# Tutorial on using genphen

Simo Kitanovski, Daniel Hoffmann,
Bioinformatics and Computational Biophysics,
University of Duisburg-Essen, Essen, Germany

April 30, 2018

## Contents



This tutorial gives you some of the technical background underlying genphen that should enable you to understand and use this tool.

## 1 genphen quantifies genotype-phenotype associations

Genome wide association studies (GWAS) have become an important tool to understand the association between genotypes and phenotypes. With GWAS we try to answer questions such as "what are the genotypes in the human genome which predispose to a disease?" or "what are the genotypes in certain strains of mice which confer resistance against a specific virus?". The advances in high- throughput sequencing technology have provided massive genetic data and thus potentially countless applications of genotype-phenotype association studies. The genotype can be a set of single

nucleotide polymorphisms (SNPs) or a set of single amino acid polymorphisms (SAAPs) identified in a group of individuals, whereas the phenotype can be any continuous or discrete individual quantity or characteristics.

To conduct GWAS, frequentist statistical methods are typically used, relying on simple and often inadequate methods to capture the complex and potentially non- linear genotype-phenotype association. Moreover, these methods often use P- values to quantify the strength of association, bringing with them a set of disadvantages, some of which include poor interpretation, difficulty to compare between different studies, as well as massive multiple hypothesis problems.

With `genphen` we provide a hybrid method which reaps the benefits of sophisticated statistical learning approaches such as random forest (RF) and support vector machine (SVM) to capture complex genotype-phenotype associations, on the one hand, and Bayesian inference on the other hand, to accurately quantify the strength of association using models consistent with the data. The results of `genphen` are multiple association scores for each genotype. Visualizing these scores together can present the researcher a meaningful guide to selecting the most promising association.

Furthermore, `genphen` provides a set of procedures including a test for phylogenetic bias (used to discover biases in the data due to the population structure), procedure for data reduction (used for removing non-informative genotypes and thereby simplifying the GWAS), data augmentation (used to augment small sample datasets) and methods for gene prioritization based on network diffusion algorithms using functional network data.

# 2 Methods

## 2.1 Input

Two data types are necessary to perform a genetic association study:

- genotype data (e.g. set of 1,000 SNPs found along the aligned genomes of 10 individuals)
- phenotype data (experimental measurement made for each individual such as body mass index, immune response, survival, case-control, etc.)

More generally, we can think of the genotype data as a character matrix with dimensions $N \times M$, whereby the $M$ columns represent different SNPs or SAAPs, and the $N$ rows represent different individuals or sequences for which we have some measured phenotype. Therefore, we can think of the

phenotype as a numerical vector of length $N$, where each phenotype corresponds to a particular individual. Moreover, the phenotype can be of specific type (e.g. quantitative, dichotomous, etc.), imposing requirements on the type of statistical test that can be be used for the association analysis.

## 2.2 Association Scores

Between each genotype (SNP/SAAP) and phenotype, `genphen` computes several measures of association, each of which is explained in the following paragraphs.

**Classification accuracy($CA$)** $CA$ measures the degree of accuracy with which one can classify (predict) the alleles of an SNP from the phenotype measurements. If there exists a strong association between a particular SNP and the phenotype, one should be able to build a statistical model which accurately classifies the two alleles of that SNP solely from the phenotype data ($CA \approx 1$). Otherwise, the classification accuracy of statistical model should be approximately similar to that of simple guessing ($CA \approx 0.5$).

To estimate a robust $CA$ estimate, `genphen` uses cross-validation (CV), obtaining a distribution of possible $CAs$. During the CV procedure a subset (e.g. 70%) of the genotype-phenotype data is selected at random for training the classifier, followed by testing based on the remaining data. The following confusion matrix represents the result of one CV step:

| | | **Real** | |
|---|---|---|---|
| | | $allele_1$ | $allele_2$ |
| **Predicted** | $allele_1$ | a | b |
| | $allele_2$ | c | d |

Table 1: Confusion matrix resulting from a classification analysis

The CA of the cross-validation step $i$ is then estimated as:

$$CA_i = \frac{a+d}{a+b+c+d}$$

The final $CA$ after 1000 CV steps is then estimated as:

$$CA = \frac{1}{1000}\sum_{i=1}^{1000} CA_i$$

In addition to estimating $CA$, the distribution of $CAs$ enables us to also compute the 95% highest density interval (95% HDI) of CA. The mutations with $CA \approx 1$, with narrow HDI have the strongest associations.

