# netChoose

*Alex Gutteridge*

*19/03/2019*

This R Markdown document is made interactive using Shiny. Unlike the more traditional workflow of creating static reports, you can now create documents that allow your readers to change the assumptions underlying your analysis and see the results immediately.

To learn more, see Interactive Documents (http://rmarkdown.rstudio.com/authoring_shiny.html).

# Introduction

Multiple groups have shown through empirical analysis that the selection of drug targets using genetic evidence improves the liklihood of successful demonstration of clinical efficacy. However it is also the case that many successful drug targets are not identified by genetic evidence due to the lack of appropriate genetic instruments, pleiotropic effects or our power to detect effects and many genes identified by genetic evidence are not suitable as drug targets due to tractability and pleiotropy. Given this background it is common practice to attempt to infer indirect genetic associations by the use of so called proxy genes.

Proxy genes are typically used in one of two scenarios: Firstly a genetic study (e.g. genome wide association study (GWAS)) identifies a small number of genes linked to a disease of which a large proportion are non-tractable or otherwise not suitable as drug targets and the aim of using proxy genes is to infer other potential targets that may be considered to have genetic evidence for the disease in question. Secondly a pre-existing target hypothesis may be under consideration (e.g. the target of an existing pipeline asset that could be repositioned) and the aim of using proxy genes is to understand what diseases the existing target could be linked to through indirect genetic evidence.

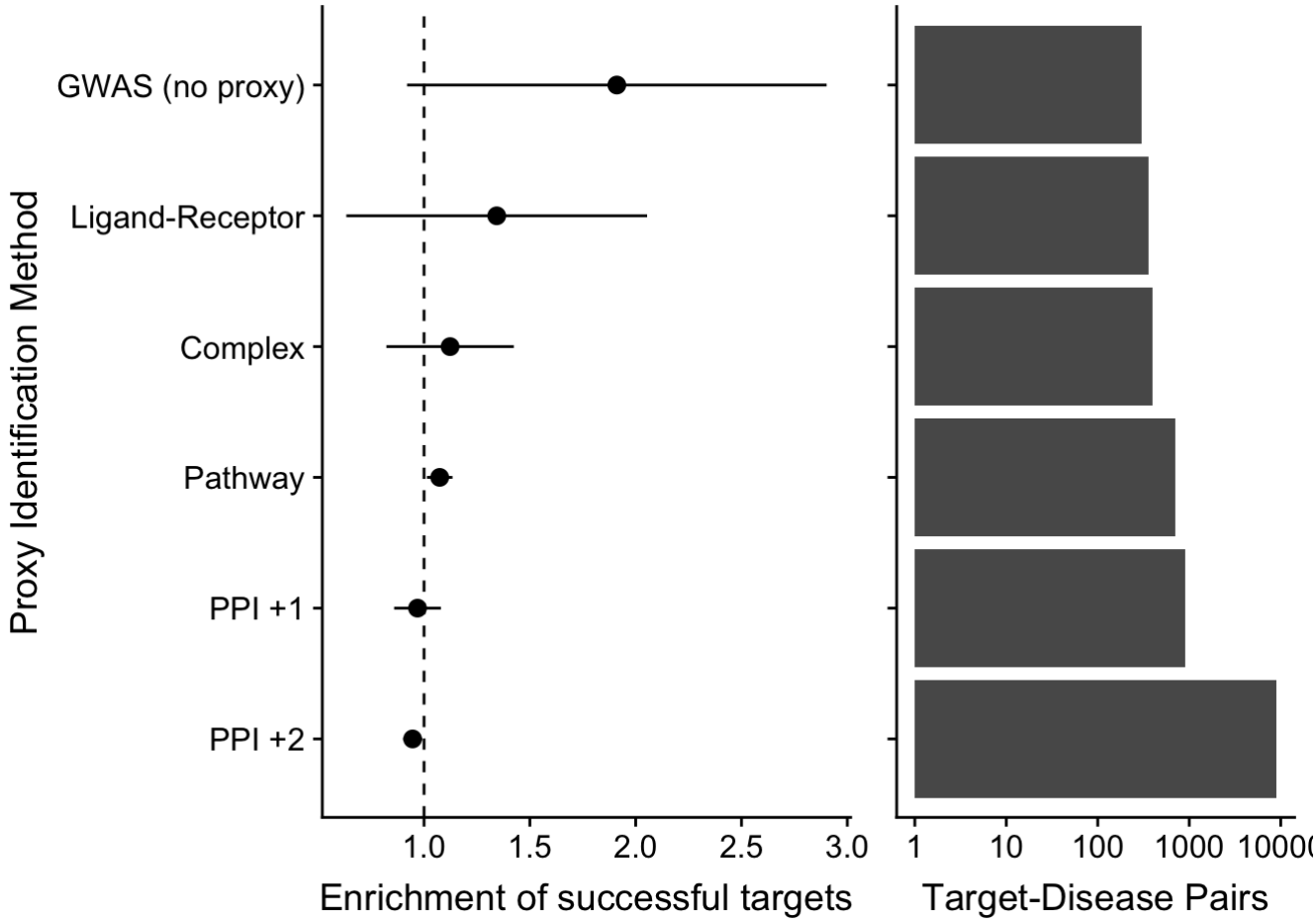
In either scenario the set of proxy genes is typically assembled using prior biological knowledge that comes from pre-existing databases of gene-gene or protein-protein interactions. The simplest example of such an analysis would be to take receptor-ligand pairs and where a ligand is genetically associated to a disease make the inference that the known cognate receptor is therefore indirectly associated to the same disease (or vice-versa). The same logic can be applied to stable complexes of proteins (heteromeric complexes for example), signalling pathways, first and second order neighbours in a protein-protein interaction network or indeed any other predefined collection of gene sets.

