Fewer permutations, more accurate P-values
Theo A. Knijnenburg1,∗, Lodewyk F. A. Wessels2, Marcel J. T. Reinders3 and Ilya Shmulevich1
1Institute for Systems Biology, Seattle, WA, USA, 2Bioinformatics and Statistics, The Netherlands Cancer Institute, Amsterdam and 3Information and Communication Theory Group, Delft University of Technology, Delft, The Netherlands

ABSTRACT
Motivation: Permutation tests have become a standard tool to assess the statistical significance of an event under investigation. The statistical significance, as expressed in a P-value, is calculated as the fraction of permutation values that are at least as extreme as the original statistic, which was derived from non-permuted data. This empirical method directly couples both the minimal obtainable P-value and the resolution of the P-value to the number of permutations. Thereby, it imposes upon itself the need for a very large number of permutations when small P-values are to be accurately estimated. This is computationally expensive and often infeasible.

Results: A method of computing P-values based on tail approximation is presented. The tail of the distribution of permutation values is approximated by a generalized Pareto distribution. A good fit and thus accurate P-value estimates can be obtained with a drastically reduced number of permutations when compared with the standard empirical way of computing P-values.

Availability: The Matlab code can be obtained from the corresponding author on request.

Contact: tknijnenburg@systemsbiology.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Permutation tests (also called randomization tests) are non-parametric procedures for determining statistical significance based on rearrangements of the labels of a dataset (Edgington, 1980). A test statistic, which is computed from the dataset, is compared with the distribution of permutation values. These permutation values are computed similarly to the test statistic, however, under a random rearrangement (permutation) of the labels of the dataset.

Permutation tests have become a widely used technique in bioinformatics. The non-parametric nature of these tests rationalizes their usability and popularity, since in many bioinformatics applications there is no solid evidence or sufficient data to assume a particular model for the obtained measurements of the biological events under investigation.

For example, Significance Analysis of Microarrays (SAM) (Tusher et al., 2001) and Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), which detect differentially expressed genes and gene sets, respectively, are two well-known techniques that use permutation tests to compute statistical significance. In these approaches, the class labels of samples from which gene expression measurements are taken, are randomly rearranged to obtain the permutation values. Besides randomization over the set of samples, permutation tests have also been performed by randomizing over the set of genes (Bretz et al., 2004; Smyth, 2004). In this case, the labels are binary indicator variables that indicate whether a gene belongs to a particular gene set or not. Efron and Tibshirani (2007) suggested to employ both permutation types to test the significance of gene sets. Other examples of permutation tests in bioinformatics include, but are not limited to: QTL detection (Doerge and Churchill, 1996), allelic association analysis (Zhao et al., 2000) and modeling ChIP sequencing (Zhang et al., 2008). In the latter case, each permutation corresponds to the simulation of a complete ChIP-seq experiment.

As in all statistical hypothesis tests, the significance of a permutation test is represented by its P-value. The P-value is the probability of obtaining a result at least as extreme as the test statistic, given that the null hypothesis is true. In permutations tests, the null hypothesis is defined as: the labels assigning samples to classes are interchangeable (Edgington, 1980). Significantly, low P-values indicate that the labels are not interchangeable and that the original label configuration is relevant with respect to the data. The P-value is assessed by performing all possible permutations and computing the fraction of permutation values that are at least as extreme as the test statistic obtained from the unpermuted data.

However in practical situations, it is (by far) not feasible to perform all possible permutations. For example, class labels that represent two classes with 50 samples each can be permuted in \(100! / 50! \times 50! \) different ways. Therefore, the P-value is approximated by computing a limited number of permutations, say \(N\), and then computing the fraction of the \(N\) permutation values that are at least as extreme as the test statistic. Usually, a pseudocount is added to avoid P-values of zero, which occur when the test statistic is never surpassed by the permutation values. Theoretically, a P-value of zero is not possible in the context of permutation tests: the minimum is \(1/N_{all}\), where \(N_{all}\) is the number of all possible permutations. This is because one of the permuted label configurations is identical to the original one, under which the test statistic is computed.

