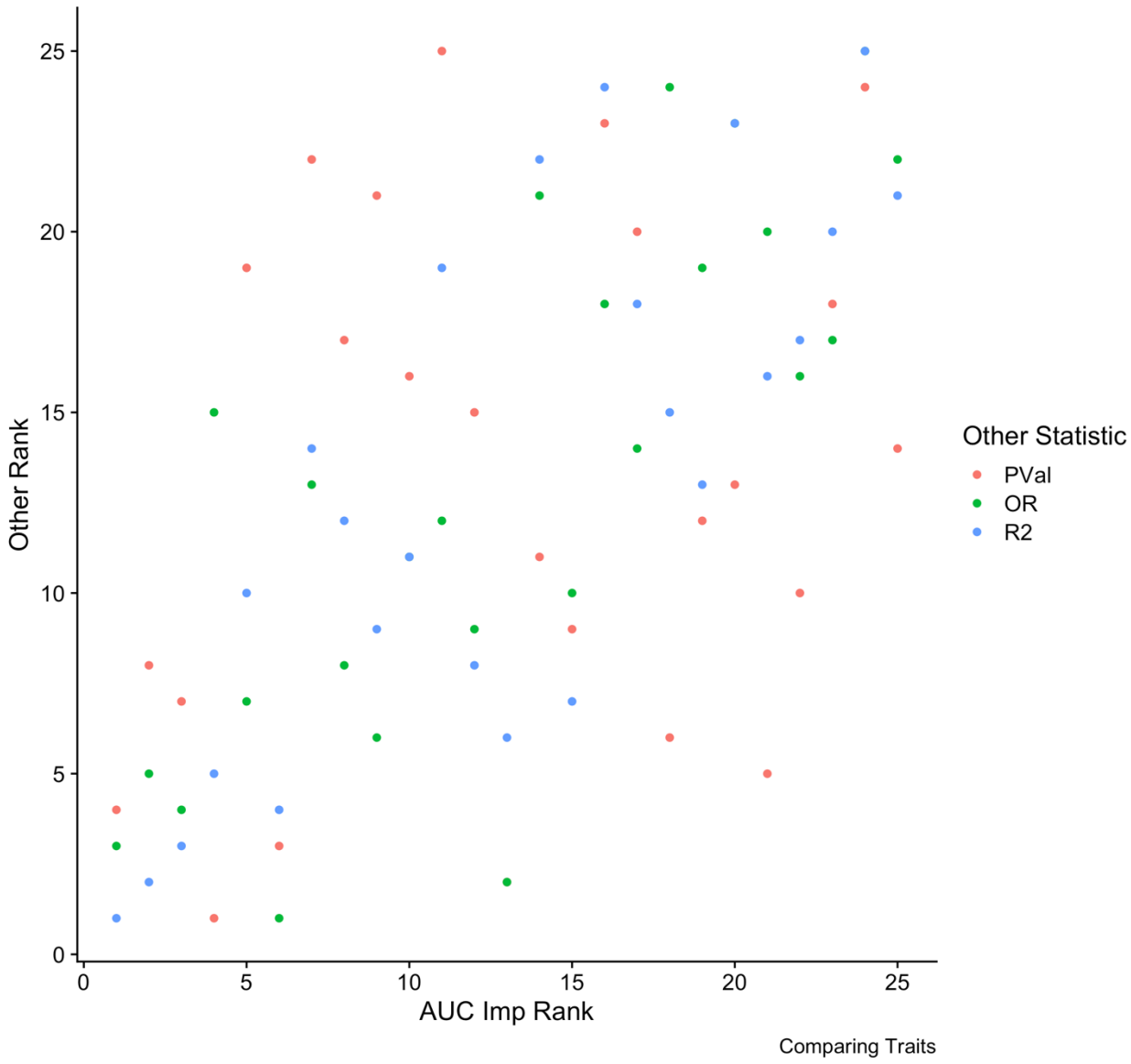

Figure 10. The variability in trait predictability ranking. The rank of each trait according to its AUC improvement value is compared against a comparable rank according to alternative statistics. All of the statistics are from the best method for each trait.

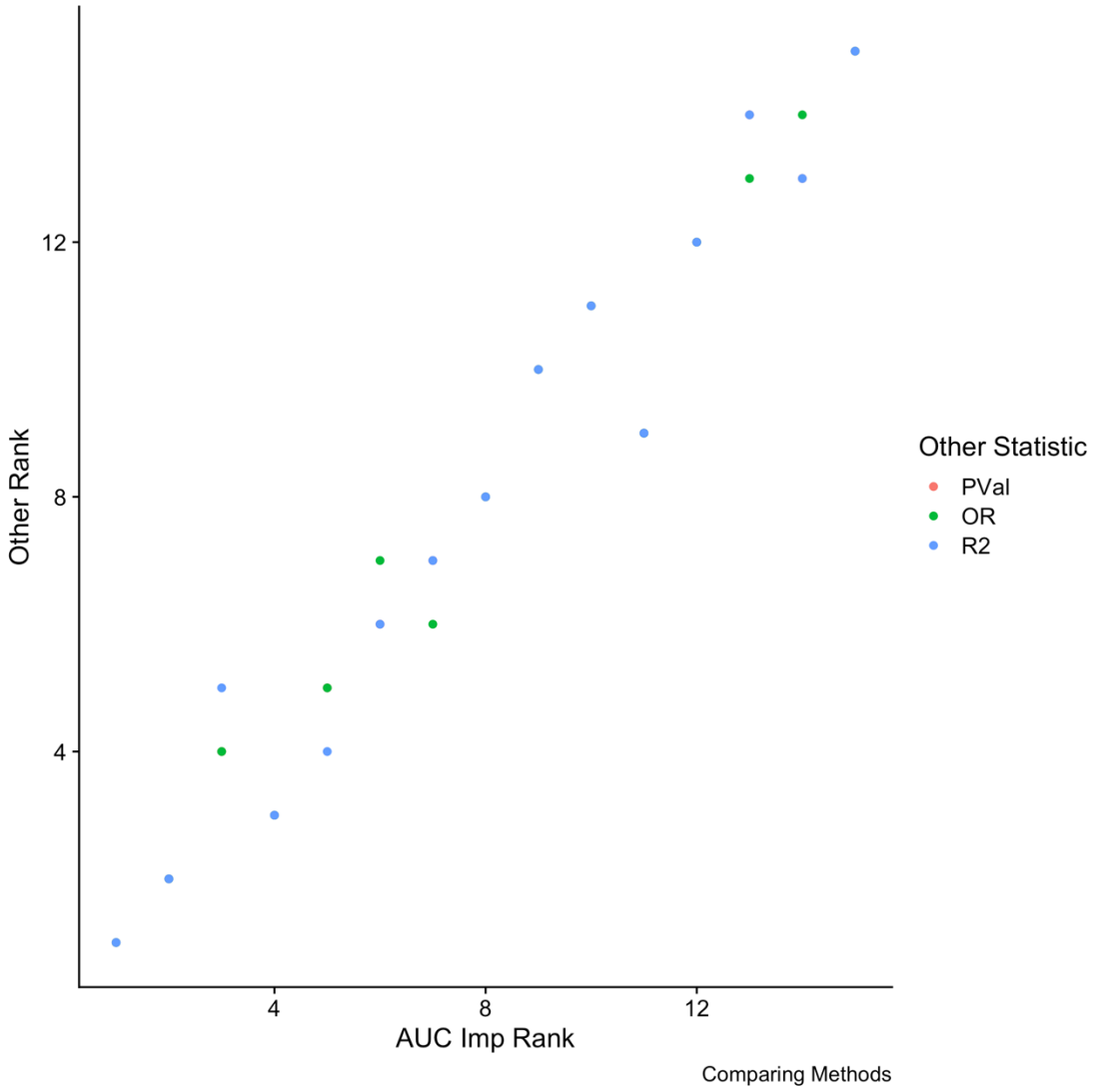

Figure 11. The variability in method predictability ranking. The rank of each method according to its average AUC improvement across traits is compared against a comparable rank according to alternative statistics.

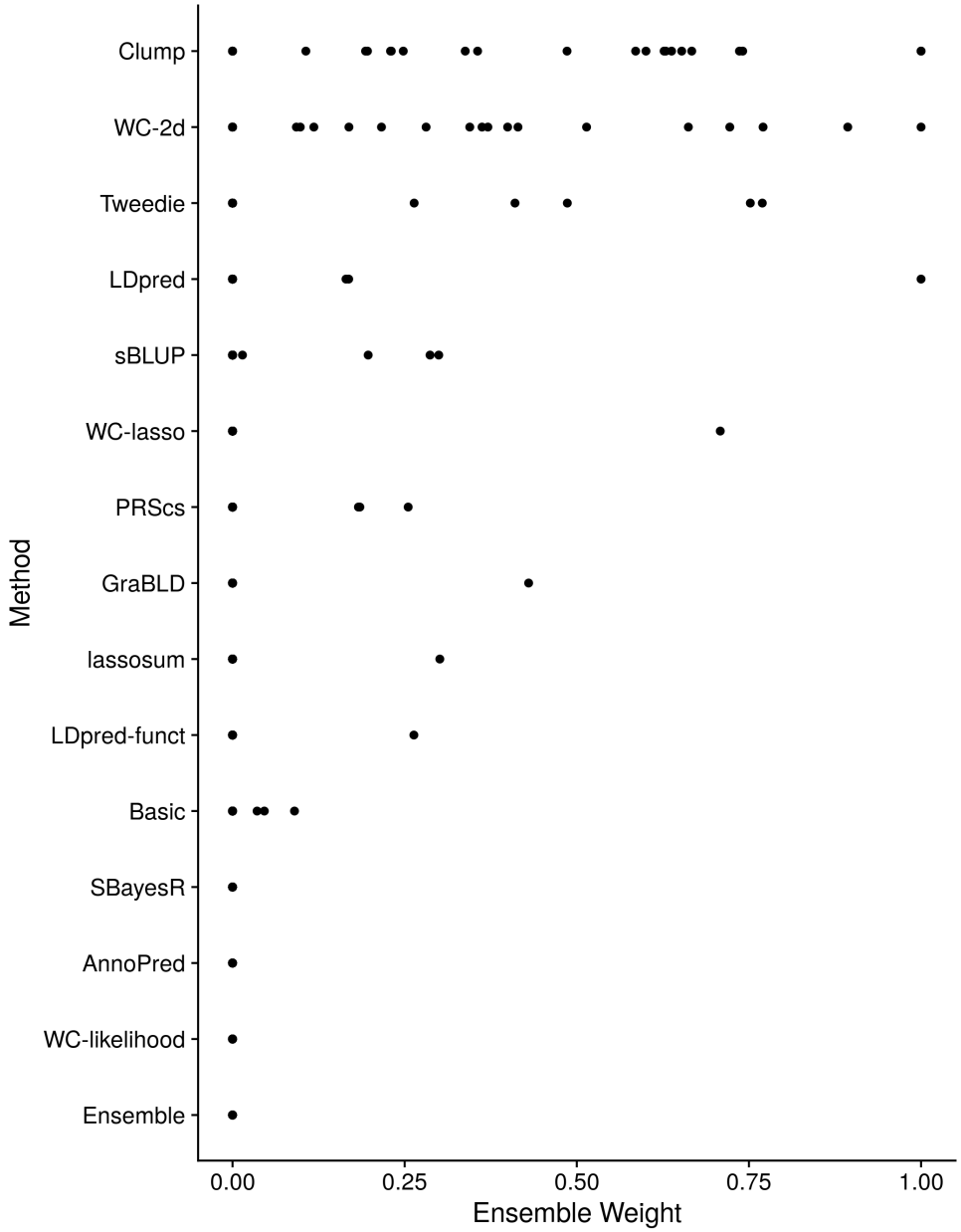

Figure 12. The ensemble weights irrespective of trait. For the various traits, the ensemble method produced weights for its 5 component polygenic risk scores. The weights of the corresponding method of each score is indicated irrespective of the trait.

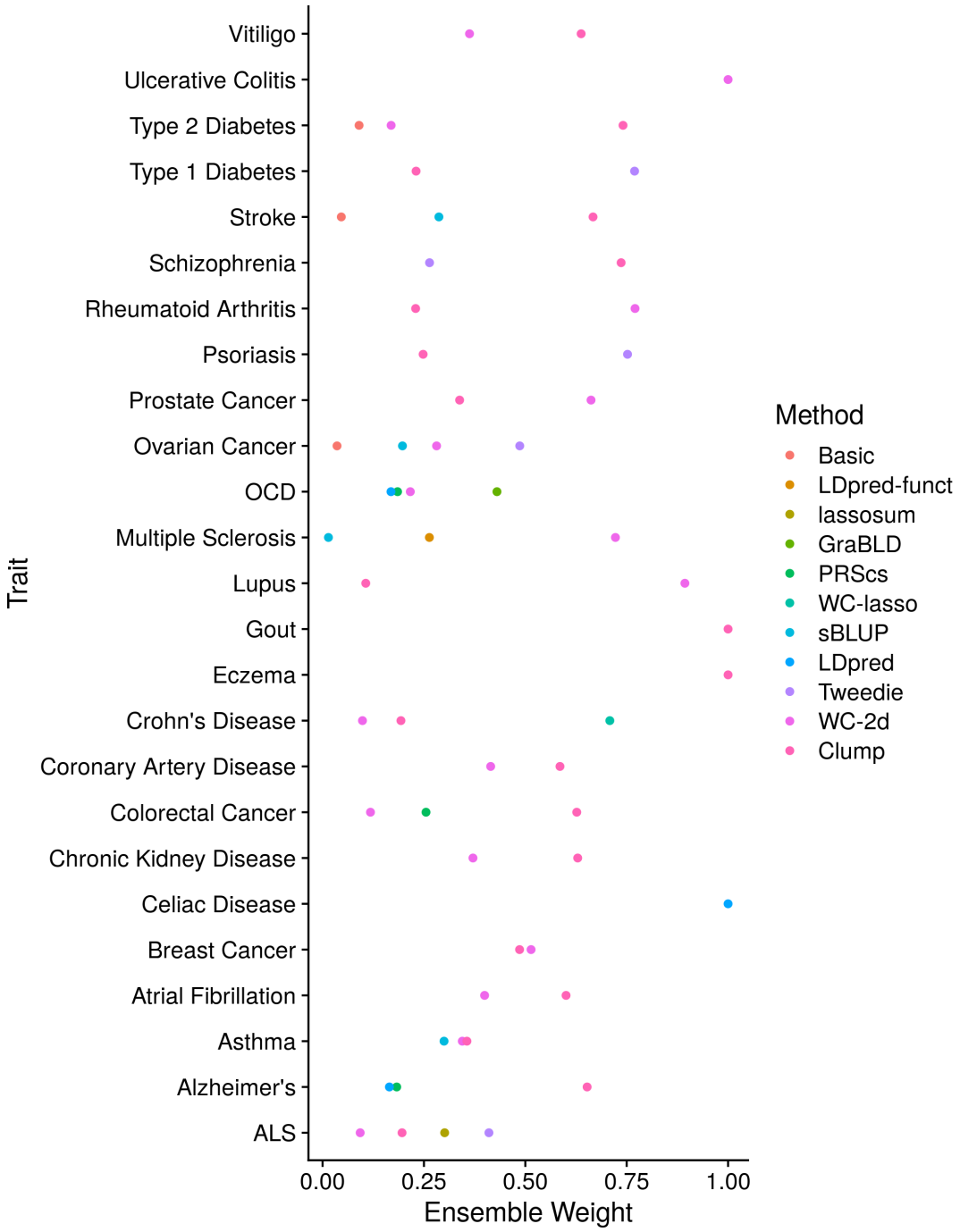

Figure 13. Ensemble weights respective of traits. This data was generated as described in the previous figure, but is now plotted with the trait the ensemble method was generated for on the y-axis and the methods that compose the scores within each ensemble model colored as described.

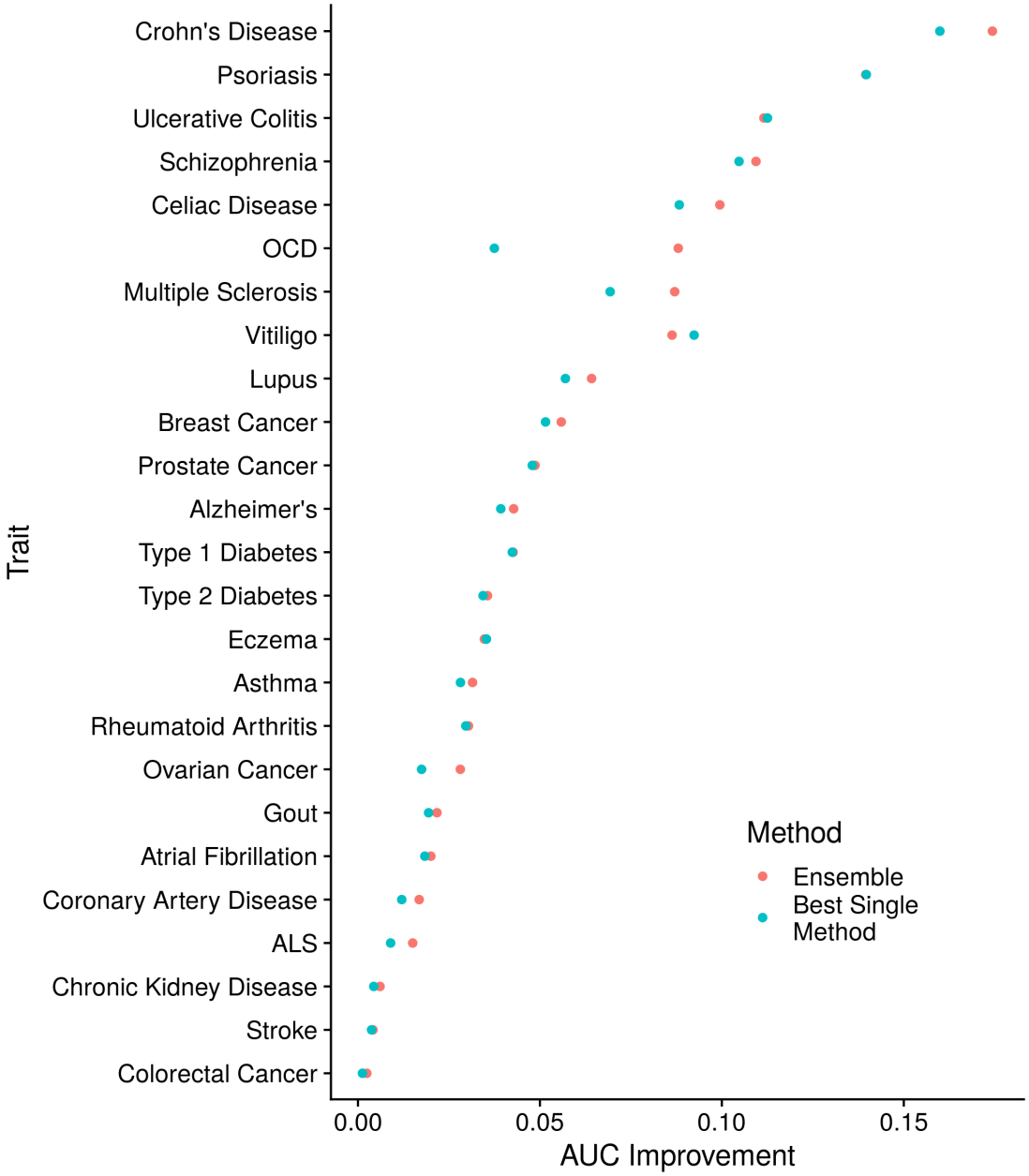

Figure 14. The improvement of the ensemble method over the next best method. For each trait the AUC improvement generated from both the score derived from the ensemble method and the polygenic risk score corresponding to the best method, are indicated.

