The development of novel experimental techniques is rapidly increasing the generation of data in many fields in biology, in particular in genomics with the advent of high-throughput sequencing (McPherson, 2009; Shendure and Ji, 2008). Chromatin immunoprecipitation (ChIP) technology combined with next-generation sequencing generates high-resolution data along the genome on DNA methylation, histone modifications, transcription factor binding and more (Hornet et al., 2010). The large amount of data generated by these techniques opens up for statistical studies of relations between genomic properties, both globally and locally along the genome. An example of such a local analysis is the study of how the relation between histone modifications and repeating elements varies across chromosomes (Pautler et al., 2009). A natural approach to such an investigation is to split the genome into bins along the genome, e.g. one bin per chromosome, cytoband or gene, and then perform a statistical test of a null hypothesis $H_0$ versus an alternative hypothesis $H_1$ for each bin. At the same time, due to the complex structural properties of the genome, it is often inappropriate to make simplified assumptions that would enable analytic evaluation of significance (Ewan Birney et al., 2007).

Instead, Monte Carlo (MC) sampling is often needed, resorting to computationally expensive reshuffling of genomic elements for each MC sample. As a consequence, Monte Carlo in multiple testing settings is rapidly becoming important (Sandve et al., 2010). With tests being performed for a large number of bins locally along the genome, the computational requirements may become extremely high, as the effort needed is basically the multiple of a very large number of test by a (possibly also very large) number of MC samples. Several papers have considered ways to reduce the number of samples during MC-based $P$-value computation for individual tests. Besag and Clifford (1991) propose a sequential MC algorithm for $P$-value computation, reducing the needed number of MC samples for tests that are anyway insignificant. Other papers consider alternative ideas for $P$-value estimation, such as controlling resampling risk (Gandy, 2009), and prediction of $P$-values using Random Forest models (Kustra et al., 2008). Also, there has been some work on Monte Carlo approaches for multiple testing in cases where $P$-values can be calculated analytically (Lin, 2005; Seaman and Müller-Myhsok, 2005).

In this article, we propose an algorithm that limits the total number of needed MC samples, regardless of how many tests are truly $H_0$. In Section 2, we describe our algorithm, in which MC sampling is stopped either according to the sequential MC stopping rule or when we reach a given multiple testing significance threshold. Then, in Section 3, we show on both simulated and real data that this method can lead to a drastically reduced total number of MC samples. Finally, Section 4 presents a discussion and some conclusions. Further details are provided in the accompanying Galaxy Pages (Goecks et al., 2010) document.

1 INTRODUCTION

The development of novel experimental techniques is rapidly increasing the generation of data in many fields in biology, in particular in genomics with the advent of high-throughput sequencing (McPherson, 2009; Shendure and Ji, 2008). Chromatin immunoprecipitation (ChIP) technology combined with next-generation sequencing generates high-resolution data along the genome on DNA methylation, histone modifications, transcription factor binding and more (Hornet et al., 2010). The large amount of data generated by these techniques opens up for statistical studies of relations between genomic properties, both globally and locally along the genome. An example of such a local analysis is the study of how the relation between histone modifications and repeating elements varies across chromosomes (Pautler et al., 2009). A natural approach to such an investigation is to split the genome into bins along the genome, e.g. one bin per chromosome, cytoband or gene, and then perform a statistical test of a null hypothesis $H_0$ versus an alternative hypothesis $H_1$ for each bin. At the same time, due to the complex structural properties of the genome, it is often inappropriate to make simplified assumptions that would enable analytic evaluation of significance (Ewan Birney et al., 2007).