This empirical approximation of computing P-values has two direct consequences. First, the resolution of obtainable P-values is \(1/N\). Second and more important, the smallest achievable P-value is \(1/N\). This means that a very large number of permutations is required to accurately estimate a small P-value. In general, \(>N\) permutations are required to reliably estimate a P-value of \(1/N\). (This is shown more extensively later in the manuscript.) Especially in bioinformatics, it is crucial to be able to accurately determine small P-values. This is due to typically huge numbers of objects [e.g. all genes, gene sets or single nucleotide polymorphism (SNPs)] that are simultaneously tested, which requires large multiple testing
corrections to prevent large numbers of false positives. Clearly, other significance scores that are based on the \( P \)-values, such as the false discovery rate (FDR), will be meaningless when the \( P \)-values are not estimated correctly.

In this work, we propose to estimate the small permutation test \( P \)-values using extreme value theory (Gumbel, 1958). The set of extreme (very large or very small) permutation values that forms the tail of the distribution of permutation values is modeled as a generalized Pareto distribution (GPD). Pickands (1975) demonstrated that the GPD approximates the distribution of the extreme values of a set of independent and identically distributed (i.i.d), random variables, i.e. those values that exceed a particular (high) threshold. Applications of the GPD to model extreme values are traditionally found in climatology to model extreme weather, such as floods, and in financial risk management to model extreme losses and insurance claims. In our case, the GPD, which is fitted on the extreme permutation values, is evaluated at the value of the test statistic to estimate the \( P \)-value of the permutation test. Both theoretical probability distributions as well as gene expression datasets are employed to demonstrate that the proposed tail approximation strategy leads to an accurate estimation of the correct \( P \)-value using far fewer permutations compared with the standard empirical approach.

2 METHODS

2.1 Problem definition

Given test statistic \( x_0 \) and set \( X \), which contains all possible permutation values, \( x_1^*, x_2^*, \ldots, x_n^* \), the permutation test \( P \)-value is defined as

\[
P_{\text{perm}} = \frac{1 + \sum_{i=1}^{N} I(x_i^* \geq x_0)}{N}
\]

where \( I(\cdot) \) is the indicator function. The goal is to approximate \( P_{\text{perm}} \) using a randomly sampled subset \( Y \subset X \), which contains \( N \) permutation values, \( y_1^*, y_2^*, \ldots, y_n^* \). Usually, \( N \ll N_0 \).

Note that the \( P \)-value calculation as described above corresponds to a right-tailed test. The \( P \)-value approximations discussed in this section will all correspond to a right-tailed test. Conversion to the left-tailed test and the two-tailed test is in all cases straightforward.

2.2 Empirical cumulative distribution function approximation

The standard approximation to \( P_{\text{perm}} \) is computed similarly to (1). Commonly, it includes a pseudocount to avoid \( P \)-values of zero:

\[
P_{\text{perm}} = \frac{1 + \sum_{i=1}^{N} I(x_i^* \geq x_0)}{N}
\]

Essentially, this approach employs the permutation values in \( Y \) to build an empirical cumulative distribution function (ECDF). The ECDF is a step function that increases by \( 1/N \) at the value of each (ordered) permutation value in \( Y \). Usually, \( N \ll N_0 \) and then adding the pseudocount of 1 at \( x_0 \). Figure 1 illustrates the concept of the ECDF by approximating an \( F \) distribution using a limited number of samples randomly drawn from this distribution.

2.3 GPD approximation

The tail of the distribution of permutation values is modeled using the GPD. The GPD has cumulative distribution function (CDF)

\[
P(z) = \begin{cases} 1 - \left(1 - k/z\right)^{1/k}, & k \neq 0 \\ 1 - e^{-z/a} & k = 0 \end{cases}
\]

The range of \( z \) is \( 0 \leq z < \infty \) for \( k \leq 0 \) and \( 0 \leq z < a/k \) for \( k > 0 \). The parameters of the GPD are \( a \), the scale parameter, and \( k \), the shape parameter. For the special values \( k = 0 \) and \( 1 \), the GPD becomes the exponential and uniform distribution, respectively. When \( k = 0 \) the GPD becomes the Pareto distribution, which has a long tail. The argument of the GPD, \( x \), are the exceedances. In our case, these are the permutation values in \( Y \) that exceed threshold \( t \), which then get subtracted by \( t \) to form the exceedances. Formally, if the values in \( Y \) are ordered, such that \( x_i^* \geq x_{i+1}^* \), we have a set \( Z \) of \( N_0 \) exceedances, \( z_1^*, z_2^*, \ldots, z_n^* \), where \( z_i^* = x_i^* - t, i = 1, \ldots, n \).