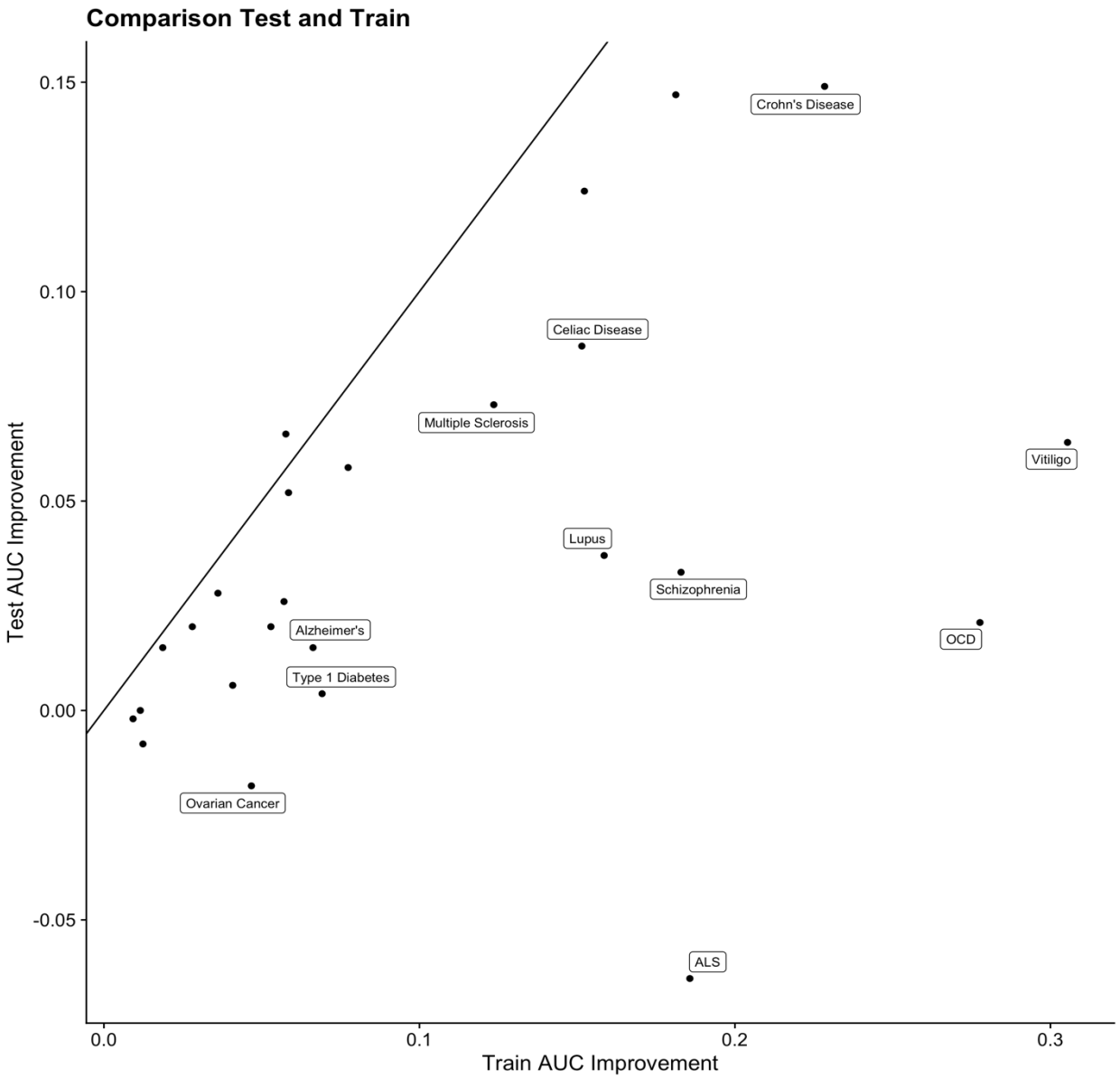

Figure 15. The comparison of training and testing AUC improvement values when considering the stackCT method. The train values are derived from the training data-set and the test values are dervied from the withheld data-set.

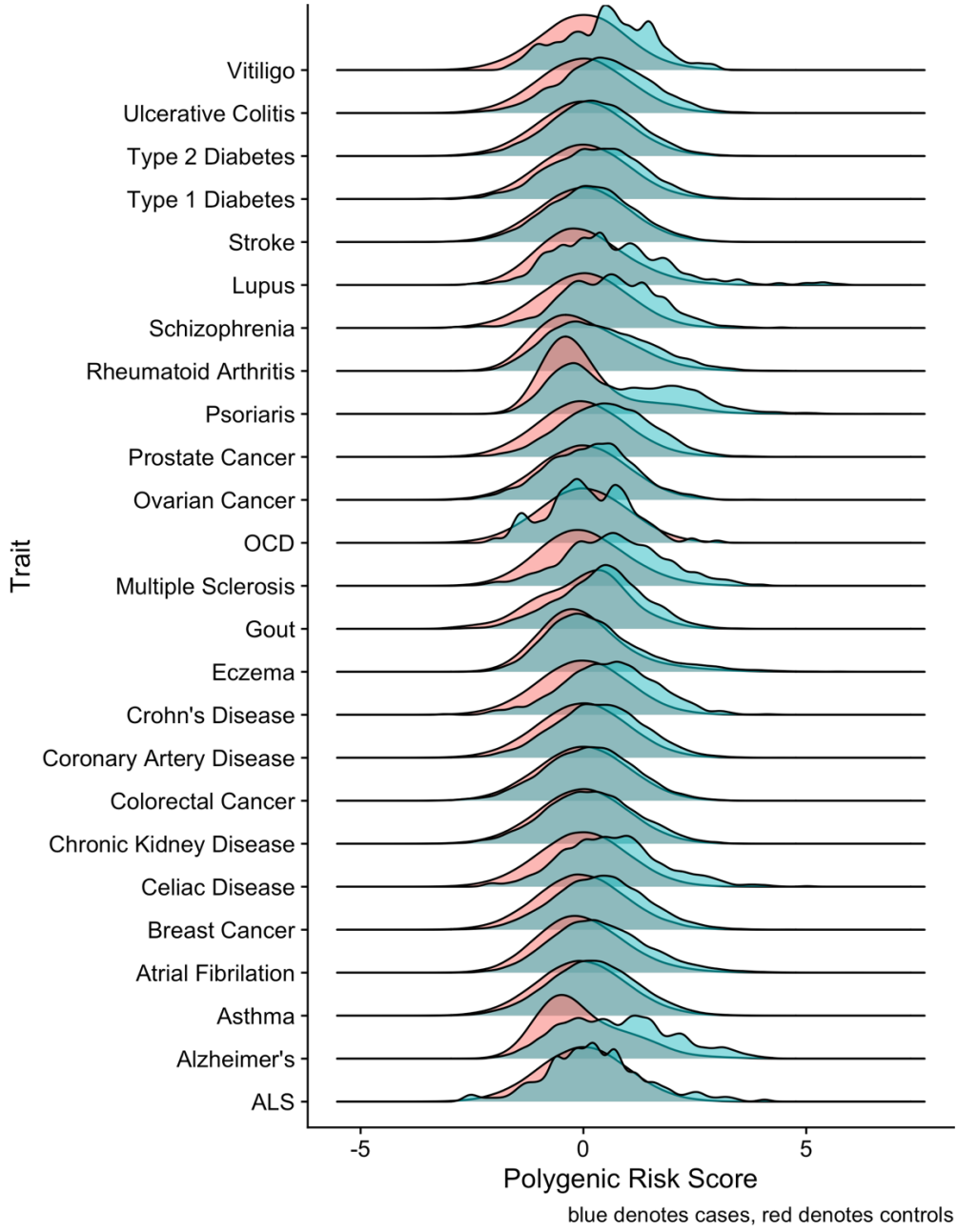

Figure 16. Case and control distribution within the withheld data-set for each trait. The blue color indicates cases, and red controls (or non-cases). The height of each shape indicates density.

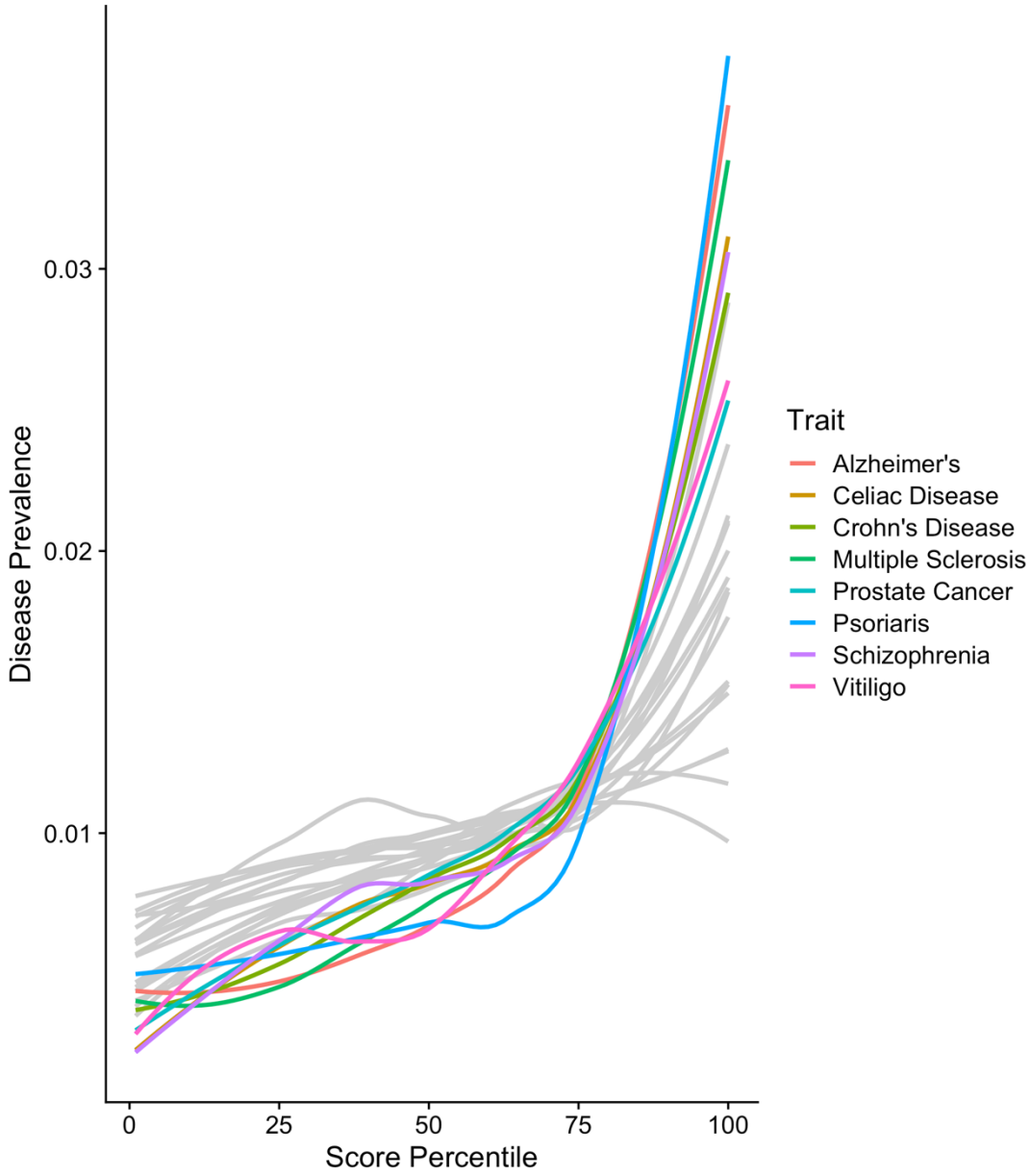

Figure 17. The prevalence of each trait at each percentile of the polygenic risk score, within the withheld data-set. The polygenic risk score is used to separate the population into 100 equally sized groups, and the prevalence or number of cases divided by total size of the group is calculated. The line plotted shows a smooth interpolation of all 100 prevalence values.

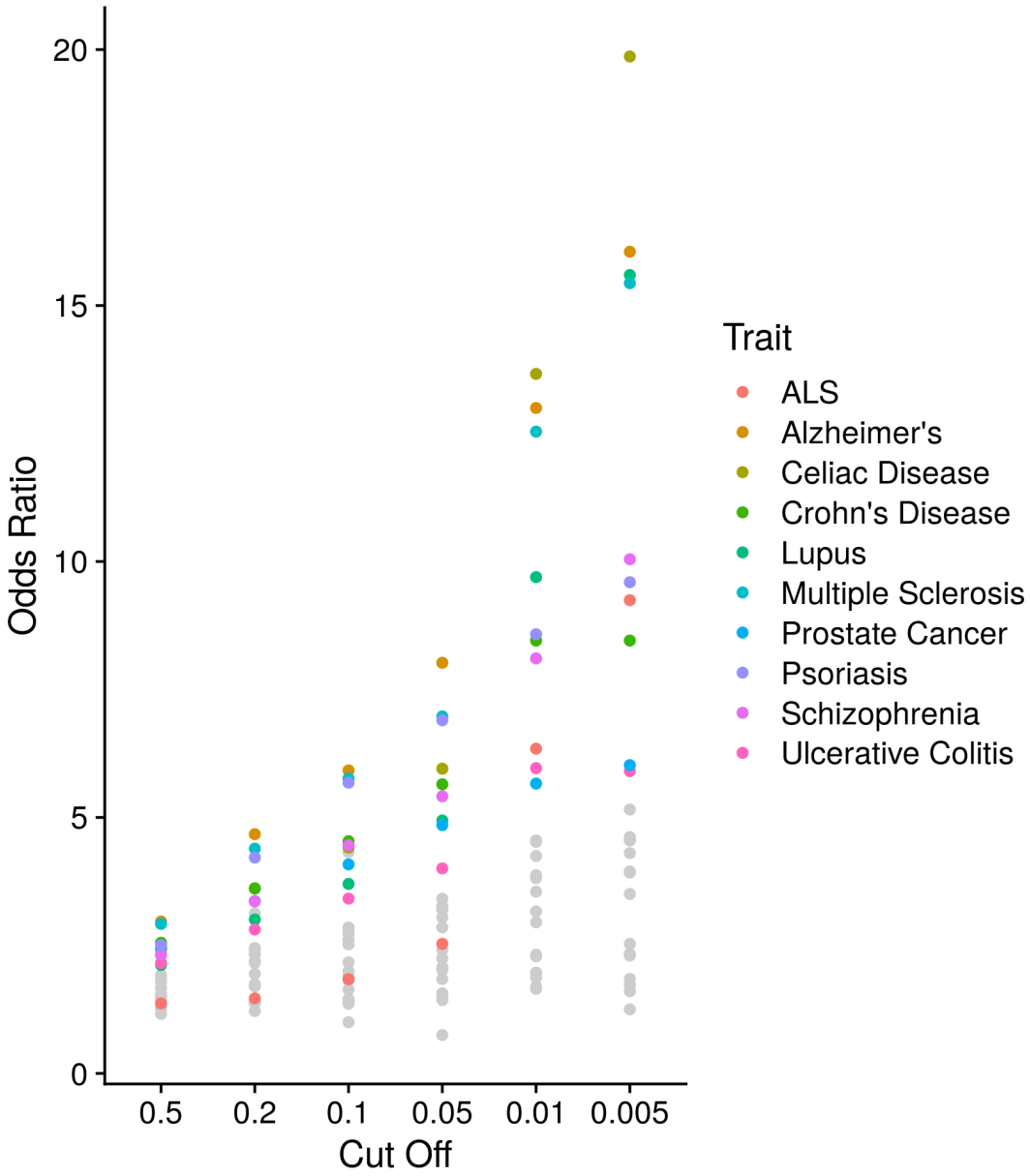

Figure 18. The odds ratios for all traits. At each cut-off point the odds ratio was calculated by considering everyone with a polygenic risk score above the cut-off percentile to be in the exposed group, and everyone below the 50th percentile to be in the non-exposed group.

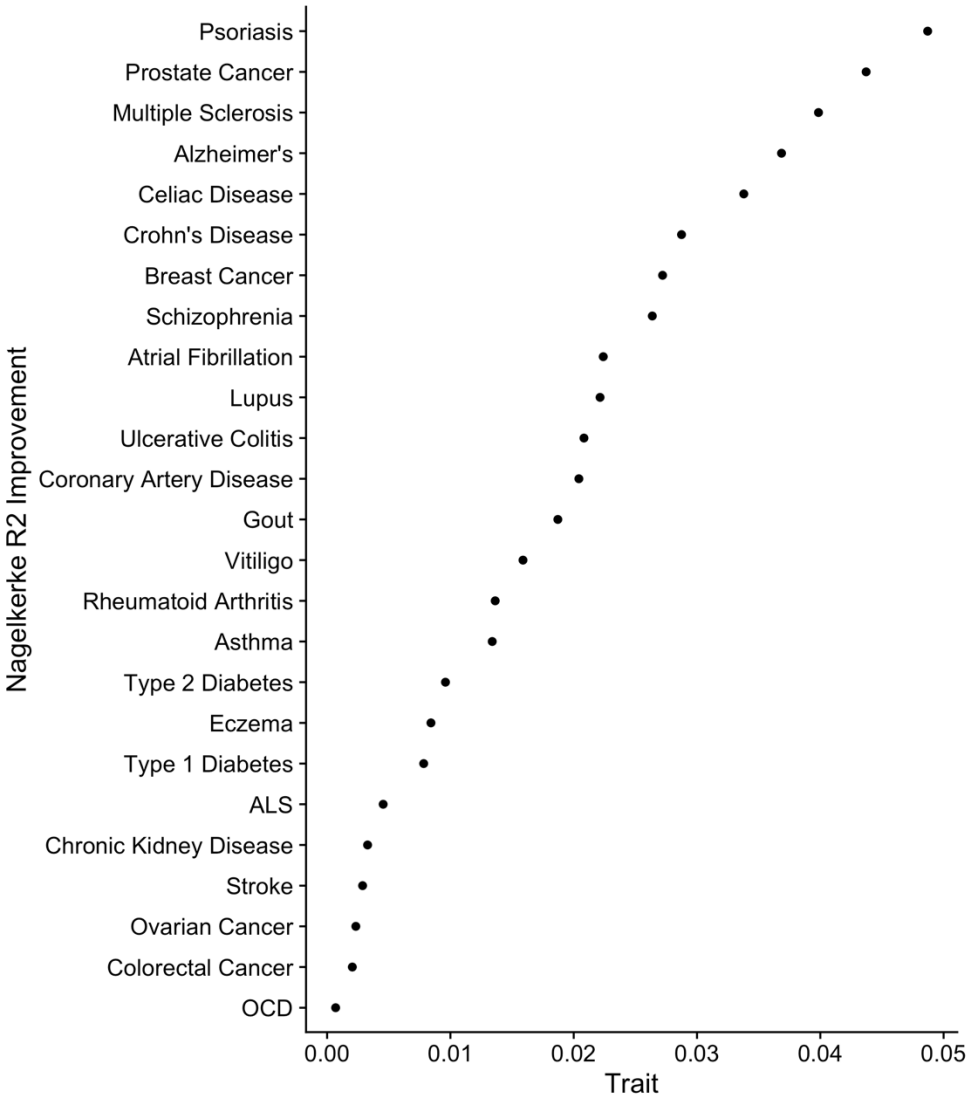

Figure 19. The improvement in Nagelgerke's R^2 through inclusion of the polygenic risk score. Specifically, the difference between the complete and covariate model, for each trait in the withheld data-set.