The figure below shows the enrichment of successful drug targets that is empirically observed when selecting targets on the basis of direct GWAS evidence alone from 77 traits with well powered GWAS and five methods of assembling proxy genes based on that GWAS evidence (hereafter described as the ‘core’ traits). We also show the final number of potential targets inferred by each method.

The five initial methods for inferring proxies used here are:

- Ligand-receptor pairs: In this algorithm if either member of a ligand-receptor pair is associated to a disease then the other member is as well.
- Complex: If any member of a known stable protein complex is associated to a disease then all other members are as well.
- Pathway: If any member of a known annotated signalling pathway is associated to a disease then all other members of the same pathway are as well.
- Protein-protein interactors (PPI) +1: For each disease associated gene every gene whose protein product is known to physically interact with the original gene’s protein product is associated to the same disease.

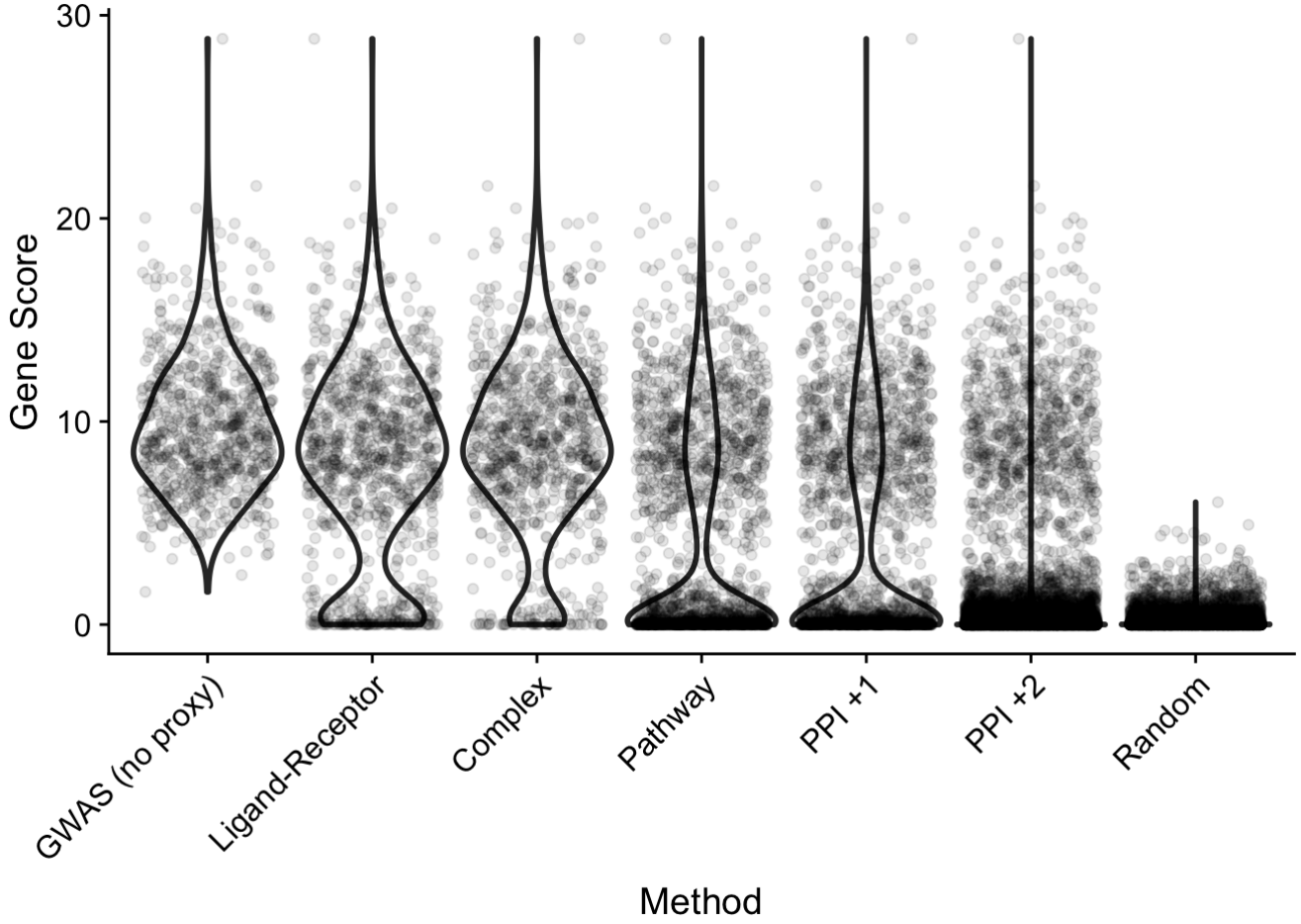
- Protein-protein interactors +2: The same approach as PPI +1 is taken but every protein interacting with a protein in the +1 set is also associated to the disease.