The metric $CA$ has the following advantages:

- simple to estimate
- simple to interpret (bound between 0 and 1)
- simple to compare across studies
- one can have high CA despite small effects
- detects strong non-linear associations as well

**Cohen's $\kappa$ statistic** There is one pitfall where the $CA$ estimate can be truly misleading, and this is the case when the analyzed SNP is composed of unevenly represented genetic states (alleles). For instance, the allele A of a given SNP is found in 90% of the individuals, while the other allele T in only 10%. Such an uneven composition of the alleles can lead to misleading results, i.e. even without learning the algorithm can produce a high $CA \approx 0.9$ simply by constantly predicting only the dominant label. The Cohen's $\kappa$ statistics can be used to estimate how much better the observed $CA$ is, compared to the classification accuracy expected by chance ($CA_{exp}$). To compute the $\kappa$ statistics, the confusion matrix shown before in Table 1 is used:

$$\kappa = \frac{CA - CA_{exp}}{1 - CA_{exp}}$$

$$CA_{exp} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} + \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$$

The $\kappa$ statistics is a quality metric, which is to be used together with $CA$. Cohen defines the following meaningful $\kappa$ intervals: [$\kappa$<0]: "no agreement", [0.0-0.2]: "slight agreement", [0.2-0.4]: "fair agreement" , [0.4-0.6]: "moderate agreement", [0.6-0.8]: "substantial agreement" and [0.8-1.0]: "almost perfect agreement". Similarly to the estimation of $CA$, the final Cohen's $\kappa$ is also estimated by averaging the individual $\kappa$ scores computed for each step of the CV. Here too, 95% HDIs are estimated.

**Cohen's $d$ effect size** Given data from two groups (two allele groups in a SNP), we ask the question: How much is one group different from the other with respect to the phenotype observed in each group? In the case of quantitative phenotypes we answer this question by computing the

Cohen's $d$ for each genotype-phenotype pair. Cohen's $d$ estimates which are significantly greater or smaller than 0 indicate that there is a large difference in the phenotype between the two genetic states of the specific genotype. Cohen (1992) defines level which define the magnitude of the effects as: $|d|<0.2$"negligible", $|d|<0.5$ "small", $|d|<0.8$ "medium", otherwise "large". The Cohen's $d$ is computed as follows:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{({\sigma_1}^2 + {\sigma_2}^2)/2}}$$

where $\mu_1$, $\mu_2$ and $\sigma_1$, $\sigma_2$ represent the mean and the standard deviations of the phenotypes in the two genetic states of the genotype.

`genphen` uses the following Bayesian inference models designed in STAN [1], to estimate each of the parameters in the Cohen's $d$ equation from the data:

$$\begin{aligned} Y_{ijk} &\sim T(\nu_i, \mu_{ijk}, \sigma_{ijk}) \\ \mu_{jk} &\sim N(\hat{\mu}, \hat{\sigma} \cdot 100) \\ \sigma_{jk} &\sim U(\hat{\sigma}/100, \hat{\sigma} \cdot 100) \\ \nu_j &\sim Gamma(2.0, 0.1) \end{aligned}$$

where $i$, $j$ and $k$ index phenotype observation $i$ at site $j$ and genotype $k$, respectively in the phenotype vector $Y$; $\mu$, $\sigma$ and $\nu$ are the mean, standard deviation and degrees of freedom parameters of the T-distribution which is used to model the phenotype observed in each allele; $\hat{\sigma}$ and $\hat{\mu}$ are the empirically estimated mean and standard deviations of the phenotype in each allele which are used to setup the broad priors for $\mu$ and $\sigma$.

For each parameter we estimate a complete posterior distribution with MCMC sampling implemented in `rstan`. Therefore, by plugging in the entire posterior distributions of the parameters into the Cohen's $d$ equation, we estimate a complete posterior distribution for $d$ as well. This also allows us to compute the corresponding 95% HDI of $d$. A complete description of the hierarchical model is provided in Krushke, 2013) [2]. For SNPs with more than two groups (e.g. 3 alleles), the Cohen's $d$ estimate is computed between each pair of alleles.