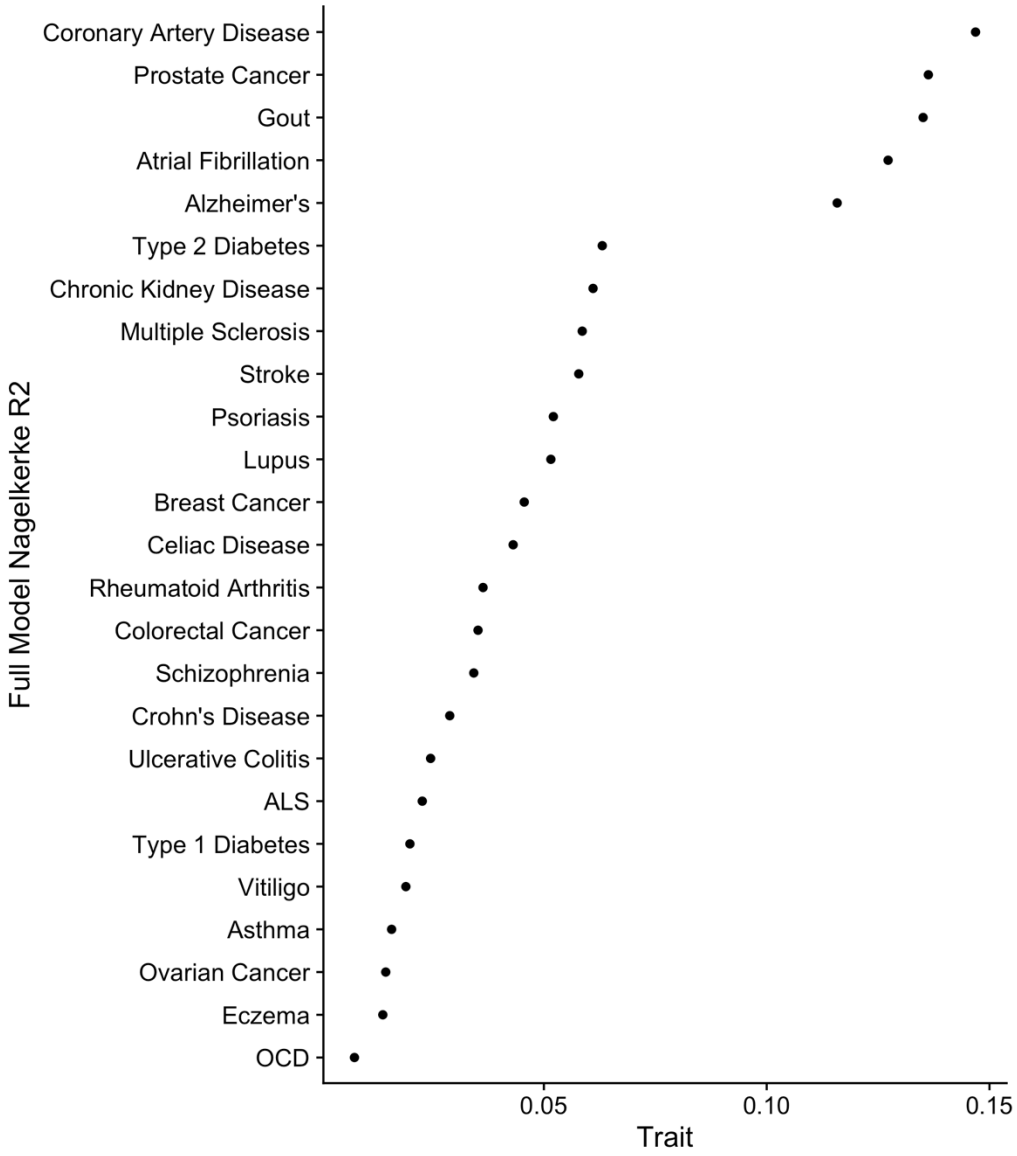

Figure 20. The total Nagelgerke's R^2 for each trait in the withheld data-set.

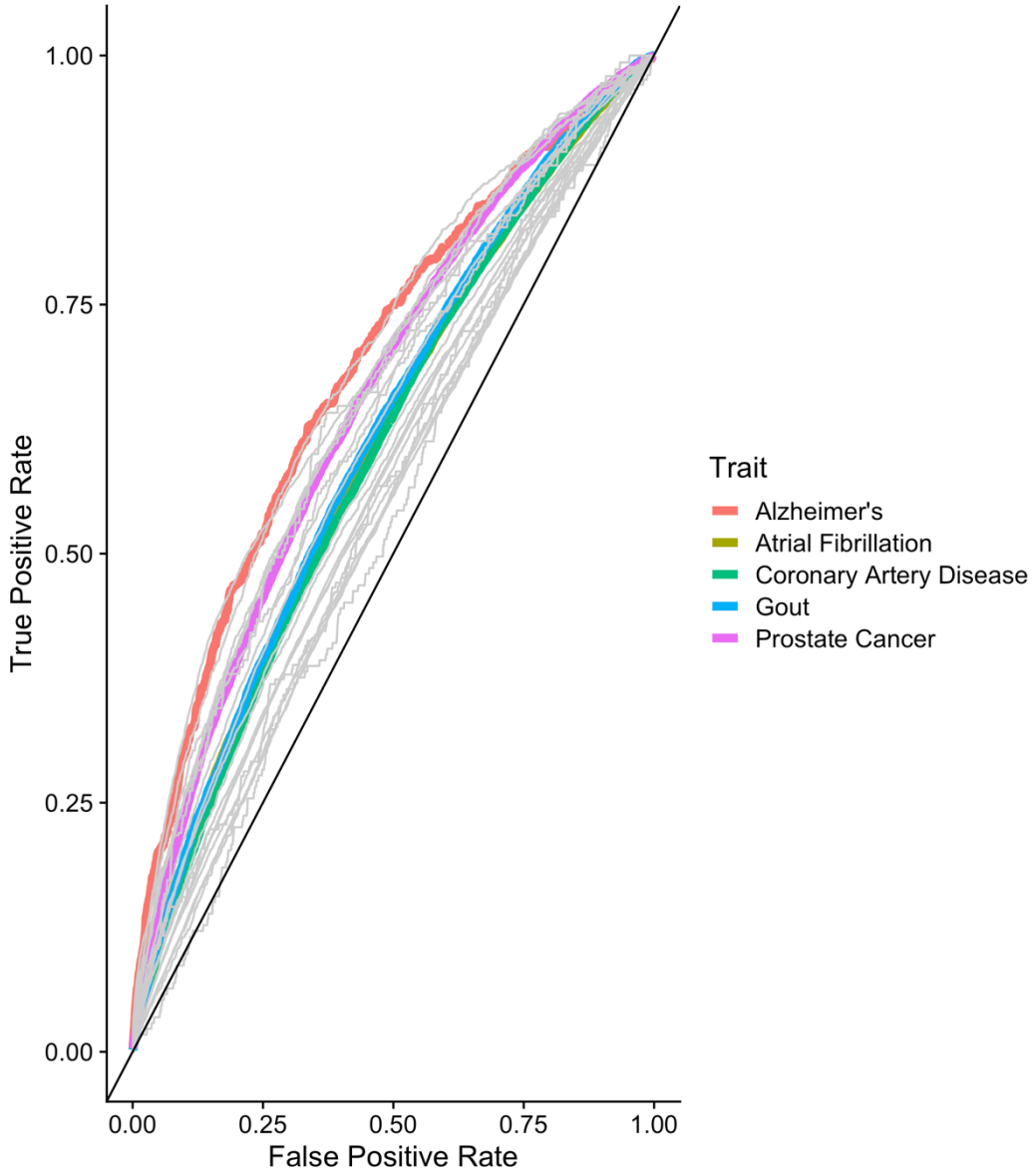

Figure 21. ROC curves for logistic regression models which only include the polygenic risk score as independent variables, for each trait in the withheld data-set.

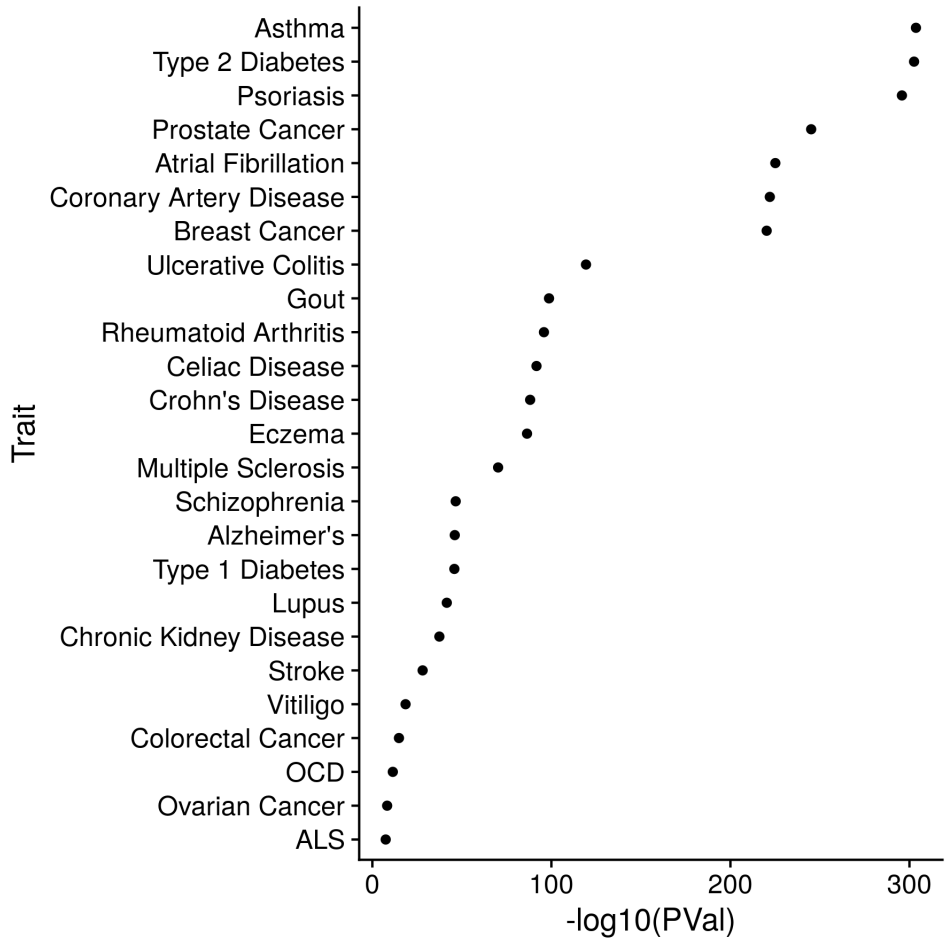

Figure 22. The P-Values corresponding to the polygenic risk score for each trait generated in a logistic regression model with all basic covariates, age, sex, and the first four principal components.

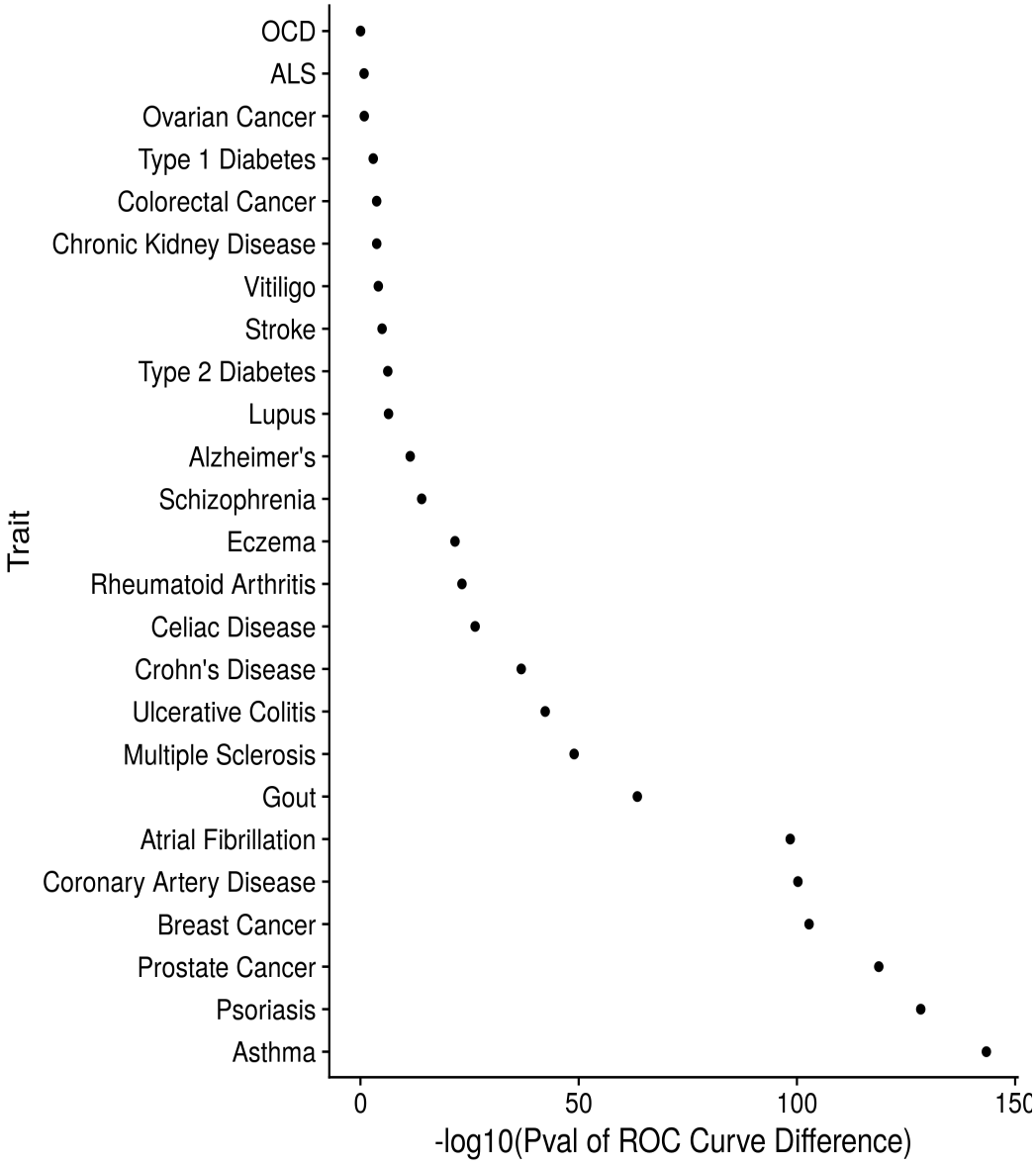

Figure 23. P-Values for each trait generated by comparing the ROC curves corresponding to the complete and covariate models within the withheld the data-set.

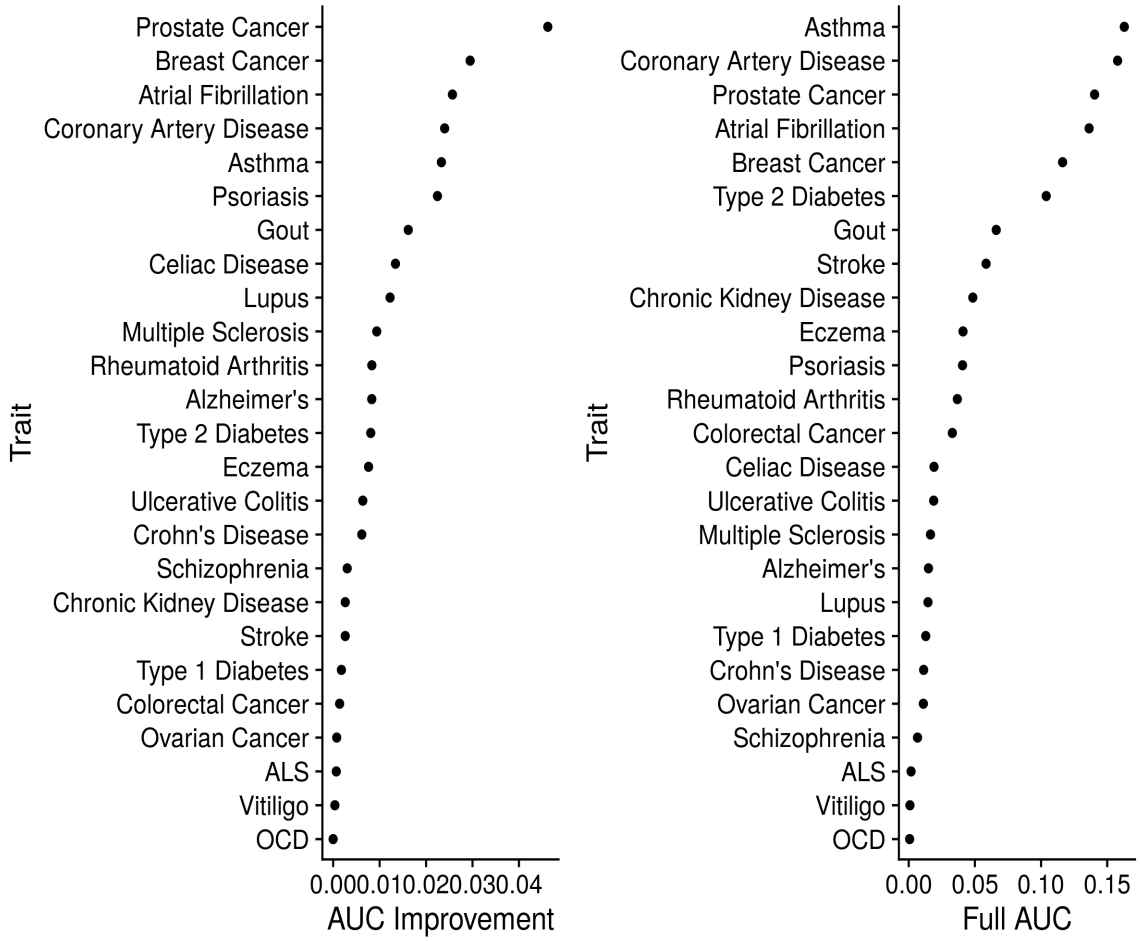

Figure 24. PR AUC Values calculated in the same fashion as the widely used ROC AUC values. Specifically, by creating two nesting logistic regression, one of which does not include the polygenic risk score, models and subtracting the difference in their outputs