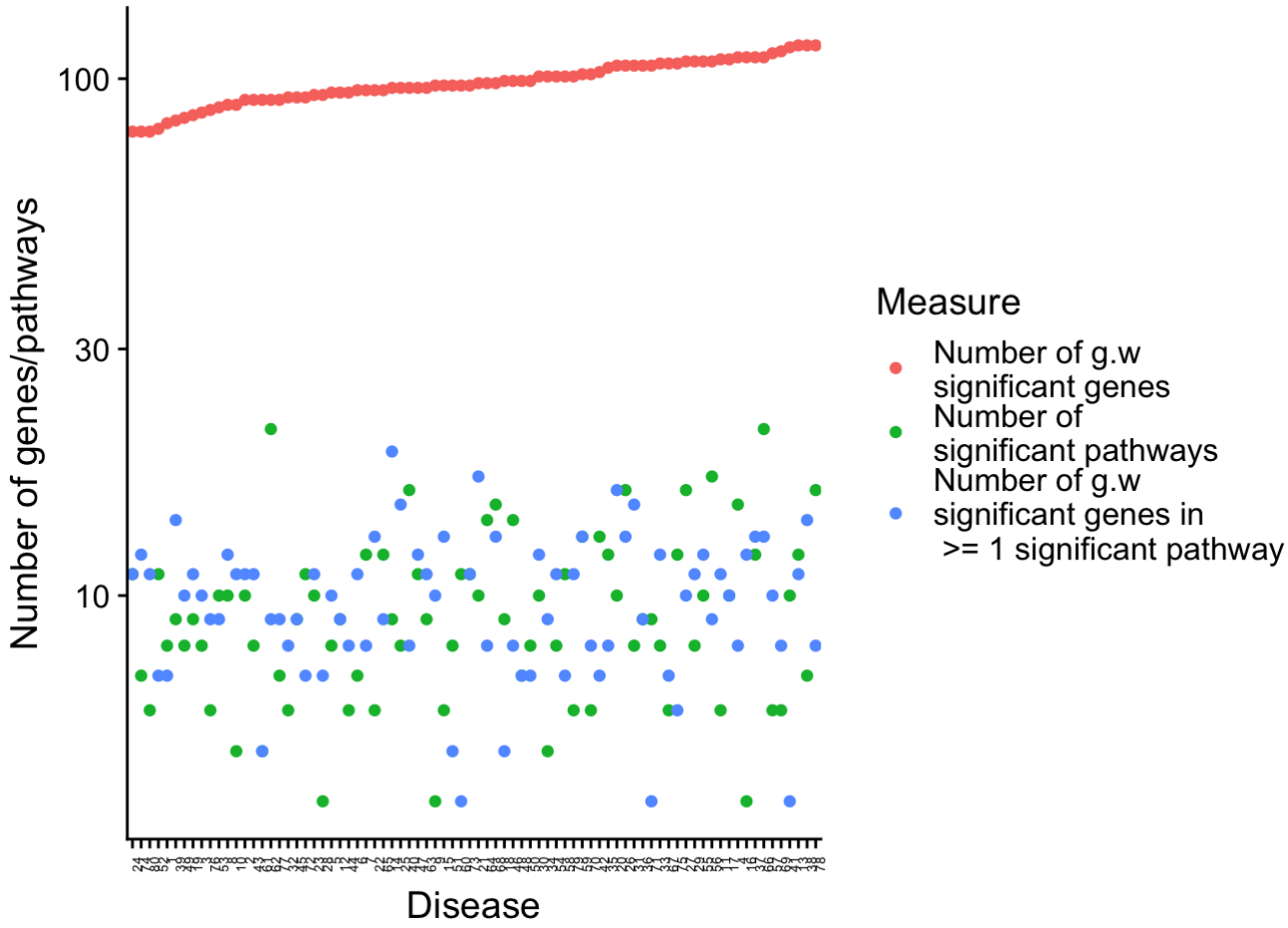

# Network and pathway methods for inferring proxy genetic evidence

The naive pathway and network methods used above result in a significantly reduced degree of enrichment for successful drug targets in the selected proxy genes compared to directly disease associated targets. For the PPI +2 proxy method in particular the algorithm is so promiscous that very large numbers of genes become associated with each disease. More sophisticated methods of using network and pathway information behave differently.

The implicit assumption behind network and pathway based methods is that there will be a clustering of genetic associations amongst genes that are functionally linked (i.e. that are involved in binding a certain ligand or are members of a certain signalling pathway) and that our prior knowledge of networks and pathways is sufficient to capture this clustering. The figure below shows the distribution of gene level scores (defined using the Pascal method) for each set of proxy genes across all the core traits compared to the directly associated genes and a random background set of 100 genes per trait. [NB: If scores vary greatly in dynamic range across the traits we could normalise within each trait such that the most significant gene is set to 1].