Maximum likelihood (ML) estimation is employed to estimate \( a \) and \( k \) given \( Z \) as explained in Hosking and Wallis (1987) and Grimshaw (1993).

Note that there exist other techniques to estimate the GPD parameters. We employed ML estimation, since this is the most commonly used technique and has overall good performance on reasonably large sample sizes as used in

![Fig. 1. ECDF approximation of an F distribution. (a) From the PDF of the F distribution, 25 samples are randomly drawn. These samples represent the permutation values. The theoretical \( P \)-value for test statistic \( x_0 \) equals the grey area, which is 0.047. (b) The theoretical CDF is approximated by the ECDF, which is based on the 25 permutation values. Since two permutation values exceed \( x_0 \), \( P_{\text{perm}} \) is \((1 + 2) / 25 = 0.12\) (including the pseudocount).](https://example.com/fig1.png)
of this test is that the exceedances come from a GPD. Small P-values (of the goodness-of-fit test) indicate that this is not the case.

We propose to use the 250 most extreme permutation values as exceedances, i.e. \( N_{\text{exc}} = 250 \). We choose such a large number, because the GPD tail approximation is frequently used for extrapolation, i.e. the test statistic is much larger than the largest permutation value. (For example, take \( s_0 = 25 \) in Figure 2, where the largest permutation value is about 17.) A highly accurate estimate of the GPD parameters is required, because small deviations of the parameters can have a huge effect in the case of large extrapolation. If the 250 largest permutation values do not follow a GPD according to the goodness-of-fit test (\( P \leq 0.05 \)), the number of exceedances is iteratively decreased by ten until a GPD good fit (i.e. \( P > 0.05 \)) is reached. If a good fit is never reached, the GPD cannot be used. However, this situation did not occur in any of the theoretical and practical cases described in this article.

The exceedances threshold \( t \) is set to \((\sqrt{s_{0}^*} + \sqrt{N_{\text{exc}}})/2 \) (assuming that the values of \( t \) are ordered from high to low as before). Thus, \( t \) is right between the smallest permutation value that is part of the exceedances, \( \sqrt{s_{0}^*} \), and the one that just falls outside of the tail of extreme permutation values, \( \sqrt{N_{\text{exc}}+1} \).

### 2.4 Proposed algorithm

The GPD approximation can only be used when the test statistic is in the range of the extreme permutation values or when it is even larger. For example, when 50 out of the 100 permutation values exceed the test statistic, the test statistic is not in the tail of the distribution of permutation values and the GPD tail approximation is useless. Furthermore, in that case the standard empirical method to compute the P-value is adequate. Therefore, we have developed a criterion to decide when to employ P-value estimation using the GPD tail approximation.

This criterion is based on the fact that the number of permutation values that exceed the test statistic follows a binomial distribution. This is because each generated permutation value can be seen as a Bernoulli trial with probability \( P_{\text{perm}} \) of success, i.e. the permutation value being larger or equal to the test statistic. Let \( M \) be the number of permutation values that exceed the test statistic, i.e.

\[
M = \sum_{i=1}^{N} I(y_i^* \geq t)
\]

Note that (from (2)) \( M = NP_{\text{perm}} \) if we would exclude the pseudocount. Let \( P_{\text{exc}} \) be \( P_{\text{perm}} \) excluding the pseudocount, i.e. \( P_{\text{exc}} = M/N \). According to the central limit theorem, if \( M \geq 10 \) one may rely on the normal approximation to the binomial distribution:

\[
M \sim N\left(NP_{\text{perm}}, N\left(1-P_{\text{perm}}\right)\right)
\]

Substituting \( P_{\text{perm}} \) by its estimate \( P_{\text{exc}} \) gives us an estimate of the confidence bounds on \( P_{\text{exc}} \):

\[
P_{\text{exc}} \sim N\left(P_{\text{exc}}\right) \in \left[ P_{\text{exc}} \pm 1.96 \sqrt{P_{\text{exc}}(1-P_{\text{exc}})/N} \right]
\]

This procedure of determining the confidence bounds on the P-value estimate is identical to the one described in Nettleton and Doerge (2000). Note that the pseudocount can be omitted, because P-values of zero can no longer occur when \( M > 0 \).

Since when \( M \geq 10 \) we can reliably compute \( P_{\text{exc}} \) and its confidence bounds, we propose the following algorithm.