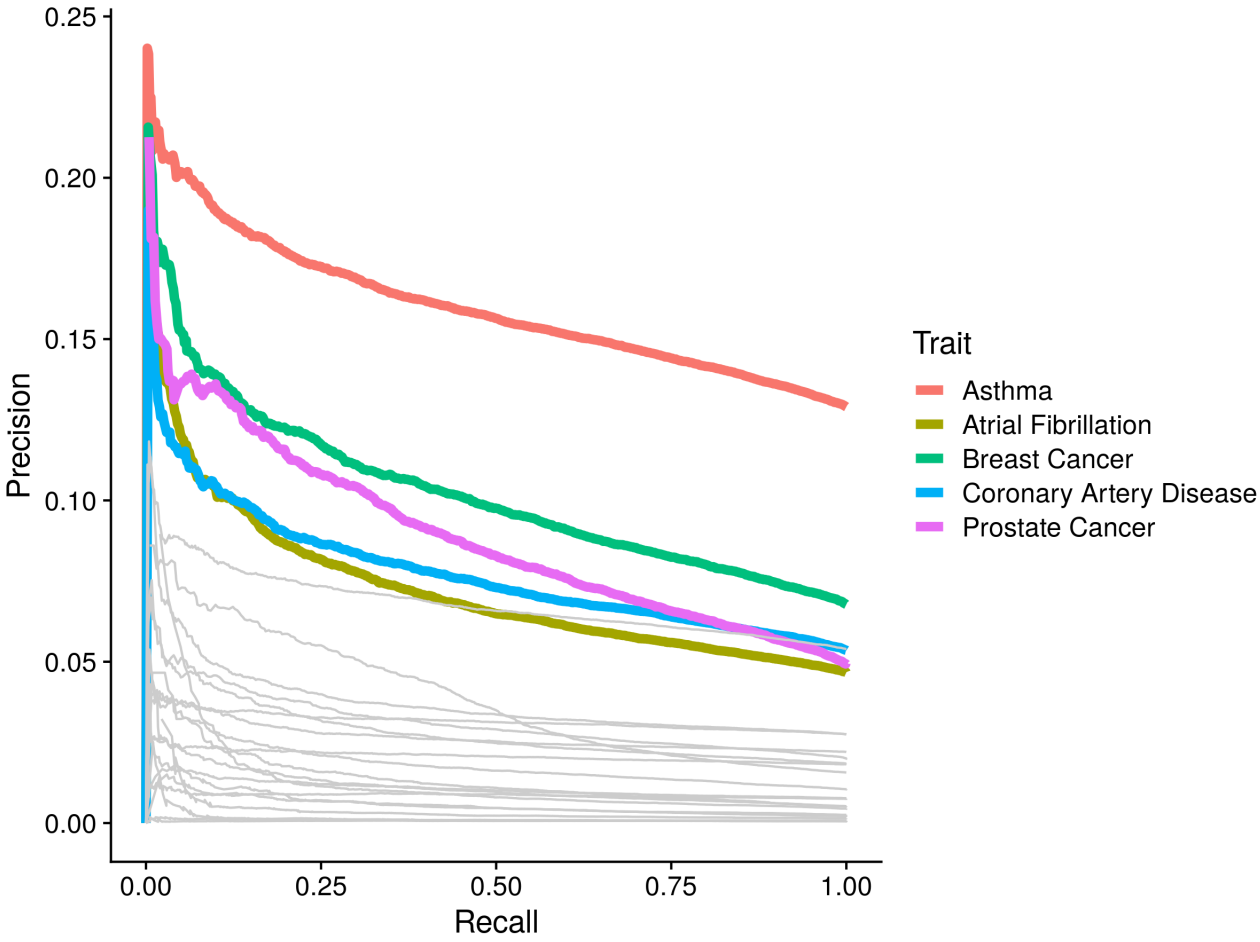

Figure 25. PR Curves generated from logistic regression models which only contained the polygenic risk score as an independent variable withheld the data-set.

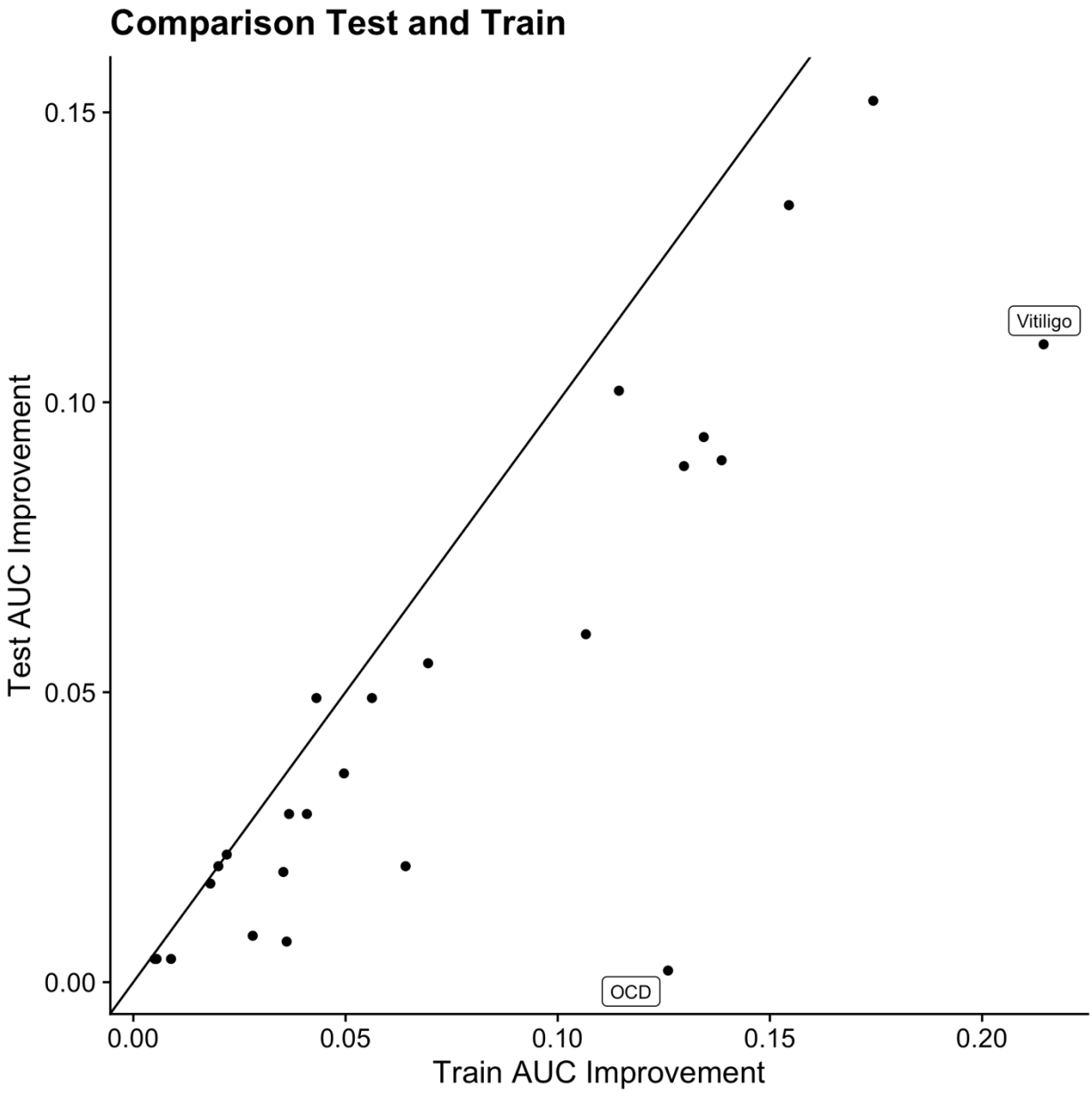

Figure 26. Comparison of the AUC improvement determined in the comparison and evaluation phase. The Train AUC Improvement is an average across three-fold cross validation within the training phase, and the test AUC improvement was generated by a logistic regression model fit on the training data and applied to the testing data. The outlier of OCD could be related to its relatively low sample size.

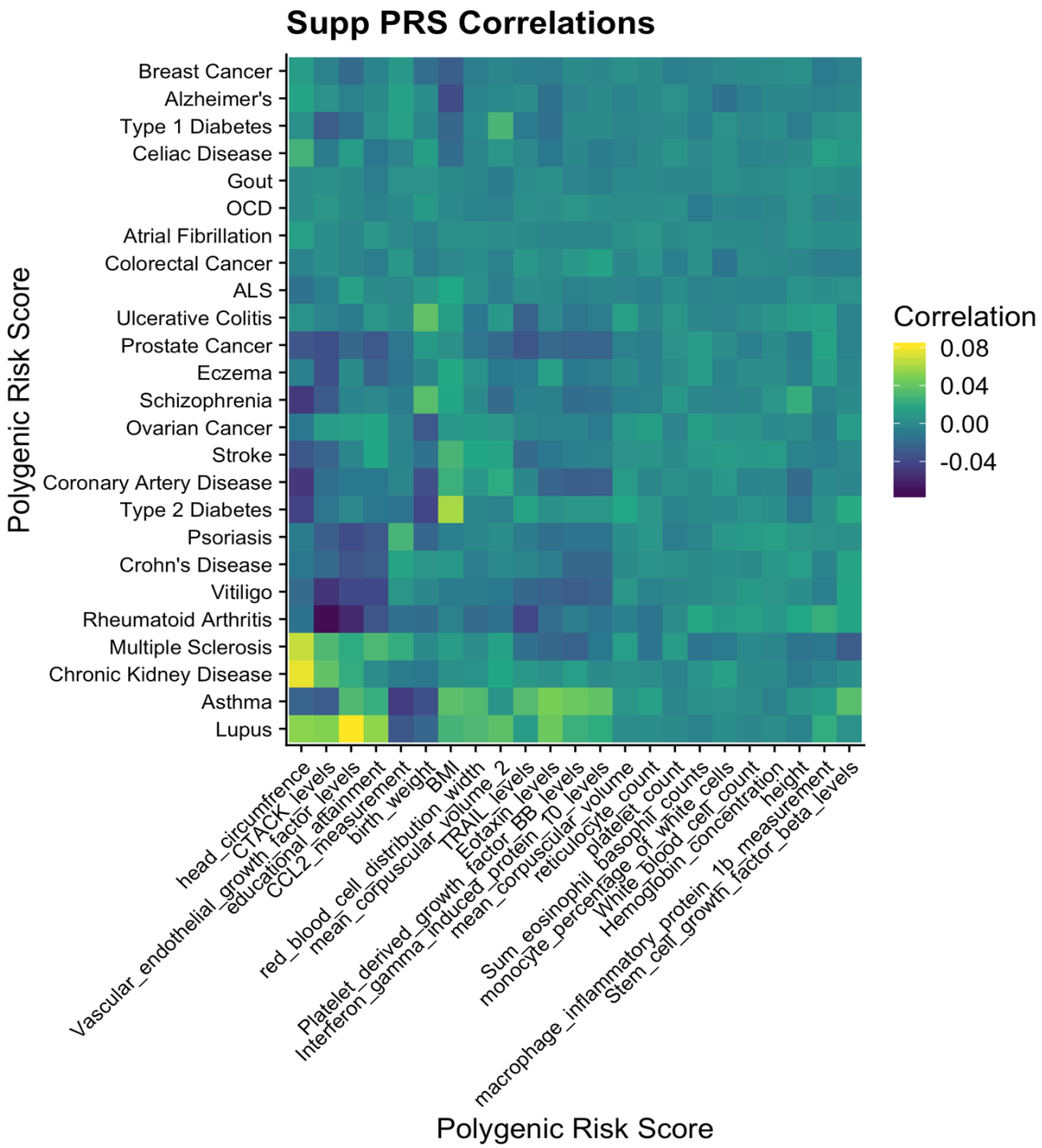

Figure 27. Pairwise correlations between the 25 primary best polygenic risk scores previously analyzed and 23 additional polygenic risk scores. The correlation is specifically Spearman's Correlation.

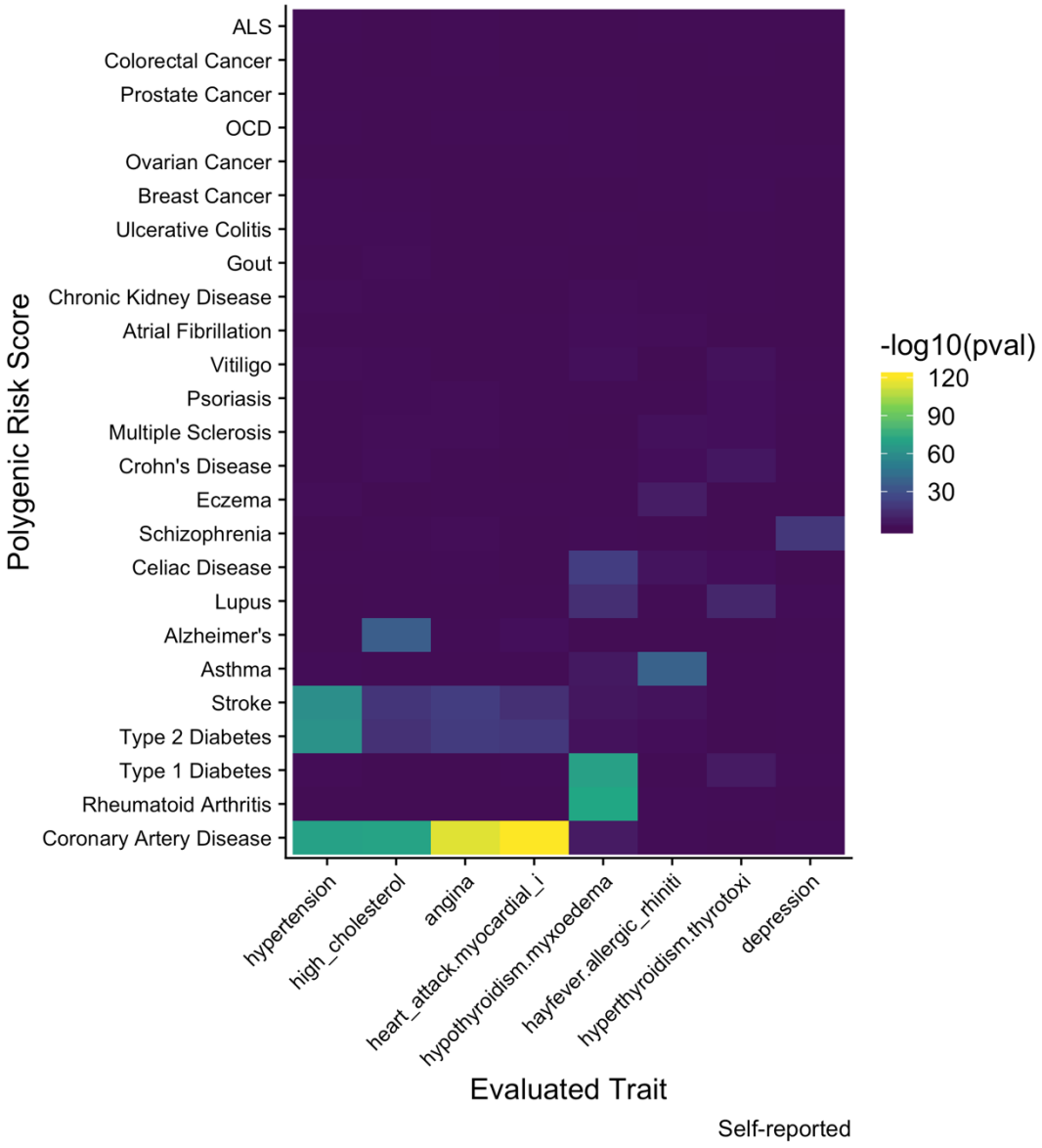

Figure 28. Association strength between the 25 main polygenic risk scores and the best 8 self-reported, unrelated evaluated traits. Within the heatmap the color of each square is determined through a logistic regression in which the polygenic risk score along with the traditional covariates are the independent variables and the evaluated trait is the dependent variable. The -log10(pval) of the polygenic risk score term from this regression is specifically indicated.

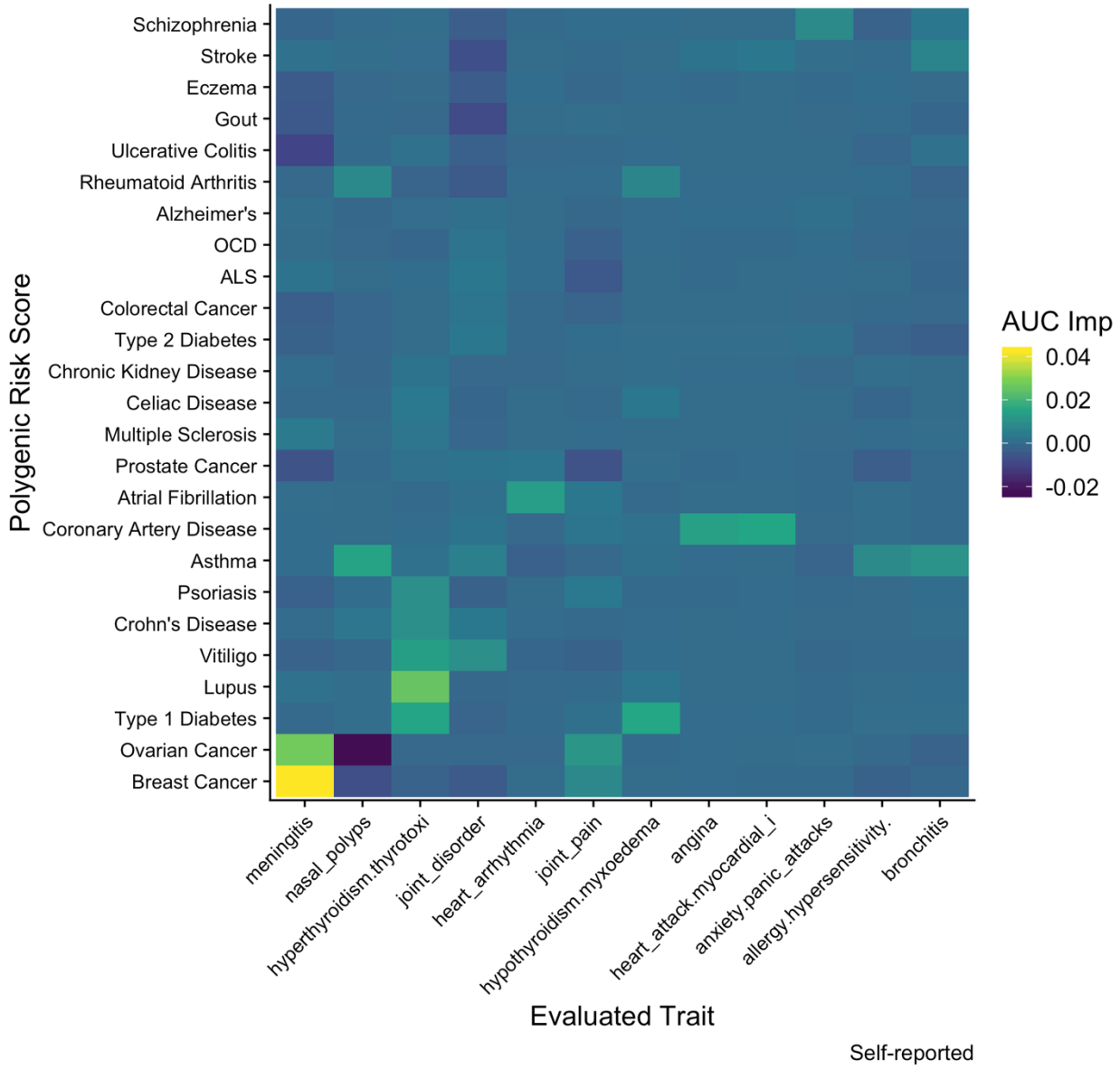

Figure 29. AUC improvement between the 25 main polygenic risk scores and the best 12 self-reported, unrelated evaluated traits. Within the heatmap the color of each square is determined through two logistic regressions with the the polygenic risk score along with the traditional covariates are the independent variables in the full model and just the traditional covariates are independent variables in the smaller model. In both models the evaluated trait is the dependent variable. The AUC improvement, determined when applying the models to the validation phase and taking the difference in AUC values, is specifically indicated.