This approach, although computationally more challenging than a simple t-test, has a few advantages related to the fact that the phenotype is modeled in a more consistent way, i.e. the distribution of the phenotype in each group is described as a T-distribution with a mean ($\mu$), standard deviation ($\sigma$) and shape parameter ($\nu$).

[1] Stan Development Team. 2017. Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. http://mc-stan.org

[2] Kruschke, John K. "Bayesian estimation supersedes the t test."Journal of Experimental Psychology: General 142.2 (2013): 573

- complete information about the credible parameter values is obtained
- handling of outliers achieved by describing the data with heavy-tailed distributions instead of normal distribution → assumption of normality is alleviated
- the T-distribution which describes the phenotype in each group has its own variance parameter → assumption of homoscedasticity is alleviated.
- in addition to computing the differences between the central tendencies of the groups, we can use the standard deviation parameters to compute differences between their variabilities

**Absolute effect size ($a$)** For dichotomous phenotypes, `genphen` simply computes the absolute difference (also known as contrast $a$) between the phenotypes of any two genetic states of each polymorphism as follows:

$$a = p_1 - p_2$$

where $p_1$, $p_2$ represent the proportion of "successes" in a given number of observed trials in the two genetic states of the given SNP/SAAP. Similar to the case of having a continuous phenotype, `genphen` uses a Bayesian inference models designed in STAN to estimate each of the parameters in the equation above from the data:

$$Y_{ijk} \sim Bern(p_{jk})$$
$$p_{jk} \sim Beta(1/2, 1/2)$$

where $i$, $j$ and $k$ index phenotype observation $i$ at site $j$ and genotype $k$, respectively in the phenotype vector $Y$; $p$, is the probability parameter of the Bernoulli distribution which is used to model the dichotomous phenotype data as a set of sucesses in a set of trials observed each allele; The prior of $p$ is a Beta distribution whose two parameters are fixed at 0.5 (Jeffrey's prior).

Plugging in the estimated posterior distributions of $p$ into the equation above allows us to compute the mean $a$ point estimate and its 95% HDI.

**Bhattacharyya coefficient ($BC$)** We further use the infered parameters needed to compute the effect size, to perform posterior predictive checks, i.e. we simulate phenotype data. The simulated data for each genotype is then used to estimate the degree of overlap between the simulated phenotype distributions in two genetic states of the mutation. The overlap is quantified

using the Bhattacharyya coefficient (BC):

$$BC(p_1, p_2) = \int_x \sqrt{(p_1(x) \cdot p_2(x))} dx$$

where $p_1$ and $p_2$ are the simulated phenotype distributions in both genetic states of a mutation. For a complete overlap BC = 1 (i.e. no difference between the phenotype distributions in the two genetic state), and BC = 0 for no overlap (significant difference).

### 2.3 Phylogenetic Bias ($B$)

To control for potential phylogenetic biases (population structure), we devised the following procedure. First, we use the complete genotype data (all SNPs) to compute a kinship matrix (NxN dissimilarity matrix between the N-individuals). Alternatively, the users can provide their own kinship matrix (e.g. estimated using more accurate phylogenetic methods). For a group of individuals which belong to a group defined by an alleles of a given SNP, we next compute their mean kinship distance using the kinship matrix data. If the individuals in the group are related, the compute mean kinship distance must be significantly lower than the mean kinship distance computed from the complete kinship matrix. Thus, we define the phylogenetic bias as:

$$B = 1 - \hat{d}_g / \hat{d}_t$$

where $\hat{d}_g$ is the mean kinship distance between the individuals who share the genotype $g$; $\hat{d}_t$ is the mean kinship distance of the complete kinship matrix. For a complete phylogenetic bias, B = 1 ($\hat{d}_g << \hat{d}_t$), and $B = 0$ (or slightly negative) for no bias. This estimate is computed for each SNP and genotype group within each SNP.