#### Algorithm 1

**Given:**
- test statistic \( t_0 \)
- \( N \) permutation values \( y_1^*, y_2^*, \ldots, y_N^* \)

**Compute:**
- \( M \)
  - if \( M \geq 10 \)
    - Compute \( P_{\text{exc}} \)
    - Compute \( P_{\text{exc}} \) and
  - otherwise
    - Compute \( P_{\text{exc}} \) and
3 RESULTS

3.1 Theoretical distributions

Seven different distribution functions, ranging from light-tailed to heavy-tailed, were employed to test the GPD approximation. See Table 1, where distribution functions are ordered from light-tailed to heavy-tailed with the most light-tailed one on the left. Permutation values are obtained by randomly drawing samples from these distributions. The theoretical permutation test \( P \)-value can be obtained by evaluating the CDF at the value of the test statistic. Since an infinite number of samples can be generated from a distribution function, this theoretical \( P \)-value is \( P_{\text{true}} \) from (1) in the limit case where \( N_{\text{perm}} \) approaches infinity. For each distribution function, we chose a set of eight test statistics such that \( P_{\text{perm}} \approx \alpha \). These are the situations, where the test statistic is intentionally focus on the situations were the GPD approximation can be useful. These are the situations, where the test statistic is larger than (almost) all permutation values, i.e. \( N_{\text{perm}} \) are large range that the \( P \)-value estimates \( P_{\text{ecdf}} \) and \( P_{\text{gpd}} \) were computed. This experiment was repeated 1000 times.

Note that \( P_{\text{gpd}} \) is computed according Algorithm 1, i.e. if \( M > 10 \) we compute \( P_{\text{ecdf}} \) in stead of \( P_{\text{gpd}} \). This, however, almost never occurs in the experiments presented in this section, since we intentionally focus on the situations were the GPD approximation can be useful. These are the situations, where the test statistic is larger than (almost) all permutation values, i.e. \( M = 0 \) or very small.

3.1.1 \( P \)-value estimates as a function of \( N \)  Figure 3 displays a typical result. The ECDF approximation converges to the correct \( P \)-value linearly with the number of permutations, \( N \). This behavior can be attributed to the effect of the pseudocount of \( 1/N \). In general, when \( N < 1/P_{\text{perm}} \), \( P_{\text{ecdf}} = 1/N \). If this pseudocount was omitted \( P_{\text{ecdf}} \) would be zero, until a sufficiently large number of permutations was performed.

In contrast, the GPD approximation converges with far fewer permutations. In Figure 3, a decent estimate of \( P_{\text{perm}} \) is obtained with \( \approx 10^3 \) permutation values (see also Table 1). However, when \( N < 1/P_{\text{perm}} \), there is a lot of variability in \( P_{\text{gpd}} \), illustrated by the large range that the \( P \)-value estimate assumes in this case. This range can even include \( P \)-value estimates of zero. This occurs when the range of \( z \) of the estimated GPD [in (3)] is limited, i.e. \( k > 0 \), and the original statistic falls outside of this range, i.e. \( \sqrt{2\pi} - 1 > a/k \).

A large variance in the \( P \)-value estimate (including \( P \)-values of zero) for small \( N \) are more frequently observed for the light-tailed distributions, where, indeed, \( k > 0 \). See Supplementary Material, which contains all figures (similar to Fig. 3) for the seven distribution functions and the eight values of \( P_{\text{perm}} \).

3.1.2 Number of permutations required for convergence Table 1 provides an overview of the number of permutations necessary before convergence to a reasonable estimate. This number, \( N_{c} \), is computed for the different distribution functions and for different