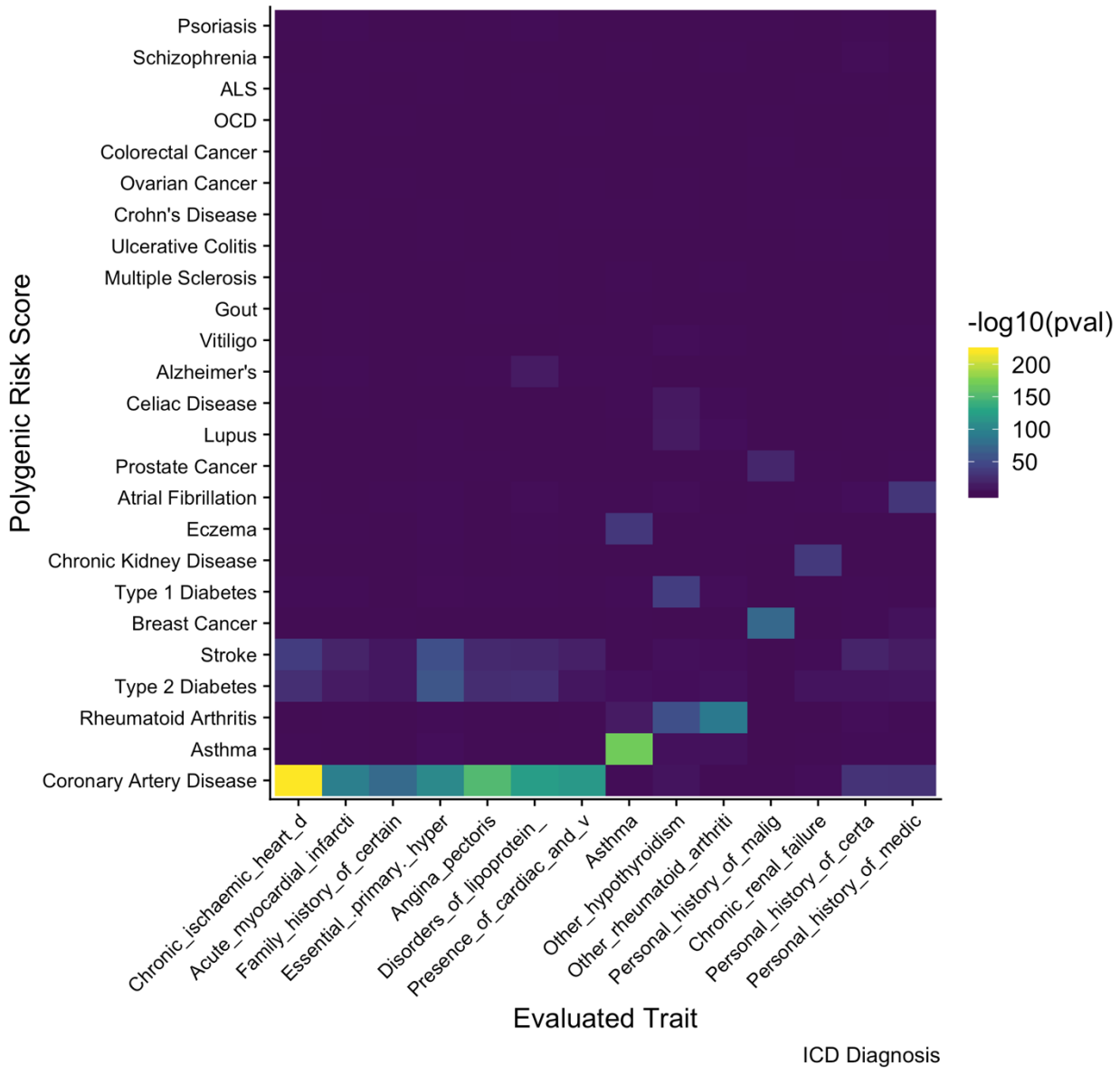

Figure 30. Association strength between the 25 main polygenic risk scores and the best 14 ICD-derived, unrelated evaluated traits. Within the heatmap the color of each square is determined through a logistic regression in which the polygenic risk score along with the traditional covariates are the independent variables and the evaluated trait is the dependent variable. The -log10(pval) of the polygenic risk score term from this regression is specifically indicated.

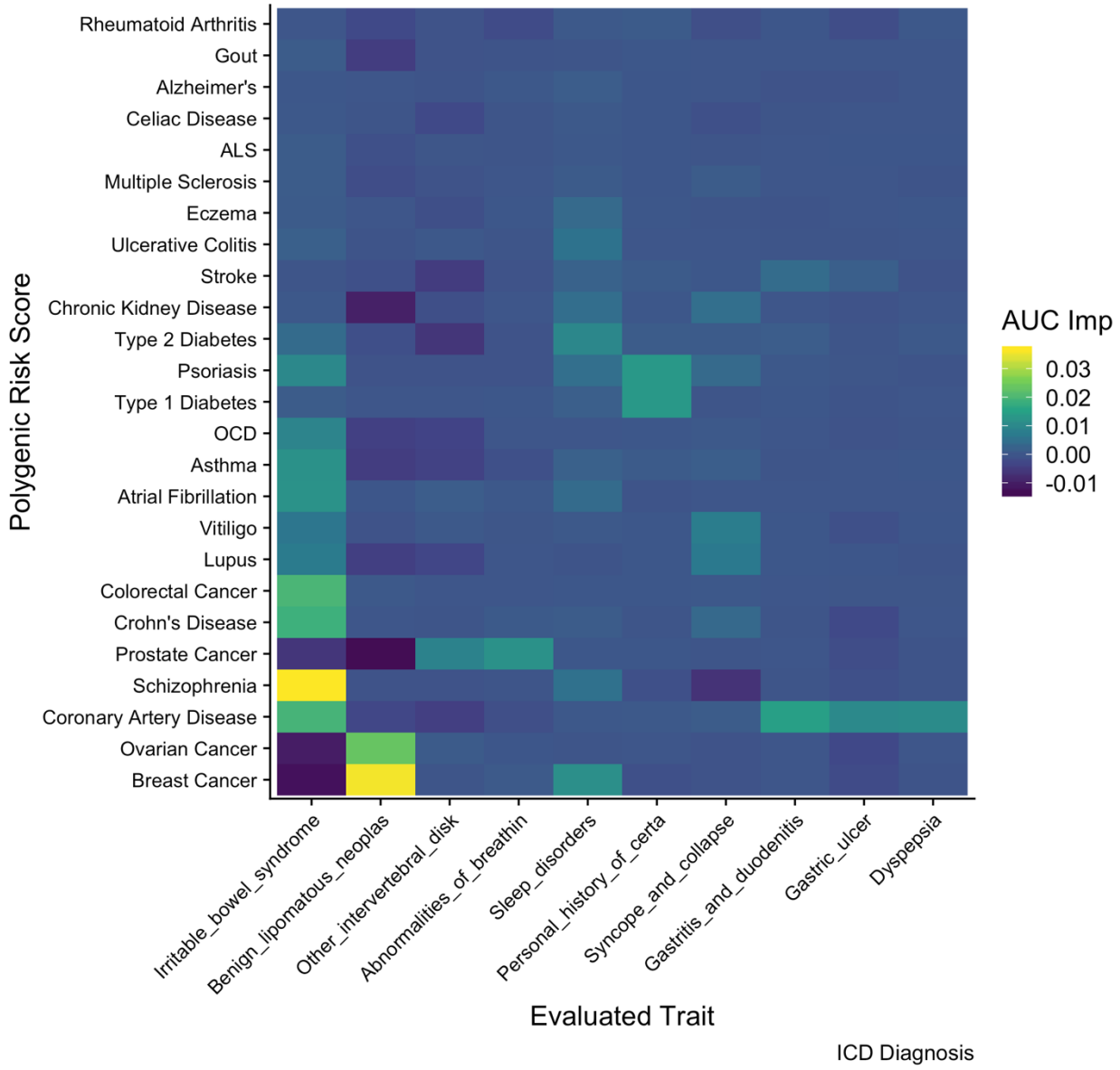

Figure 31. AUC improvement between the 25 main polygenic risk scores and the best 10 ICD-derived, unrelated evaluated traits. Within the heatmap the color of each square is determined through two logistic regressions with the the polygenic risk score along with the traditional covariates are the independent variables in the full model and just the traditional covariates are independent variables in the smaller model. In both models the evaluated trait is the dependent variable. The AUC improvement, determined when applying the models to the validation phase and taking the difference in AUC values, is specifically indicated.

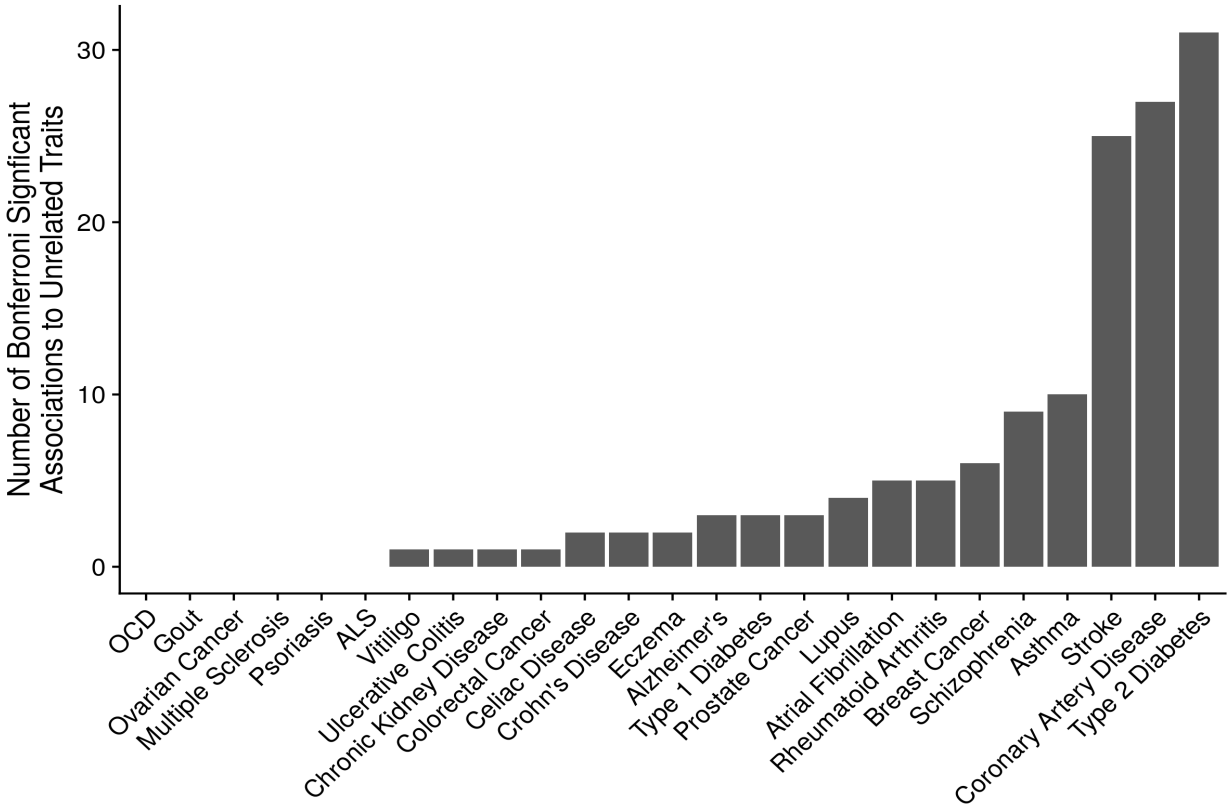

Figure 32. For each trait the best polygenic risk score was used to assess many unrelated traits within a logistic regression model that included age, sex the top 4 principal components, and the polygenic risk score. The number of polygenic risk score p-values in each of these many regressions, which were below the Bonferonni threshold, are indicated here for each trait.

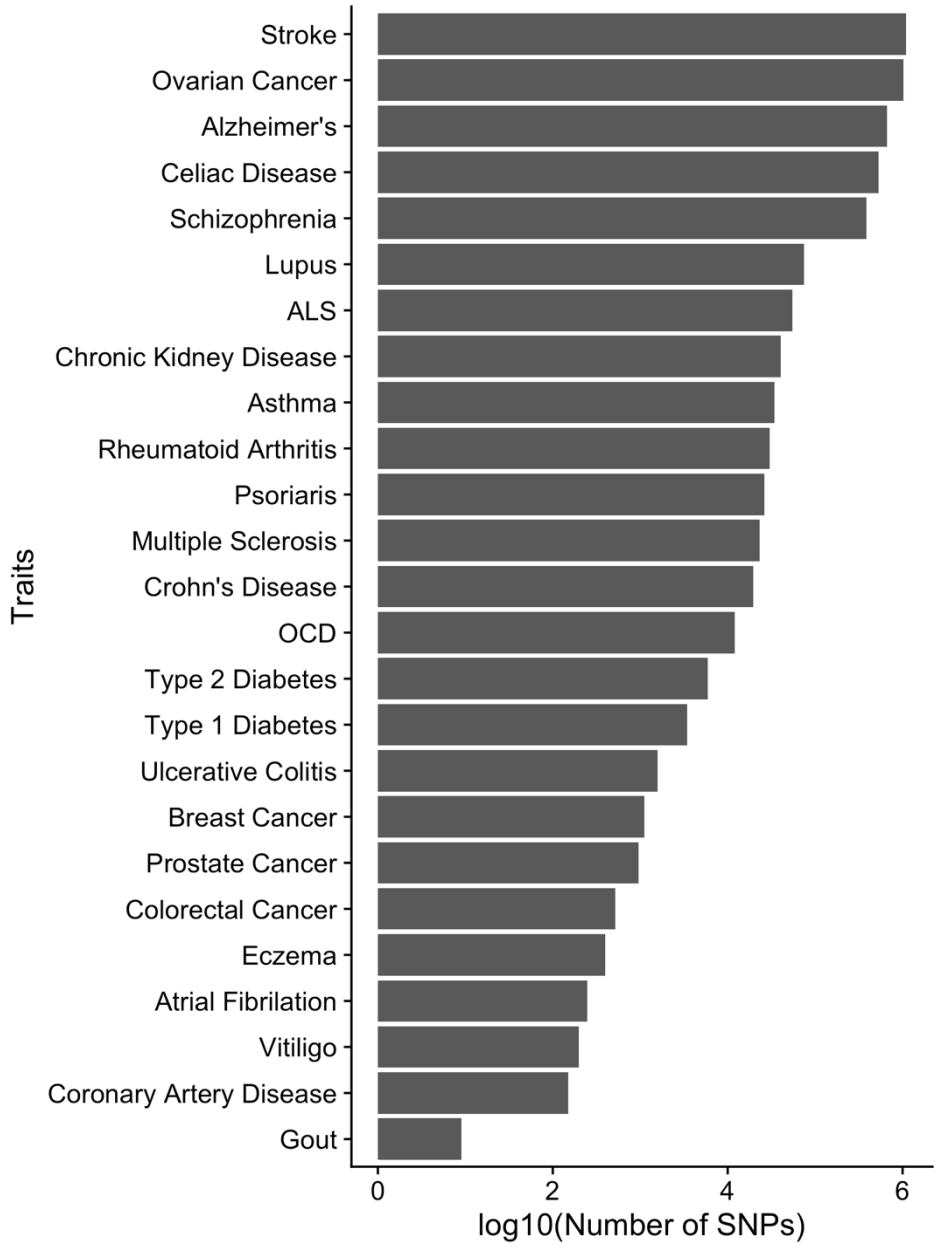

Figure 33. The number of variants within the set of data which produces the best polygenic risk score for each trait.

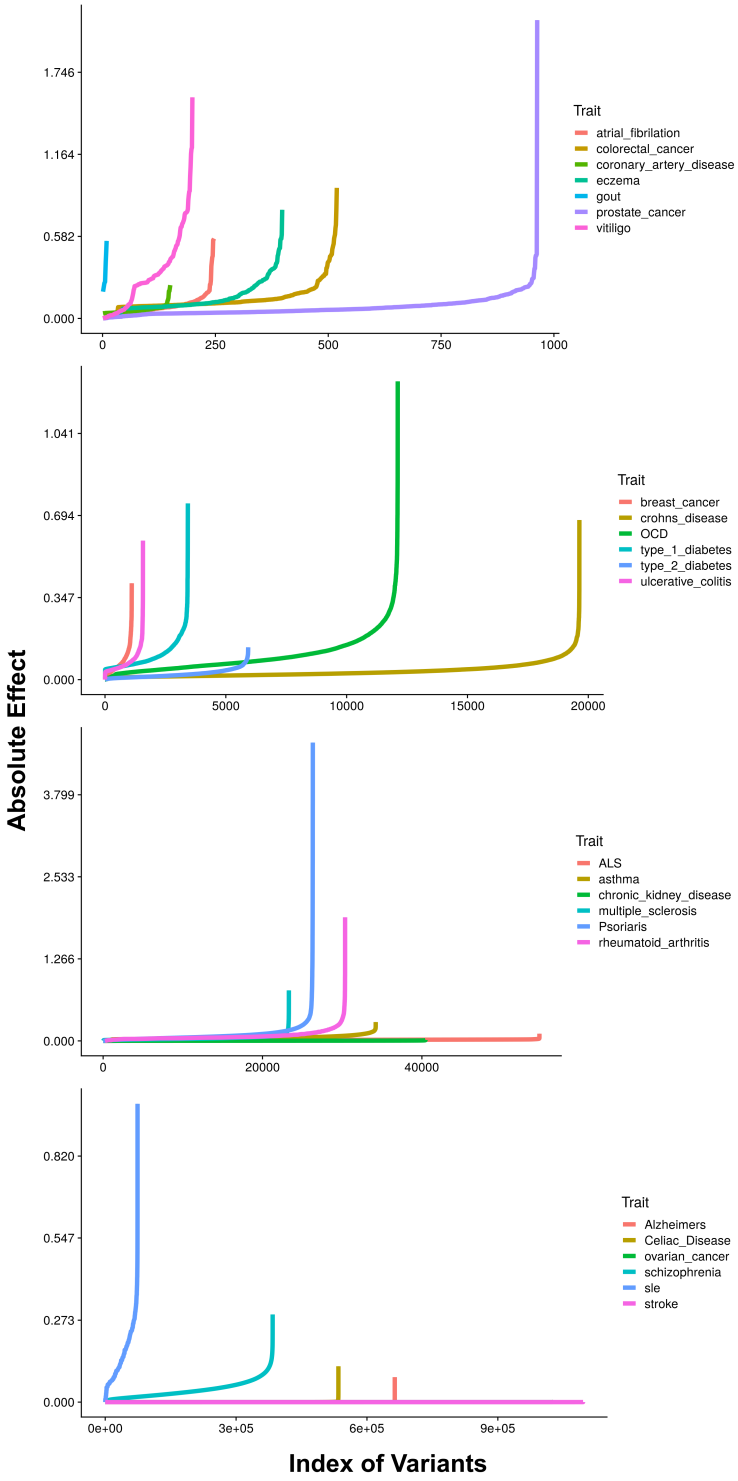

Figure 34. The effect of each variant, ordered, for each trait. The variants within the set of data which produces the best polygenic risk score are ordered according to their effect and plotted with equal distance between each variant.