To compute the phylogenetic bias associated with a mutation ($g_1 -> g_2$), we compute:

$$B = 1 - min(\hat{d}_{g_1}, \hat{d}_{g_2}) / \hat{d}_t$$

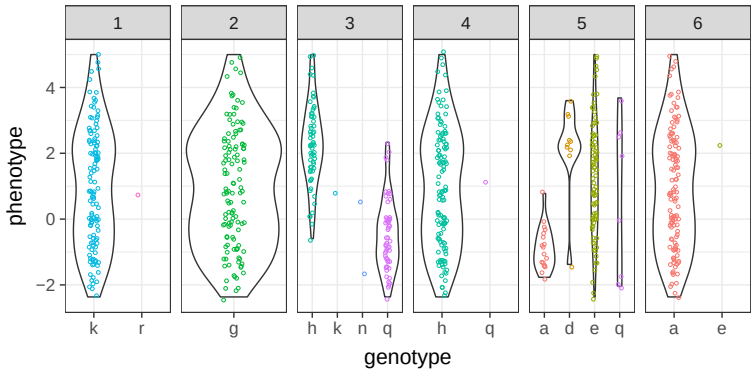
where $\hat{d}_{g_1}$ and $\hat{d}_{g_2}$ represent the mean kinship distance between the individuals who share the genotype (allele) $g_1$ and $g_2$ or a given SNP; $\hat{d}_t$ is the mean kinship distance in the complete kinship matrix. For a complete phylogenetic bias, B = 1 and B = 0 (or slightly negative) for no bias. This estimate is computed for each SNP and each pair of genotypes.

# 3 Case studies

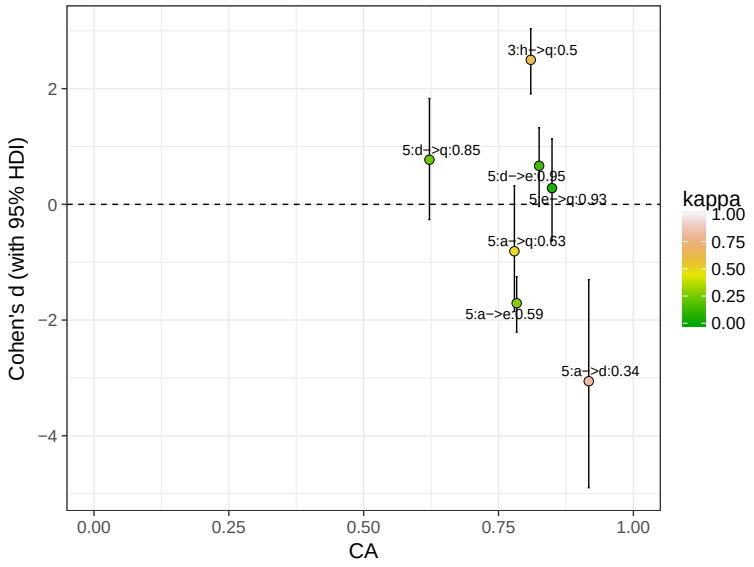
## 3.1 I: Association between SNPs and a *continuous* phenotype

In the first case study, we show a typical genotype-phenotype analysis, whereby the genotype is a protein sequence alignment composed of 6 sites and 120 individuals (sequences), and a continuous phenotype measured for each of the individuals.

**Genotype-phenotype data** First we show an overview of the distribution of the phenotype across each of the 6 studied polymorphic sites in the sequence alignment, and the underlying genotype states. `genphen` will list the mutations found at each size, followed by quantification fo the association strength as explained in 2.




**Association analysis** A typical way of visualizing the `genphen` results is with the following plot where each point represents a polymorphism (here SAAP) plotted according to x=classification accuracy ($CA$), y=Cohen's $d$, color=Cohen's $\kappa$. The most promising SAAPs have CA and $\kappa$ close to 1, with Cohen's $d$ estimate whose 95% highest density interval (HDI) does not overlap with the null effect (dashed line in figure: $d = 0$). The labels indicate the MSA site number, followed by the type of the polymorphism, and finally the BC score.




The association scores are also shown in the following table:

| site | mutation | cohens.d | cohens.d.hdi | bc | ca | ca.hdi | kappa | kappa.hdi |
|---|---|---|---|---|---|---|---|---|
| 3 | h->q | 2.50 | (1.91, 3.03) | 0.50 | 0.81 | (0.74, 0.89) | 0.62 | (0.48, 0.78) |
| 5 | a->d | -3.06 | (-4.9, -1.3) | 0.34 | 0.92 | (0.75, 1) | 0.82 | (0.48, 1) |
| 5 | a->e | -1.71 | (-2.21, -1.25) | 0.59 | 0.78 | (0.69, 0.88) | 0.28 | (-0.06, 0.51) |
| 5 | a->q | -0.81 | (-1.85, 0.32) | 0.63 | 0.78 | (0.58, 0.93) | 0.48 | (0, 0.82) |
| 5 | d->e | 0.67 | (-0.04, 1.32) | 0.95 | 0.83 | (0.72, 0.91) | 0.14 | (-0.14, 0.43) |
| 5 | d->q | 0.77 | (-0.26, 1.83) | 0.85 | 0.62 | (0.36, 0.86) | 0.24 | (-0.31, 0.6) |
| 5 | e->q | 0.28 | (-0.63, 1.13) | 0.93 | 0.85 | (0.76, 0.92) | 0.05 | (-0.13, 0.33) |

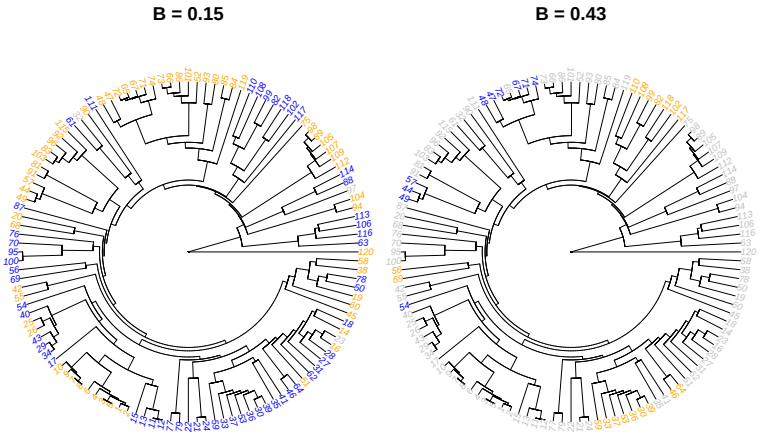
**MCMC convergence** Next, we want to check the validity our Bayesian inference model by inspecting the `genphen` output named convergence which contains information about the markov chain monte carlo (MCMC) simulation done with R package rstan including potential scale reduction factor (Rhat) and effective sampling size (ESS), as well as information concerning potential convergence issues such as divergences, tree depth exceeded warnings, etc. The small sample size of specific alleles (of a site) are often the cause of such warnings, which are then reported for the the alleles of that site.

| s | g | n | mu.rhat | sigma.rhat | mu.ess | sigma.ess | divergence | treedepth |
|---|---|---|---|---|---|---|---|---|
| 3 | h | 62 | 1.00 | 1.00 | 3000.00 | 2843.81 | FALSE | FALSE |
| 3 | q | 55 | 1.00 | 1.00 | 3000.00 | 3000.00 | FALSE | FALSE |
| 5 | a | 18 | 1.00 | 1.00 | 2736.10 | 1809.73 | FALSE | FALSE |
| 5 | d | 10 | 1.00 | 1.00 | 1722.14 | 1997.50 | FALSE | FALSE |
| 5 | e | 84 | 1.00 | 1.00 | 3000.00 | 3000.00 | FALSE | FALSE |
| 5 | q | 8 | 1.00 | 1.00 | 3000.00 | 1616.72 | FALSE | FALSE |

**Phylogenetic bias control** Next, we compute the phylogenetic bias of each mutation, shown in the table below:

| site | mutation | bias |
|---|---|---|
| 82 | h->q | 0.15 |
| 83 | h->q | 1.00 |
| 84 | a->d | 0.43 |
| 84 | a->e | 0.29 |
| 84 | a->q | 0.29 |
| 84 | d->e | 0.43 |
| 84 | d->q | 0.43 |
| 84 | e->q | 0.19 |
| 85 | a->e | 1.00 |