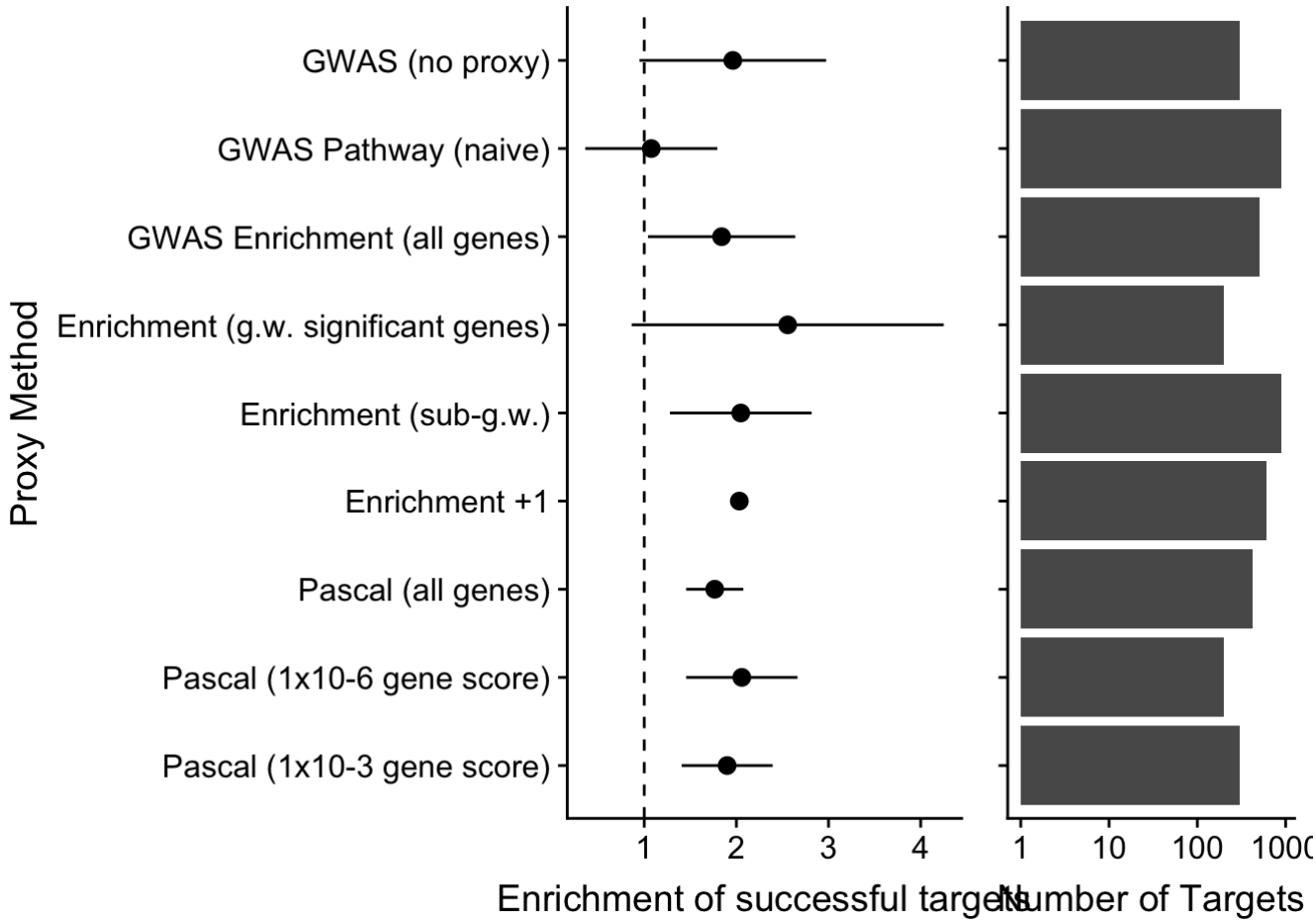

Next we look at pathway and hybrid pathway/network methods. When comparing the performance of different methods we take care to use the same underlying pathway and network data wherever possible. As a first step all of these methods define a set of pathways as enriched for genetic associations. For the GWAS enrichment method this is measured using a Fisher's Exact test and for the Pascal method this is measured using a Chi-Squared test based on the full genome wide summary statistics. The information on the level of enrichment of targets within pathways is itself useful beyond simple selection of targets as pathway enrichment results provide a basis for both grouping GWAS associations by molecular function and forming deeper mechanistic hypotheses. The number of gene and pathway level hits across the 77 core traits is shown below