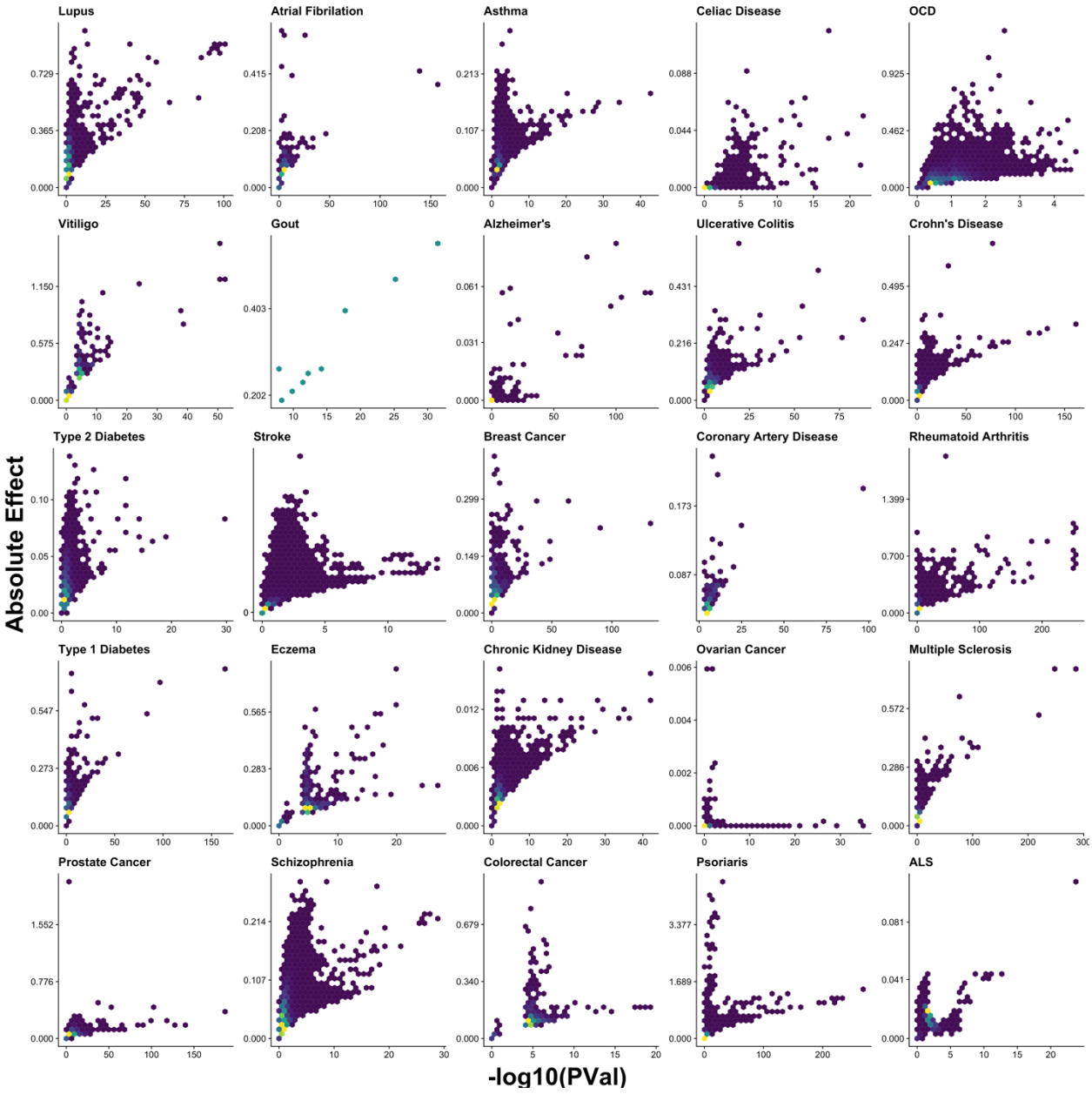

Figure 35. The effect and significance for all variants that produce the best polygenic risk score, for each trait. The density of variants that share the same approximate significance and effect are indicated with color, yellow being a higher density and dark blue being a lower density.

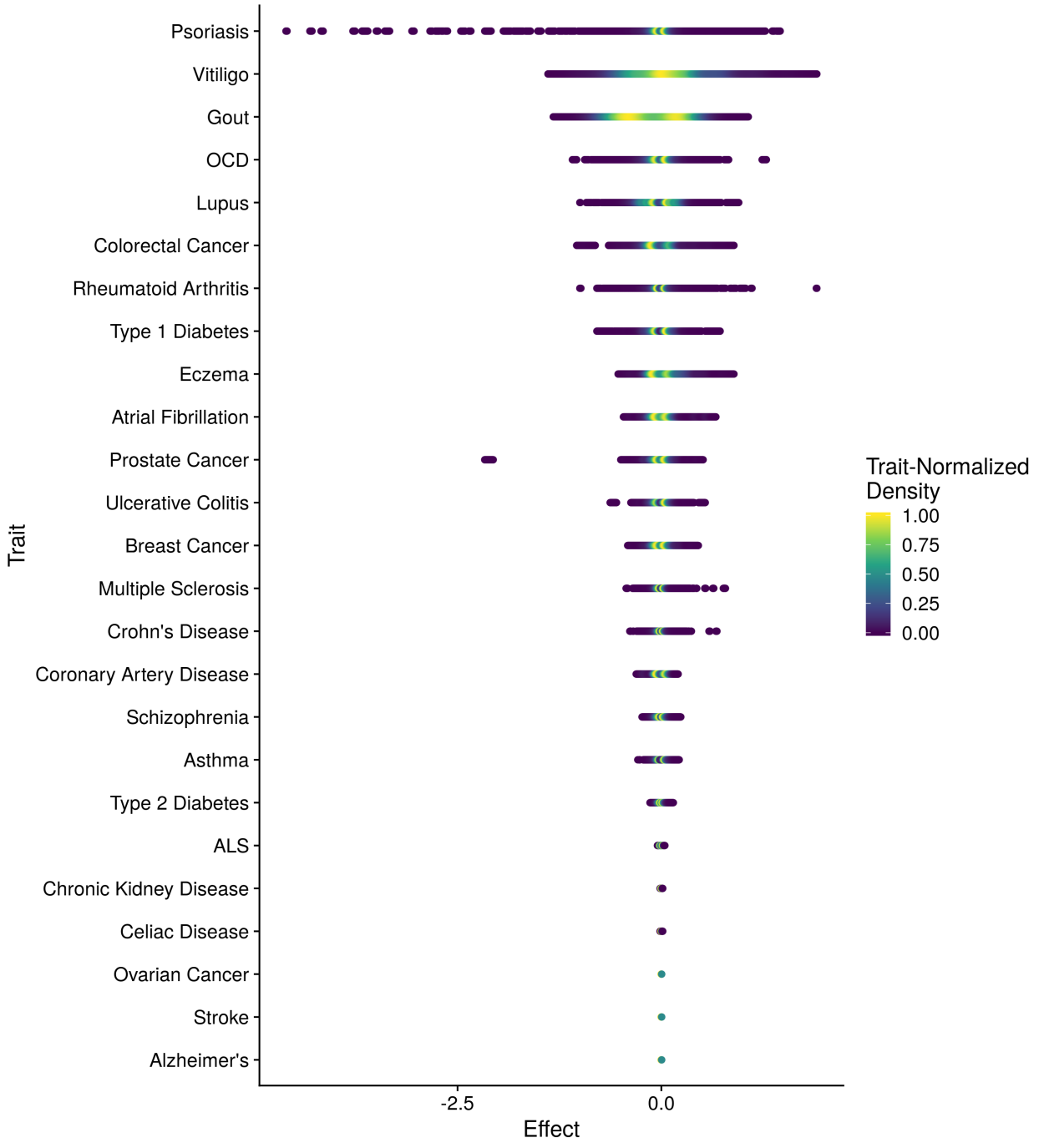

Figure 36. The effects of the variants the compose the best polygenic risk scores for each trait. The density of the effect, or the number of variants with a similar effect, is indicated with color. Bright yellow is the largest effect for the trait and dark blue is the smallest effect for the trait.

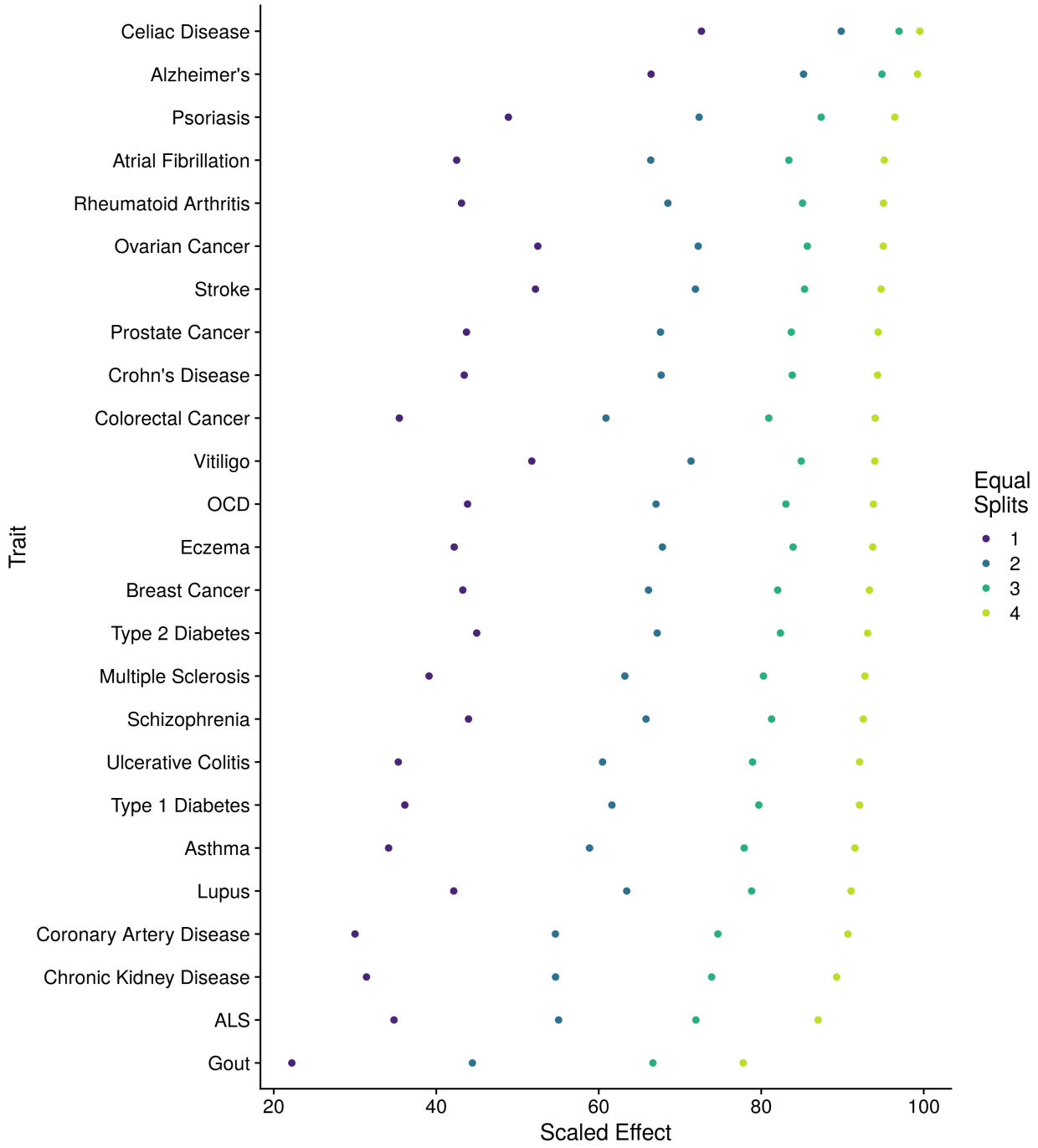

Figure 37. The values of scaled effects which equally split the total set of variants that define the best polygenic risk score into groups with equal number of variants. The scaled effect is the cumulative absolute value of all effect divided by the total sum of the absolute effects.

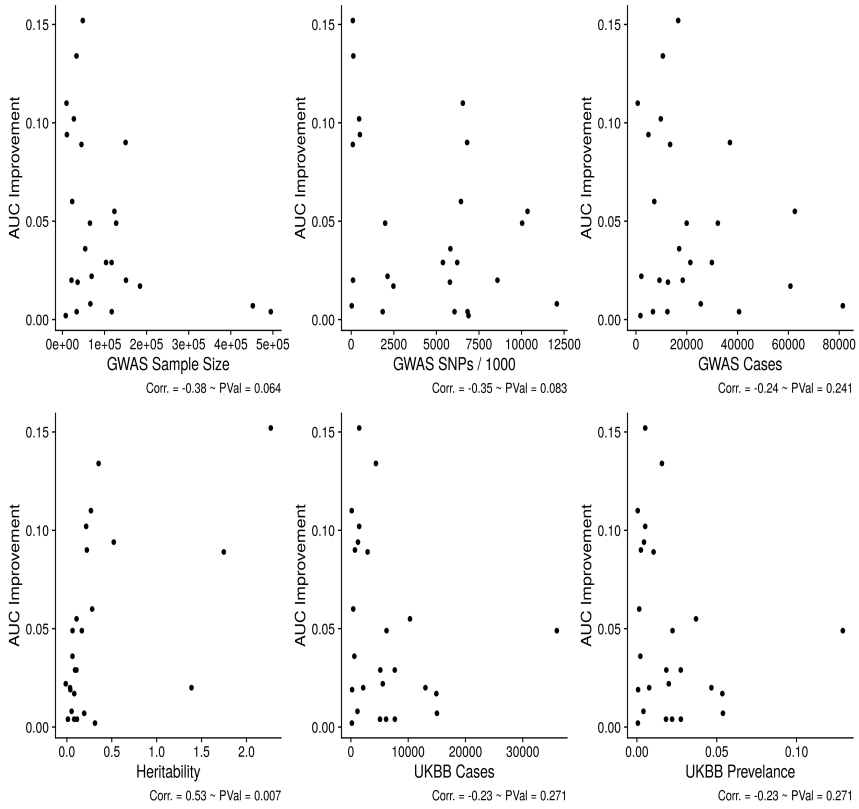

Figure 38. Relationships between the AUC improvement in the validation phase and various meta-statistics. The correlation is specifically the Pearson Correlation.

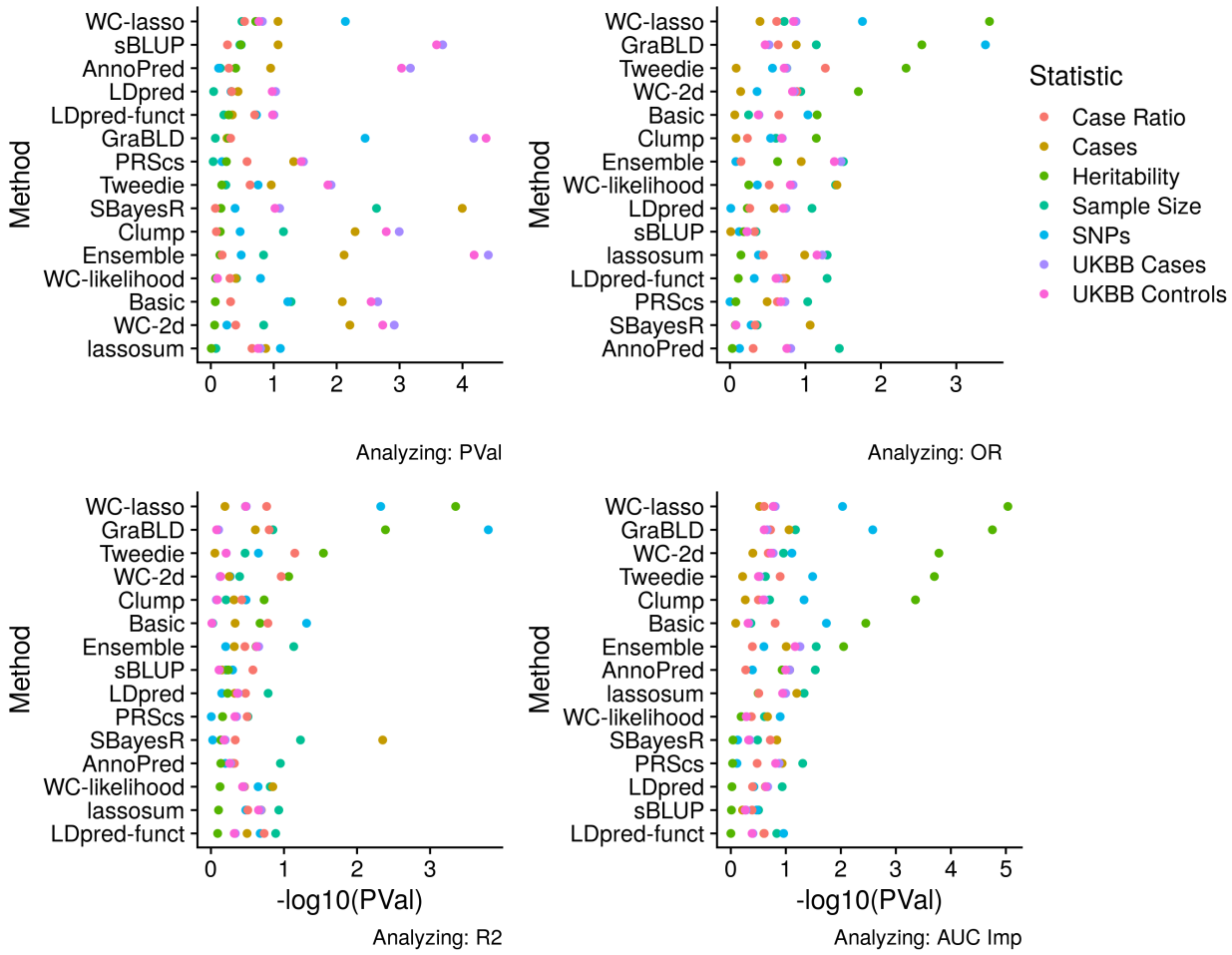

Figure 39. Relationship between the assessment statistics for all traits and various meta-statistics. The assessment statistics are generated from linear regression models between the polygenic risk score, as the independent variable, and the described meta-statistic, all generated in the training data-set.