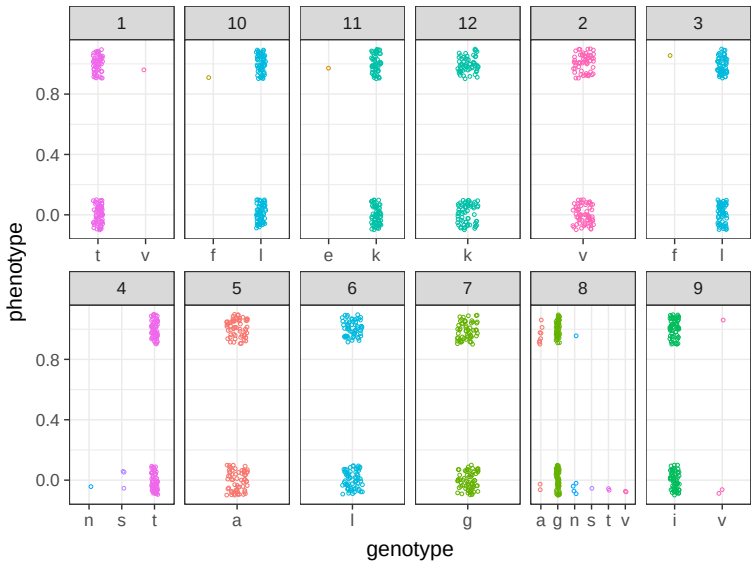
We use the kinship matrix to perform hierarchical clustering, visualizing the population strcuture and two examples (mutations) with genotype 1 marked with blue and genotype 2 marked with orange in either case. Individuals not covered by either genotype are marked with gray color. The shown examples differ in the degree of phylogenetic bias.



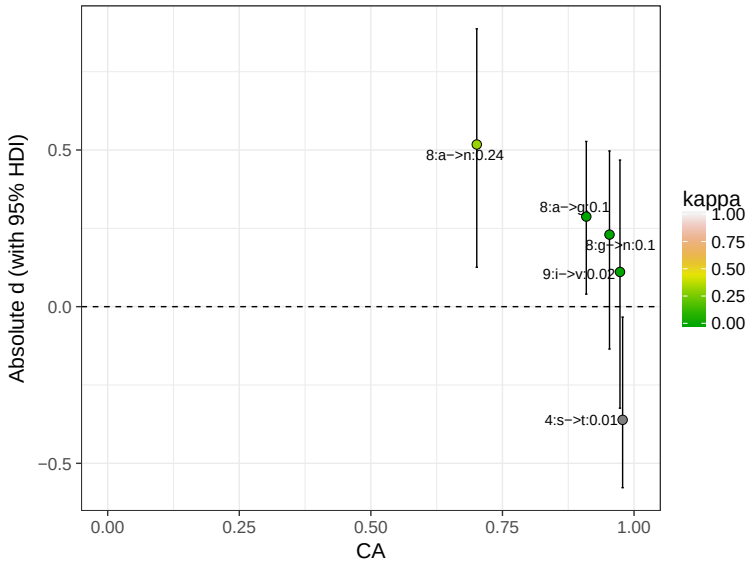

## 3.2 II: Association between SNPs and a *dichotomous* phenotype

In the second case study we show you how to use `genphen` in case the phenotype is of dichotomous type. The genotype input is a protein sequence alignment composed of 12 sites and 120 individuals (sequences), and the phenotype is a vector of 120 dichotomous values measured for each individual.

**Genotype-phenotype data** First we show an overview of the distribution of the phenotype across each of the 12 studied polymorphic sites in the sequence alignment, and the underlying genotype states. `genphen` will list the mutations found at each size, followed by quantification fo the association strength as explained in 2.




**Association analysis** A typical way of visualizing the `genphen` results is with the following plot where each point represents a polymorphism (here SAAP) plotted according to x=classification accuracy ($CA$), y=absolute $d$ (with error bars 95% HDI), color=Cohen's $\kappa$. The most promising SAAPs have CA and $\kappa \approx 1$, with absolute $d$ estimate whose 95% highest density interval (HDI) does not overlap with the null effect ($d \neq 0$). The labels indicate the MSA site number, followed by the type of the polymorphism, and finally the BC score.




**MCMC convergence** Next, we want to check the validity our Bayesian inference model by inspecting the `genphen` output named convergence. A similar analysis was performed and described in 3.1.

| s | g | n | mu.rhat | mu.ess | divergence | treedepth |
|---|---|---|---|---|---|---|
| 4 | s | 3 | 1.00 | 1888.17 | FALSE | FALSE |
| 4 | t | 116 | 1.00 | 1370.45 | FALSE | FALSE |
| 8 | a | 10 | 1.00 | 1740.39 | FALSE | FALSE |
| 8 | g | 100 | 1.00 | 2288.70 | FALSE | FALSE |
| 8 | n | 5 | 1.00 | 2617.30 | FALSE | FALSE |
| 9 | i | 117 | 1.00 | 1552.66 | FALSE | FALSE |
| 9 | v | 3 | 1.00 | 1494.71 | FALSE | FALSE |