| Distribution | Poisson | Normal | \( \chi^2 \) | Exponential | F | Log-normal | Cauchy |
|--------------|---------|--------|-------------|-------------|---|------------|--------|
| Arguments    | \( \lambda = 10^6 \) | \( \mu = 0 \), \( \sigma = 1 \) | \( d_1 = 3 \) | \( \lambda = 1 \) | \( d_1 = 5 \), \( d_2 = 10 \) | \( \mu = 0 \), \( \sigma = 2 \) | \( t = 0, i = 1 \) |
| \( k_{\text{perm}} \) | \( 10^3 \) | \( 10^4 \) | \( 10^5 \) | \( 10^6 \) | \( 10^7 \) | \( 10^8 \) | \( 10^9 \) |
| \( P_{\text{perm}} \) | \( 9.6 \times 10^9 \) | \( 9.0 \times 10^8 \) | \( 8.7 \times 10^5 \) | \( 9.1 \times 10^3 \) | \( 9.7 \times 10^1 \) | \( 9.7 \times 10^0 \) | \( 9.6 \times 10^0 \) |
| \( P_{\text{ecdf}} \) | \( 4.4 \times 10^3 \) | \( 4.1 \times 10^4 \) | \( 2.2 \times 10^5 \) | \( 2.0 \times 10^3 \) | \( 1.7 \times 10^1 \) | \( 1.1 \times 10^0 \) | \( 8.4 \times 10^0 \) |
| \( P_{\text{gpd}} \) | \( 3.5 \times 10^4 \) | \( 3.6 \times 10^3 \) | \( 3.5 \times 10^2 \) | \( 3.5 \times 10^1 \) | \( 3.6 \times 10^0 \) | \( 3.6 \times 10^0 \) | \( 3.6 \times 10^0 \) |
| \( P_{\text{ecdf}} \) | \( 2.2 \times 10^3 \) | \( 2.3 \times 10^4 \) | \( 5.8 \times 10^5 \) | \( 8.0 \times 10^3 \) | \( 5.4 \times 10^1 \) | \( 3.3 \times 10^0 \) | \( 1.1 \times 10^0 \) |
| \( P_{\text{gpd}} \) | \( 2.6 \times 10^6 \) | \( 2.6 \times 10^3 \) | \( 2.6 \times 10^2 \) | \( 2.6 \times 10^1 \) | \( 2.6 \times 10^0 \) | \( 2.6 \times 10^0 \) | \( 2.6 \times 10^0 \) |
| \( P_{\text{ecdf}} \) | \( 7.1 \times 10^4 \) | \( 6.2 \times 10^3 \) | \( 3.9 \times 10^4 \) | \( 3.0 \times 10^3 \) | \( 1.4 \times 10^2 \) | \( 2.1 \times 10^1 \) | \( 1.3 \times 10^0 \) |
| \( P_{\text{gpd}} \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) |
| \( P_{\text{ecdf}} \) | \( 5.4 \times 10^6 \) | \( 5.6 \times 10^5 \) | \( 1.7 \times 10^4 \) | \( 1.4 \times 10^3 \) | \( 3.3 \times 10^2 \) | \( 1.5 \times 10^1 \) | \( 1.5 \times 10^0 \) |
| \( P_{\text{gpd}} \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) |
| \( P_{\text{ecdf}} \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) |
| \( P_{\text{gpd}} \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) | \( > 10^6 \) |

The top three rows of the table state the names of distribution functions, their arguments and their ranges of the estimated scale parameter \( k \). \( N_{c} \) is given for both estimators (\( P_{\text{ecdf}} \) and \( P_{\text{gpd}} \) for a range of different \( P \)-values \( P_{\text{perm}} \).
This behavior is not observed for the ECDF approximation, where 170 microarrays of yeast chemostat cultivations (Knijnenburg et al. 2009) employed oxygen regime, i.e. in 80 of these arrays yeast was grown anaerobically; for the other 90 arrays yeast was grown aerobically. In this experiment, we focused on the 132 genes annotated with MIPS (Meuwese et al., 1997) function category ‘respiration’, since we expected to find many differentially expressed genes in this group. It is computationally infeasible to compute \( P_{\text{gpd}} \) from (1), since \( N_{\text{gpd}} = \sum_{i} 10^{N_i} \). Therefore, we applied the following strategy to approximate \( P_{\text{gpd}} \): for each gene, we generated permutation values until \( M \), the number of permutation values that exceeds the test statistic, \( \geq 25 \). Then, using (8) we can reliably estimate \( P_{\text{gpd}} \). For 69 of the 132 genes, \( M < 25 \) even after \( > 3 \) billion \( N_i \times 10^{3} \) permutations. The other 63 genes, for which \( P_{\text{gpd}} \) was reliably estimated, were used in the rest of the experiment. For different values of \( N_i \), we computed \( P_{\text{gpd}} \) and \( P_{\text{ecdf}} \) for the 63 genes. This experiment was repeated 200 times. Figure 4 visualizes the results for \( N = 10^{5} \). As expected, the ECDF approximation is adequate for genes with \( P_{\text{gpd}} < 10^{-6} \). However, for smaller \( P \)-values, the estimate is way off and bounded by \( 1/N \). The GPD approximation provides a better approximation for the small \( P \)-values. However, for \( P \)-values \( > 10^{-6} \), the variance of the estimate becomes quite large and frequently \( P \)-value estimates of zero are encountered. This behavior was also observed with the light-tailed theoretical distributions. Accordingly, the range of the shape parameter \( k \) estimated on the permuted SAM statistics was \([0, 0.27]\), similar to range of the light-tailed distributions.