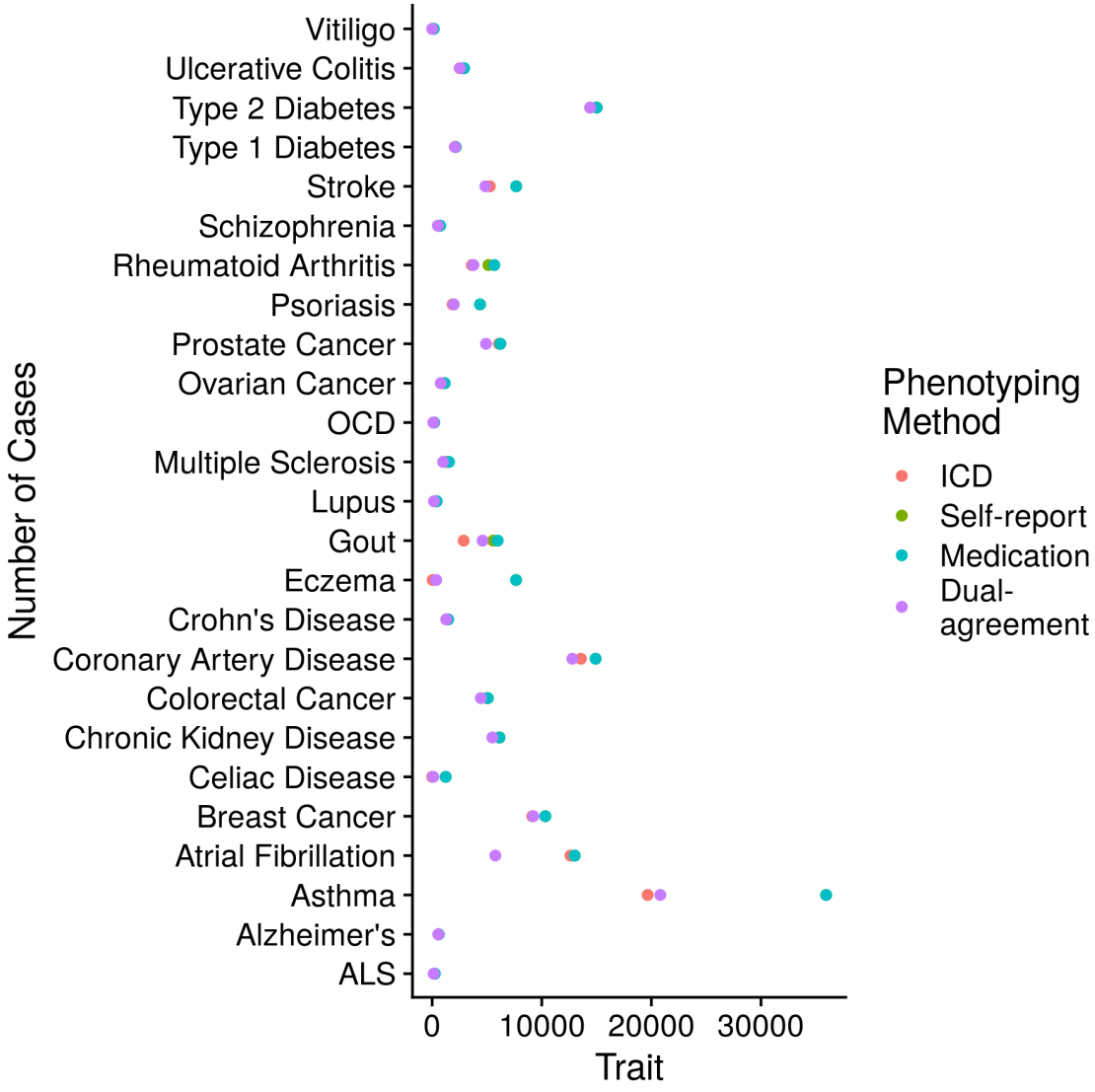

Figure 40. The number of cases for each trait according to each phenotyping methods in the withheld data-set.

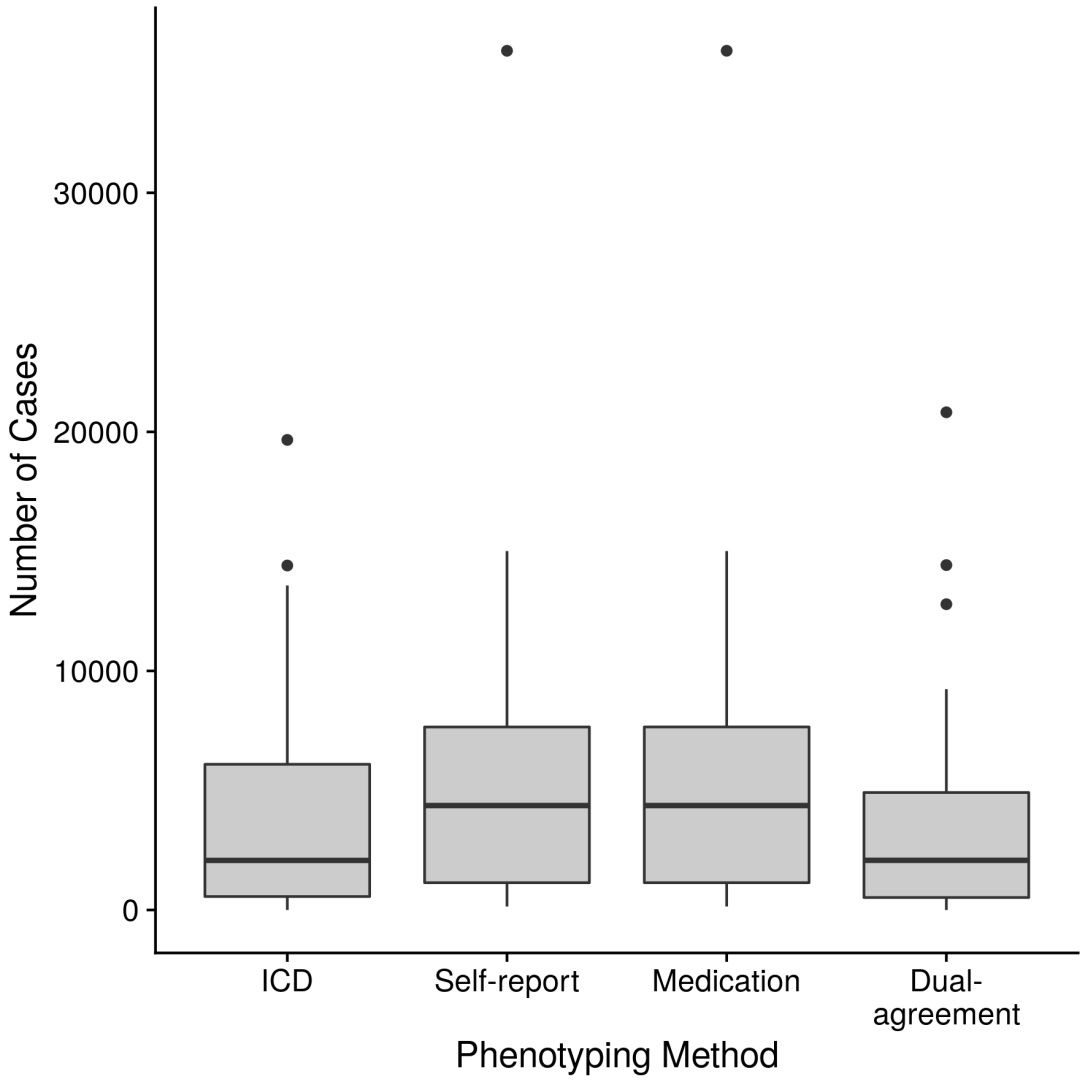

Figure 41. The number of cases for each phenotyping method in the withheld data-set.

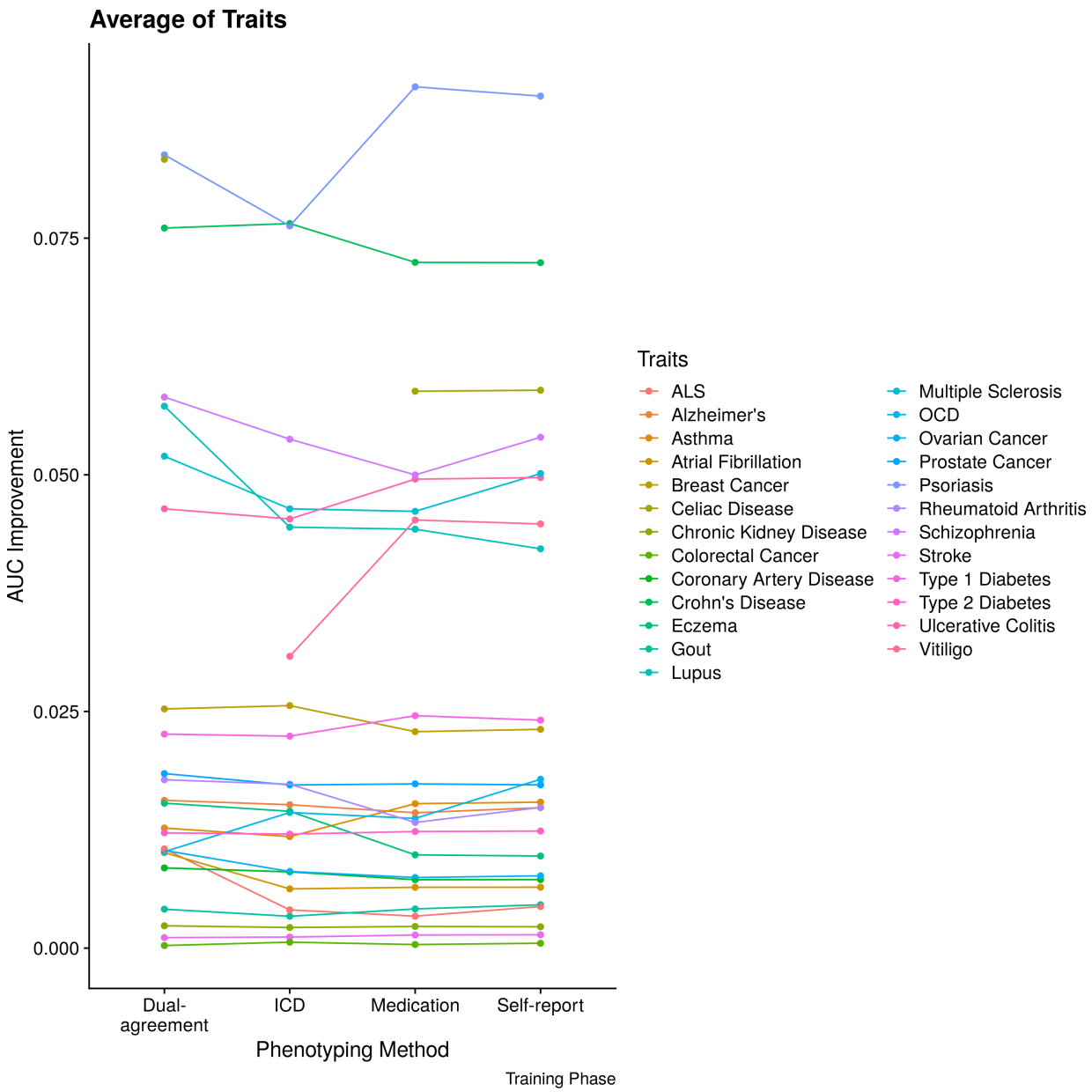

Figure 42. The AUC improvement for each trait and phenotyping method. The phenotyping method is indicated on the x-axis and the traits are color coded, and connected with a line. Each point is the best polygenic risk score for the trait, within the training phase. Missing values appear when the phenotyping method produces too few cases to analyze.

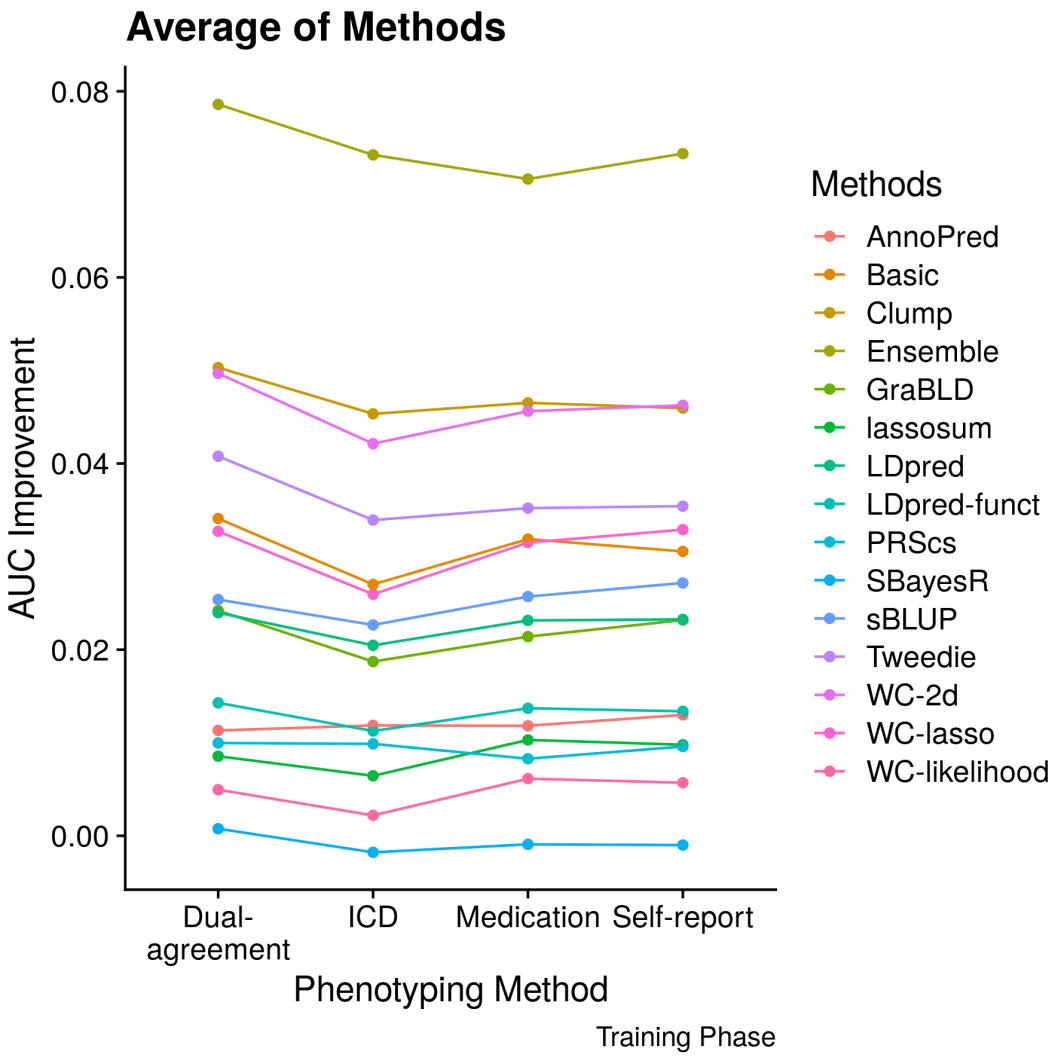

Figure 43. The AUC improvement for each method and phenotyping method. The phenotyping method is indicated on the x-axis and methods are color-coded and connected with a line. Each point is the average over all traits within the training phase.

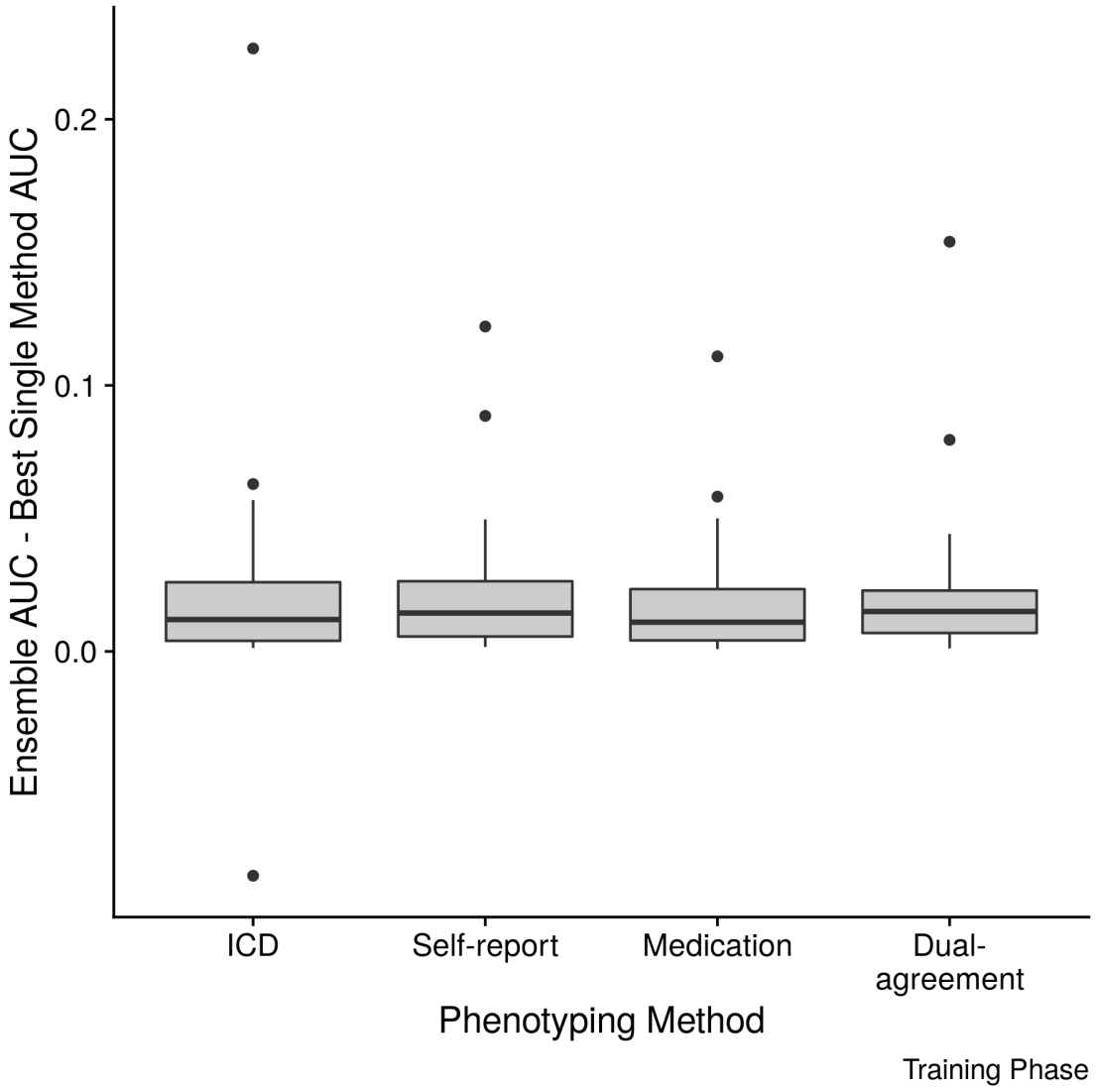

Figure 44. The difference between the AUC improvement of the ensemble and best non-ensemble method. The difference values for each trait are summarized within a boxplot for each phenotyping method.