# 4 Extra Utilities

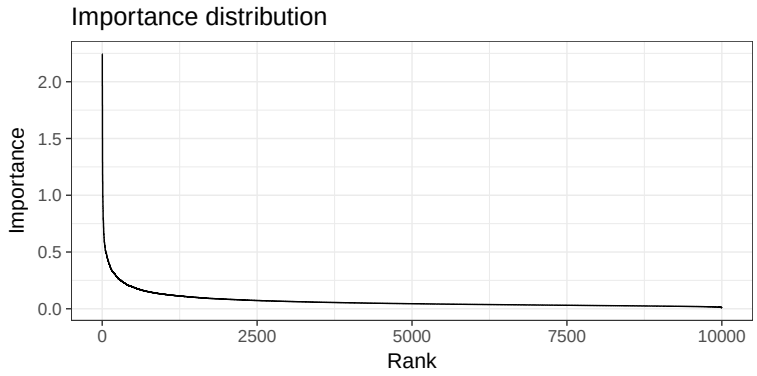
## 4.1 Data Reduction

The methods implemented in `genphen` are statistically superior to the ones implemented by classical tools for GWAS. This however comes at the cost of increased computational burdain. Therefore, using `genphen` to study the association between hundreeds of thousands of SNPs and a phenotype can be quite costly. Motivated by the biological fact that a major fraction of the SNPs are non-informative (genetic noise) with respect to the selected phenotype, we implemented a low-weight diagnostics procedure in `genphen`, which allows us to quickly scan the SNP space and quickly discard large portion of the non-informative SNPs.

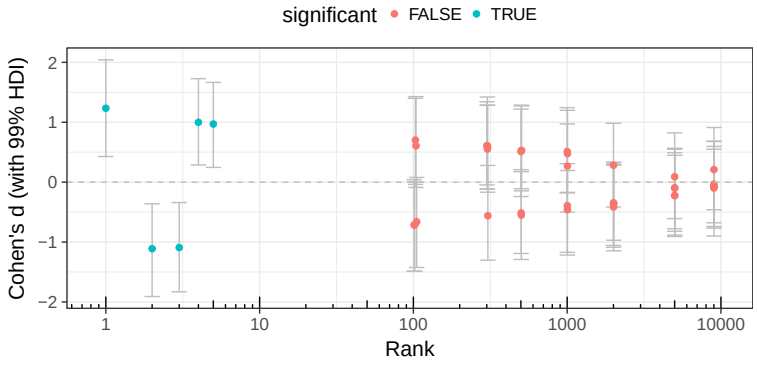
The data reduction procedure includes the following steps:

1. using random forest and their variable importance measures, we obtain one importance value for each SNP.
2. next, we rank the SNPs by their importance and use the importance distribution as a 'rough' guide to sample and evaluate SNPs (so-called anchor SMPs) using a lighter-weight version of the standard genphen approach.
3. using the previously explained `genphen` association scores.
4. we can therefore determine the importance rank at which the SNPs no longer carry any information and would be discarded, reduce the data and perform the analysis on the remaining data using the standard `genphen` method.

Using a case study, we explain the typical data reduction steps in more detail. First we use random forest to get the distribution of variable (SNP) importance, shown in the figure below.




We then select a set of anchor points (SNPs) to analyze. In this particular example we selected 25 anchor points with ranks in the interval 1-5, 101-105, 301-305, 501-505, 1001-1005, 2001-2005, 5001-5005 and 9001-9005 from the total set of 10,000 genotypes. We then visualize the estimated Cohen's d effect size (and 99% HDI) as shown in the figure below. The anchor points (SNPs) are shown as points, colored green if the HDI (gray error bar) does not include the 0 effect, and green otherwise. We observe that the non-informative SNPs become prevalent past rank 100. We can therefore select all SNPs with rank higher than 100 and perform the main analysis with only 100 SNPs (yielding 99% data reduction). We might also select a more conservative threshold such as 500 (yielding 95% data reduction).


significant
FALSE
TRUE
Cohen's d (with 99% HDI)
2
1
0
−1
−2
1
10
100
1000
10000
Rank