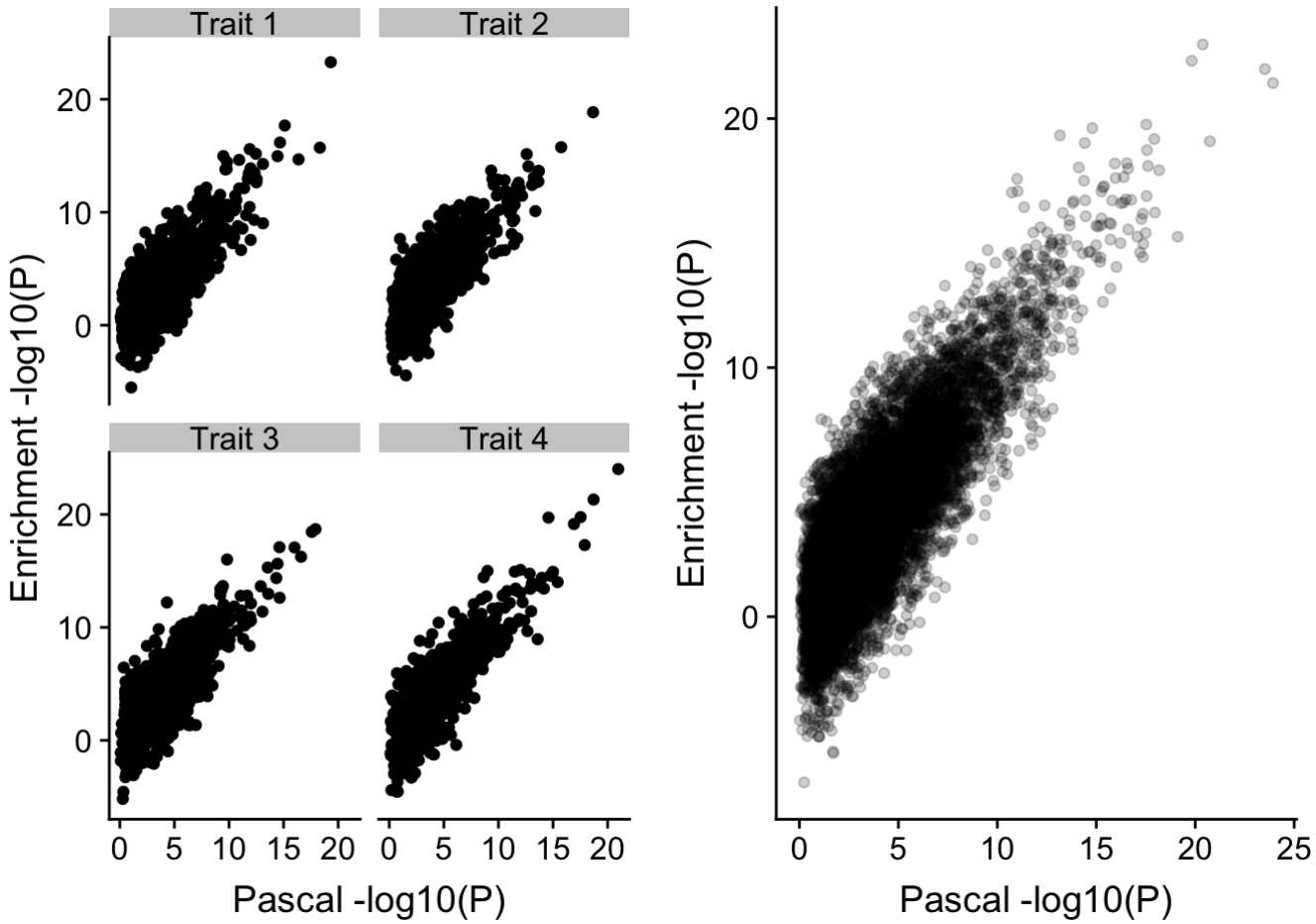

To quantitatively assess the ability of these methods to prioritise drug targets we use several different secondary analyses that select individual genes within the pathways as potential targets. The secondary methods used are:

- GWAS Enrichment (all genes): This method assigns all genes within a pathway detected as enriched by Fisher's Exact test (but only those enriched pathways) as potential targets.
- GWAS Enrichment (genome wide significant genes): This method only assigns those genes that are themselves associated to the disease in question at a genome wide level of significance. Note that this is the only method tested that results in a smaller number of potential targets for a given GWAS than the original input gene list.
- GWAS Enrichment (sub-genome wide): This method assigns genes that are in enriched pathways and have a gene level association greater than […].
- GWAS Enrichment +1: This method is the same as the genome wide significant method but also includes genes that are immediate neighbours of the genome wide significant genes that are also members of the enriched pathway. This method is therefore a hybrid pathway/network approach.
- Pascal (all genes, 1x10-6, 1x10-3): These methods are analagous to the above methods but use Pascal to define enriched pathways using the full genome-wide summary statistics.

For example… [Can we find one example where the number of pathways is significantly lower than the number of individual associations, but also where a significant fraction of the associations are captured within a pathway?].




The two pathway enrichment methods we have tested are a Fisher's Exact based method and a ChiSq based method (Pascal). A comparison of the performance of the two methods for the detection of enriched pathways is shown below for four sample traits and for all traits combined.



# Network propagation and subnetwork detection methods

We next consider a further class of methods that only model gene-gene associations as networks (not as genesets or pathways). These methods themselves fall into two classes: network propagation and subnetwork detection. In the first class genetic association scores are mapped onto genes within a network and subsequerntly propagated through the network typically by a modelled diffusion or random walk process to create a new set of scores. In the second class of methods a set of subnetworks enriched for high scoring nodes are retrieved from the network. Both methods tested here (HotNet2 and Hierarchical HotNet) perform an implicit score propagation step first.