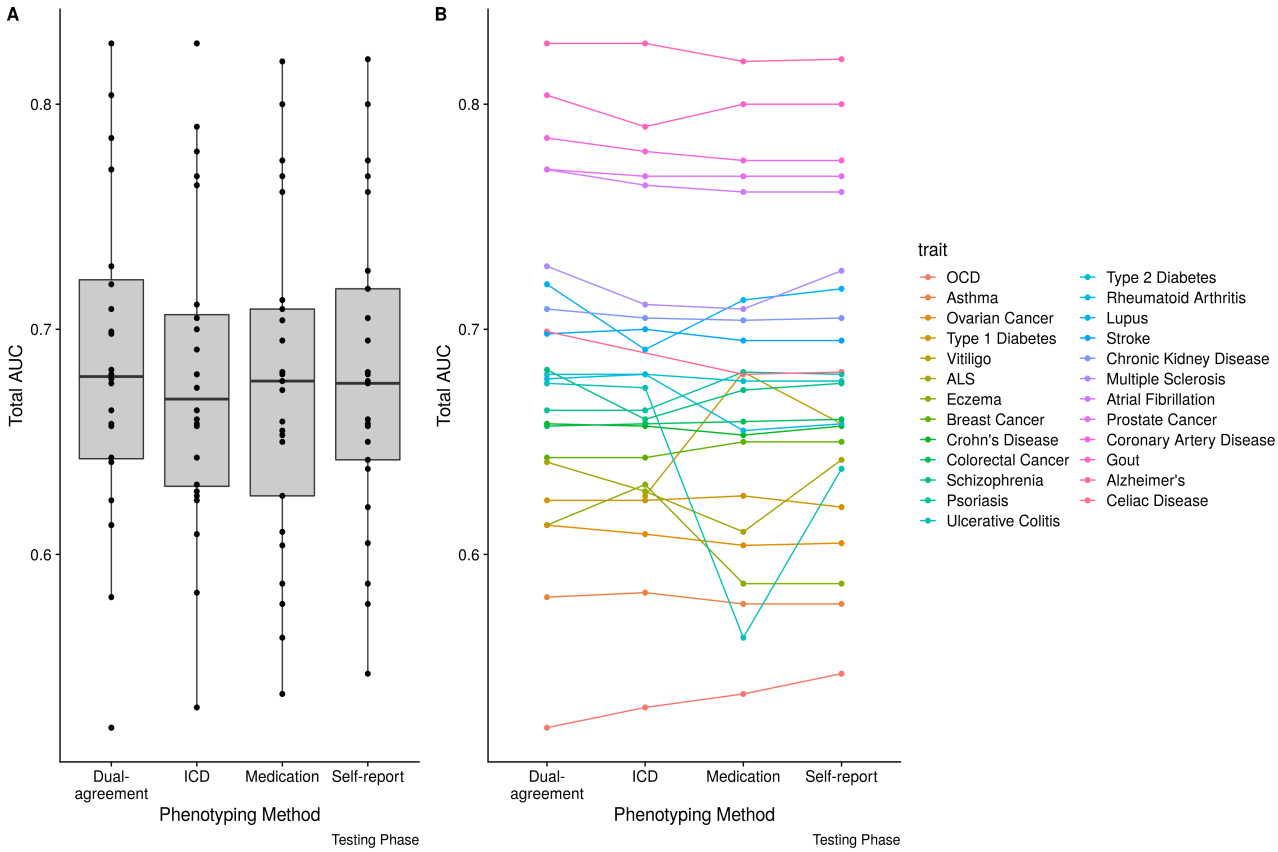

Figure 45. The distribution of AUC improvement for each trait in the withheld data-set, for each phenotyping method. In panel A an average improvement is shown through the boxplots, and in panel B a stratified improvement is shown for each trait.

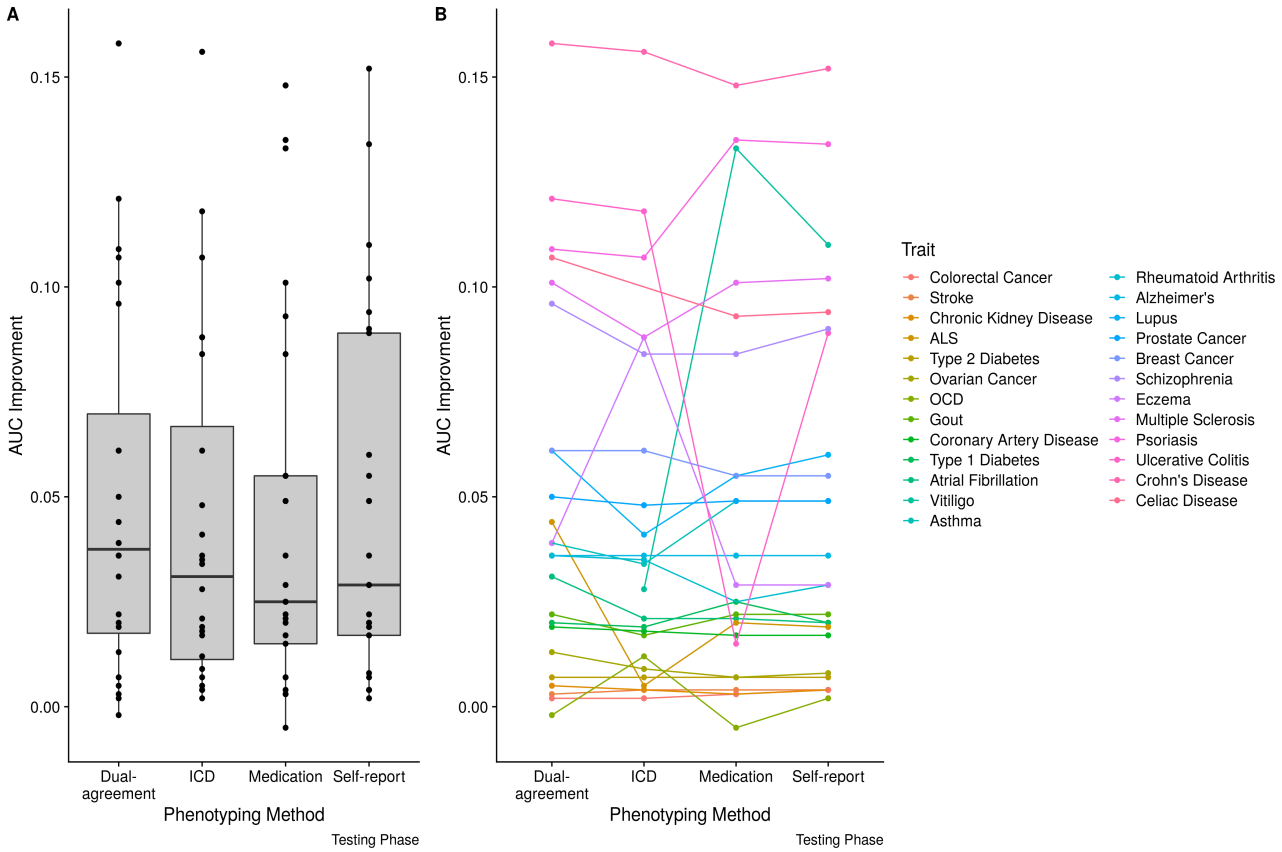

Figure 46. The distribution of AUC improvement for each trait in the withheld data-set, for each phenotyping method. In panel A an average improvement is shown through the boxplots, and in panel B a stratified improvement is shown for each trait.

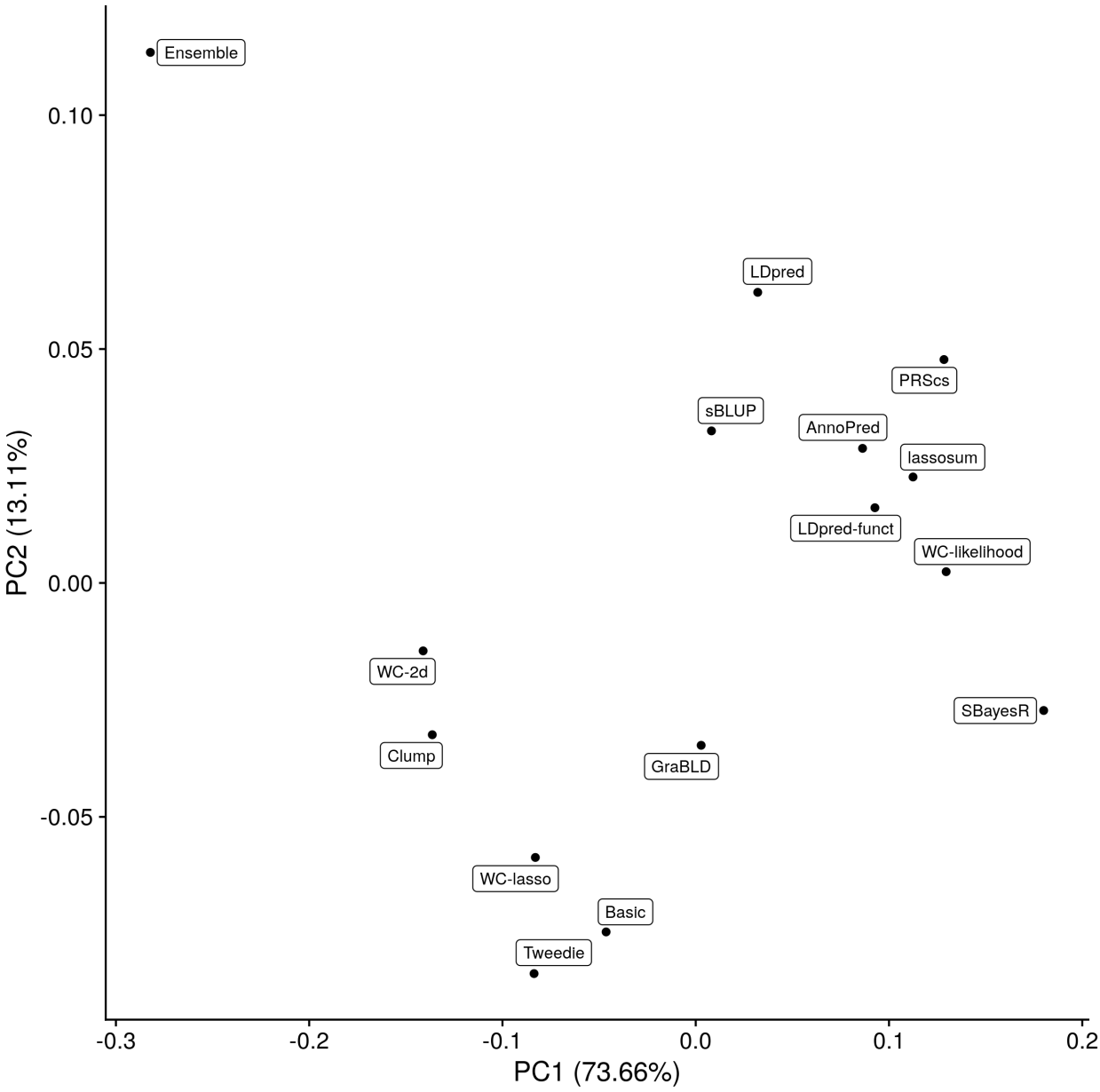

Figure 47. Principal components analysis of a matrix containing the ranking of each trait according to AUC improvement across all methods. Two clear clusters appear between the simple and complex models.

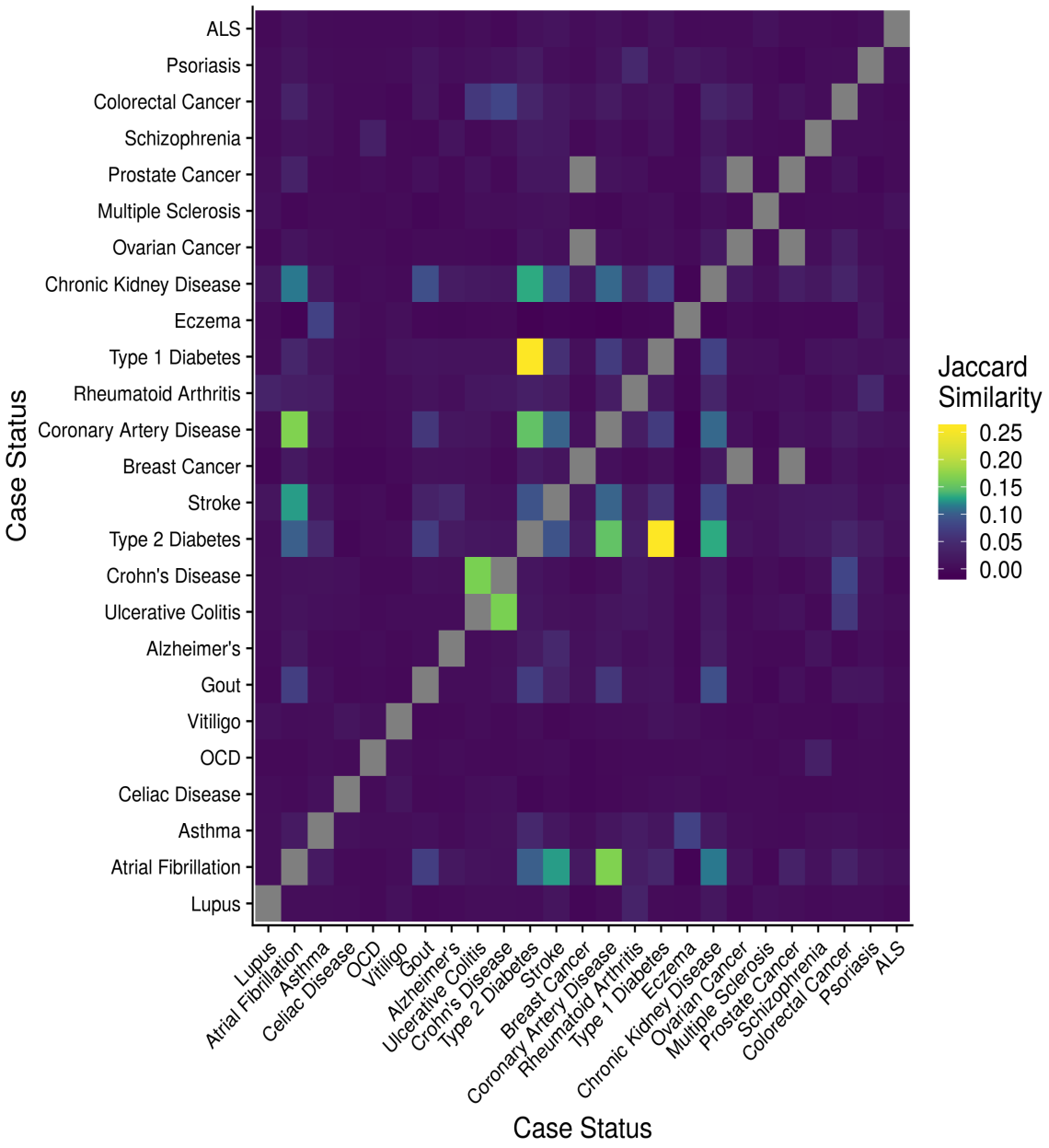

Figure 48. Pairwise Jaccard similarities between all of the 25 main traits analyzed. These similarities can help inform the purely genetic component of the heatmap in figure 3.

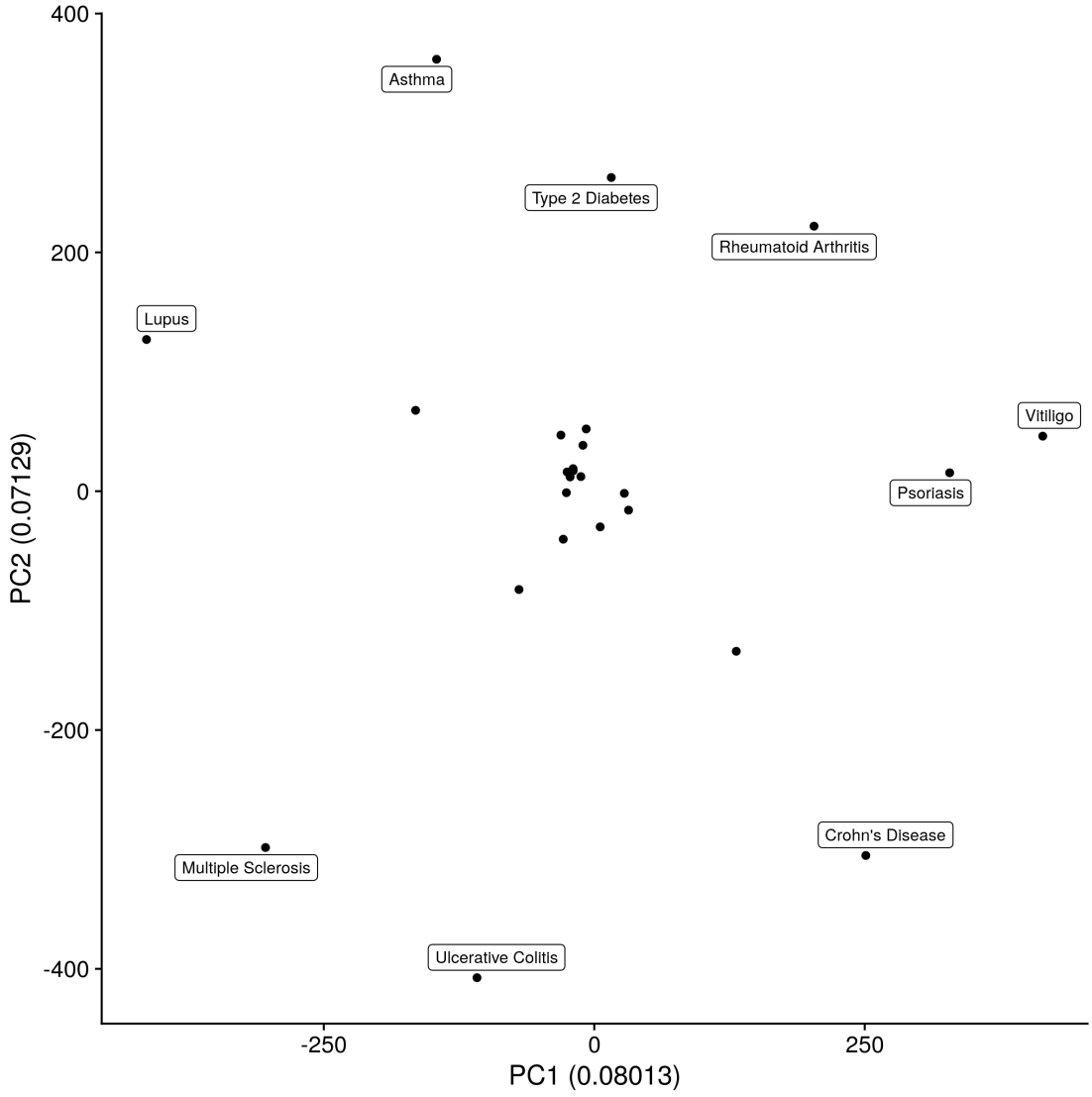

Figure 49. Principal component analysis of a matrix with all individuals examined against all traits, filled with the respective polygenic risk score. Most of the variation of this matrix is not captured within the first two principal components.