In a follow-up experiment, we aimed to transform the test statistic and its permutation values, such that \( k < 0 \), i.e. the tail becomes more heavy. Basically, any strictly increasing (and thus order preserving) function can be applied to the test statistic and its permutation values without changing (the definition of) \( P_{\text{gpd}} \). Also, such a transformation would not influence the computation of \( P_{\text{ecdf}} \). And, if the tail of the transformed permutation values still follows a GPD, which is tested using the goodness-of-fit test, they can be used to estimate \( P_{\text{gpd}} \). We raised all test statistics and corresponding permutation values to the power three, i.e. \( d'_{i} = (d_{i})^{3} \), \( d''_{i} = (d'_{i})^{3} \), \( \forall i, n \) and recomputed \( P_{\text{gpd}} \). After the transformation, the estimated range of the shape parameter \( k \) was \([-0.6, 0]\). Furthermore, \( P_{\text{gpd}} \) based on the transformed statistic proved to be a better estimate with much less variance (Fig. 5). Now a reasonable estimate of \( P \)-values \( < 10^{-7} \) can be made using only \( 10^{5} \) permutations.

More discussion about transforming the permutation values will follow in Section 3.2.3.

3.2.2 Enrichment of gene sets Another popular application of the permutation test on expression data is to uncover enriched gene sets. These are a priori defined groups of genes, e.g. functionally related genes, that show concurrent differential expression across the different conditions or classes of a microarray experiment. In this case, the gene-set specific test statistic is compared with the permutation values, which are obtained by permuting the class labels and recomputing the statistic.

GSEA (Subramanian et al., 2005) is the most widely used method that follows this approach. The GSEA statistic is a weighted version of the Kolmogorov–Smirnov statistic, \( \epsilon_{i} \), where \( i \) is the index of a gene set. GSEA uses the empirical way of estimating the permutation \( P \)-value based on \( \epsilon_{i} \) and its permutation values \( \epsilon'_{i}, \epsilon''_{i}, \ldots, \epsilon_{i}^{\prime \prime \prime} \).

In comparison with SAM, GSEA is computationally very expensive, as it involves sorting all genes based on correlation with the class labels and then computing a running sum statistic. For some
bioinformatics applications, dynamic programming approaches can be used to build a suitable null distribution based on the statistic (Newberg and Lawrence, 2009). One such approach has been developed for GSEA (Keller et al., 2007) in order to avoid the expensive permutations. However, this method only applies to the unweighted version of GSEA (Mootha et al., 2003), which is not the default setting of the GSEA algorithm.

We applied (default) GSEA to the van de Vijver et al. (2002) breast cancer dataset consisting of >24,000 gene expression measurements of 295 patients, 180 with poor prognosis and 115 with good prognosis. All gene sets from gene ontology (GO) (Ashburner et al., 2000) and kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa and Goto, 2000) containing >10 genes were analyzed.

$P_{\text{ecdf}}$ was computed similarly to the previous section, i.e. for each gene set we generate permutation values until $M > 25$. Since the GSEA statistic is so computationally expensive, not more than 1,000,000 permutations ($N = 10^6$) were performed. We found 89 gene sets for which $M > 25$ within the $10^5$ permutations and had a $P_{\text{perm}} < 0.01$.

In this experiment, we focused on a different aspect of the outcome of the permutation test, namely the order of the gene sets based on the estimated $P$-value. Often, researchers are interested in the top of the list of significant genes, gene sets, SNPs, etc., more or less regardless of the associated significance scores themselves. These objects will then be analyzed or investigated in order of significance, starting with the most significant. Here, we compare the correctly ordered list of 89 gene sets for which $P_{\text{ecdf}}$ was larger or smaller than (or equal to) $P_{\text{gpd}}$ out of the 100 repeats.