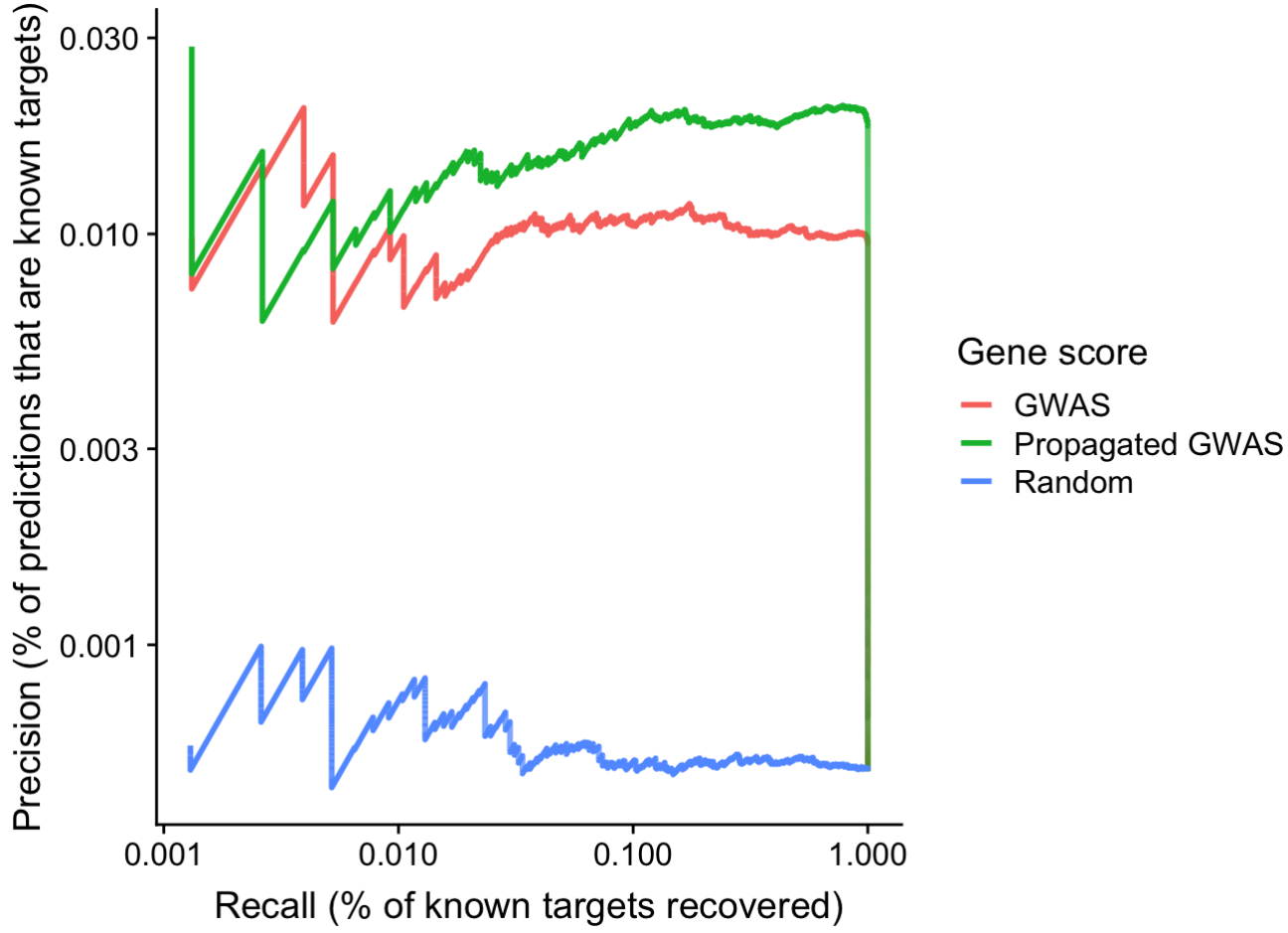
The performance of these methods are shown below. To assess the network propagation methods we use three arbitrary cutoffs of the ranked scores (top 50, top 100 and top 500 genes) to define the set of drug target hypotheses for each trait. Note that in this conception the concept of proxy genes is no longer strictly applicable.

[..Anotehr enrichment Figure Here..]

These methods are largely agnostic of the type of network used. Below we show the performance of a selection of these methods across different input networks, including different forms a single network type (PPI) and different network types. We know from previous studies [cite] that networks with higher connectivity, but likely higher noise, perform better when given the task to retrieve known drug targets. Since all these methods rely on the assumption that genetic associations cluster in the network this is a useful assumption to test. In the figure below we show the distribution of modularity scores for X traits (just core or all?) using each specific network as input and those genes that are g.w. significant (or Pascal gene score

[..Networks performance comparison..]

Because the network propagation methods natrually provide a ranked list for genes it is also possible to assess their performance (recall of known targets) as a function of the score cutoff (precision). As a baseline we compare the propagated scores to a binary score based on whether the gene is or isn't associated to a trait at a genome wide significant level and a non-propagated continous gene score derived from Pascal.




```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 pr_auc  binary        0.0102
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 pr_auc  binary        0.0191
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 pr_auc  binary      0.000520
```

# Network methods as a framework for integrating Mendelian and complex, common disease information

All the methods and performance comparisons above.