The results are visualized in Figure 6a. For values of $N < 250$ the GPD approximation performs worse than the standard empirical approach. This could be expected, since for such small values of $N$,
Enhanced P-value estimation for permutation tests

Table 2. The 25th and 75th percentile values of \( N_c \) and the corresponding P-value estimates (\( P_{\text{perm}} \) and \( P_{\text{gpd}} \)) for five different genes

| Gene  | \( P_{\text{perm}} \) | \( \log_{10}(P_{\text{perm}}) \) | \( \log_{10}(P_{\text{gpd}}) \) |
|-------|-----------------|-----------------|-----------------|
| YHR011W | 9.4 \times 10^{-9} | -4.03 | -4.3 |
| YDR079W | 9.9 \times 10^{-9} | -5 | -4.6 |
| YIL045W | 9.4 \times 10^{-7} | -6.03 | -6.4 |
| YML125C | 6.7 \times 10^{-8} | -7.17 | -7.2 |
| YLR044C | 1.2 \times 10^{-8} | -7.91 | -7.8 |

The power \( Z \) to which the original statistic and its permutation values were raised before GPD approximation is indicated on the left side of the table.

Table 2 displays the results for the five genes, for which \( P_{\text{perm}} \) was the closest to \( 1 \times 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8} \), respectively. These genes were chosen to present results for a large range of different \( P_{\text{perm}} \) values.

From the table, it is clear that fewer permutations are necessary to reach an accurate P-value estimate when employing the GPD approximation compared with the standard empirical approach. For the first two genes ~5–10 times fewer permutations are necessary. For the last three genes, convergence was not reached with the empirical approach within \( 10^6 \) permutations. However, based on the \( N_c \) values from the \( P_{\text{gpd}} \) estimates that did converge, we can infer that orders of magnitude fewer permutations are sufficient for a reliable estimate for these smaller P-values.

Additionally, we observed that the behavior of the confidence bounds on an individual \( P_{\text{gpd}} \) estimate (based on one set of permutation values) is comparable with the behavior of the confidence bounds (percentile values) based on many \( P_{\text{gpd}} \) estimates. That is, when too few permutations have been performed for an accurate estimate, \( P_{\text{gpd}} \) will tend to have large confidence bounds, and thus much uncertainty in the estimate.

With the convergence criteria, it is possible to analyze the effect of order-retaining transformations on the test statistic and its permutation values. From both the theoretical distributions and the application on gene expression data, we observed that the shape of the tail of permutation values influences the GPD estimation process. That is, for light-tailed distributions (\( k > 0 \)), where the GPD has a finite range, the estimate appears unstable and less accurate. Small changes in \( k \) have a huge effect on \( P_{\text{gpd}} \). Especially, in the case of large extrapolation, i.e. \( N \ll 1 / \text{perm} \), there is a large variance (or large confidence bounds) on the estimator and frequently P-value estimates of zero are encountered.

Order-retaining transformations can be applied to change the shape of the tail of the distribution of permutation values. While such a transformation does not affect \( P_{\text{perm}} \) and the computation of \( P_{\text{ecdf}} \), it might provide a more robust and accurate estimate (and prevents P-values of zero) for the GPD approximation. We applied power transformations to the SAM statistic in order to reduce \( k \).

The SAM statistic, which can assume both positive and negative values, was raised to the power \( Z = 1 \) (no transformation), 3 and 5. The ranges of \( k \) for \( Z = 1, 3 \) and 5 are [0, 0.27], [−1.6, −0.6] and [−1.6, −0.2], respectively. Table 2 also displays the results for \( Z = 3 \) and 5.

There are too few samples to accurately estimate the tail and extrapolate this to the correct P-value. However, for larger values of \( N \), the list ordered based on \( P_{\text{gpd}} \) leads to a higher correlation with the optimal ordering compared with the list ordered based on \( P_{\text{ecdf}} \). When \( N = 1000 \) (the standard number of permutations in GSEA), the Spearman correlation \( C_{\text{gpd}} \) is significantly > \( C_{\text{ecdf}} \). The difference between \( C_{\text{gpd}} \) and \( C_{\text{ecdf}} \) becomes more obvious in Figure 6b. For each of the 100 repetitions, we counted how many times the one correlation was higher than the other. When \( N = 1000 \), for almost all of the 100 trials, the GPD approximation led to a higher correlation, and thus to a more correctly ordered list compared with the list based on the empirical approach, which was computed on the same permutation values. When \( N \) increases, the difference between \( C_{\text{ecdf}} \) and \( C_{\text{gpd}} \) diminishes and in many cases they lead to identically ordered lists. This can be attributed to the fact that amongst the 89 genes, only few have a \( P_{\text{gpd}} \leq 10^{-4} \) and none is \( \leq 2 \times 10^{-4} \). Consequently, \( P_{\text{ecdf}} \) becomes a good approximation of \( P_{\text{perm}} \) when \( N \) approaches \( 10^6 \).

3.2.3 Choosing \( N \) In practical applications, \( P_{\text{perm}} \) is not known and no repeats are performed, i.e. only one set of permutation values is generated for a test statistic. The convergence criteria developed in Section 3.1.2 to decide when enough permutations have been performed can be slightly altered to suit practical applications:

- \( \log_{10}(P_{\text{perm}}(N)) \leq 0.1 \times \log_{10}(P_{\text{perm}}(N_c)) \) if \( N_c \leq N \)
- \( \log_{10}(P_{\text{est}}(N_c)) \leq 0.1 \times \log_{10}(P_{\text{est}}(N)) \) if \( N_c \leq N \)

where \( P_{\text{perm}}(N) \) is the estimated P-value (either \( P_{\text{ecdf}} \) or \( P_{\text{gpd}} \)) after \( N \) permutations, and \( P_{\text{est}}(N_c) \) is the \( \alpha \)% confidence bound on the estimated P-value. \( N_c \) is the minimum amount of permutations at which these criteria are met. For \( P_{\text{ecdf}} \) it is not possible to compute confidence bounds. In that case, only the first criterion applies. \( P_{\text{ecdf}} \) and \( P_{\text{gpd}} \) estimates (including confidence bounds) were computed for the 63 genes of Section 3.2.1 based on their SAM statistic and corresponding permutation values. The number of permutations was increased until the convergence criteria were met or when the maximum number of permutations (\( N = 10^6 \)) was reached. This experiment was repeated 25 times.
Although the transformation is not necessary to outperform the ECDF approximation, it does lead to convergence with even fewer permutations. This gain in improvement does not continue for larger values of \( Z \) (not in the table), where \( N_r \) starts to increase again. From this analysis, we can conclude that a transformation can be useful when the shape of the tail of the distribution of permutation values is transformed from light-tailed (\( k > 0 \)) to heavy-tailed (\( k < 0 \)), where our estimator has more stable and accurate performance.

4 DISCUSSION

The non-parametric nature of the permutation test rationalizes its least to a more sensible division of the total number of permutations), the total number of permutations, and thus computational time (or at number of permutations to reliably compute their large (and hence, its permutation values could play.

investigating the possible role that transformations of the test statistic relationship between the shape of the tail of extreme permutation estimation technique is under development.

For example, if 600 of the 1000 permutation values exceed the different test statistics require different numbers of permutations. Permutation tests are commonly performed in batches for large numbers of different test statistics, e.g. for all genes or all gene sets. In these permutation schemes, the same number of permutations is performed for each test statistic. This number is usually selected a priori (possibly based on an estimate of computational time or complexity). Such an approach can be highly inefficient, since different test statistics require different numbers of permutations. For example, if 600 of the 1000 permutation values exceed the test statistic, another 1000 permutations are not necessary, since the \( P \)-value can already be determined with great accuracy \( P = 0.6 \pm 0.016 \) according to (8)). However, if only one permutation value exceeds the test statistic, more permutations are necessary to accurately determine the corresponding \( P \)-value. In this work, we have shown that simple convergence criteria and confidence bounds on the estimate can be used to indicate when enough permutations have been performed to have certain statistical confidence in the \( P \)-value estimate. In most applications, only a small fraction of the test statistics will be significant, i.e. they will require a lot of permutations to reliably estimate their small \( P \)-values. The large majority of test statistics will require only a small number of permutations to reliably compute their large (and hence, insignificant) \( P \)-values. Such an approach can lead to a decrease in the total number of permutations, and thus computational time (or at least to a more sensible division of the total number of permutations), while producing more accurate \( P \)-value estimates.

In future research, we will more elaborately explore the relationship between the shape of the tail of extreme permutation values and the accuracy of the estimator. This will include investigating the possible role that transformations of the test statistic and its permutation values could play.

A web interface for the proposed permutation test \( P \)-value estimation technique is under development.