Instead, Monte Carlo (MC) sampling is often needed, resorting to computationally expensive reshuffling of genomic elements for each MC sample. As a consequence, Monte Carlo in multiple testing settings is rapidly becoming important (Sandve et al., 2010). With tests being performed for a large number of bins locally along the genome, the computational requirements may become extremely high, as the effort needed is basically the multiple of a very large number of test by a (possibly also very large) number of MC samples. Several papers have considered ways to reduce the number of samples during MC-based $P$-value computation for individual tests. Besag and Clifford (1991) propose a sequential MC algorithm for $P$-value computation, reducing the needed number of MC samples for tests that are anyway insignificant. Other papers consider alternative ideas for $P$-value estimation, such as controlling resampling risk (Gandy, 2009), and prediction of $P$-values using Random Forest models (Kustra et al., 2008). Also, there has been some work on Monte Carlo approaches for multiple testing in cases where $P$-values can be calculated analytically (Lin, 2005; Seaman and Müller-Myhsok, 2005).

In this article, we propose an algorithm that limits the total number of needed MC samples, regardless of how many tests are truly $H_0$. In Section 2, we describe our algorithm, in which MC sampling is stopped either according to the sequential MC stopping rule or when we reach a given multiple testing significance threshold. Then, in Section 3, we show on both simulated and real data that this method can lead to a drastically reduced total number of MC samples. Finally, Section 4 presents a discussion and some conclusions. Further details are provided in the accompanying Galaxy Pages (Goecks et al., 2010) document.

2 METHODS

A commonly used multiple testing analogue to the classical $P$-value is the so-called $q$-value (Storey, 2002). The $q$-value of an individual test is defined as the minimal false discovery rate (FDR) (Benjamini and Hochberg, 1995) at which the test is called significant. $P$-values relate to $q$-values by a factor which is proportional to the number of tests for which $H_0$ is true. Denote an MC sample of the test statistic as ‘extreme’ if it is further to the $H_1$-tail of the null distribution than the observed test statistic.

When most tests are from $H_0$, the correction factor is large. Then, for an individual test to become significant, many MC samples are needed. Note, however, that in such a situation most tests would have many extreme samples (as they come from $H_0$), implying that for these tests the MC
sampling would have been stopped early if a sequential MC stopping criterion were adopted. In the opposite scenario, when most tests are from \( H_0 \), the multiple testing correction factor is small, and fewer MC samples are needed in order to obtain a significant \( q \)-value for each individual test. But then few tests have many extreme samples, and few tests would have been stopped early by sequential MC.

Consider testing \( m \) null hypotheses \( H_0_1, \ldots, H_0_m \) with corresponding test statistics \( T_1, \ldots, T_m \), where large values of \( T_i \) constitutes evidence against \( H_0_i \). Each test is performed by MC simulation: for test \( i = 1, \ldots, m \), we simulate \( n_i \) independent datasets under the null hypothesis, yielding simulated test statistics \( T_i^j \) for \( j = 1, \ldots, n_i \). Let \( k_i \) be the number of simulated test statistics that are greater than or equal to \( T_i \). The Monte Carlo \( P \)-value \( p_{\text{MC}} \) (Davidson and Hinkeyl, 1997; Phipson and Smyth, 2010) is then given by

\[
p_{\text{MC}} = \Pr(T_1 \geq T_k \mid H_0) = \frac{k+1}{n+1}
\]

since under \( H_0 \), all \( n+1 \) values \( 0, 1, \ldots, n \) are equally likely values of \( T_i \) and \( k_i+1 \) of these are greater than or equal to the observed \( t_i \).

Our concern is with choosing the number of needed MC samples, \( n_i \), for each test \( i \) (cf. the discussion in Hope (1968)). To begin, consider the case of a single test, \( m = 1 \), and assume that we observe \( k_1 = 0 \). Then \( P_{\text{MC}} = (n_1 + 1)^{-1} \), and for a given significance threshold \( \alpha \), we need \( n_1 \geq \frac{1}{\alpha} - 1 \) to have a possibility of rejecting \( H_0 \). Thus, the stricter we make the significant threshold, the more MC samples are needed. For example, for \( \alpha = 0.05 \), we only need \( n_1 \geq 19 \), but for \( \alpha = 0.001 \) we need \( n_1 \geq 999 \).

For a moderate to large number of tests, computational problems arise for two reasons: first, simply because \( m \) itself is large; second, because of the need to correct for multiple testing. As an example, assume that \( n_1 = n \) for each \( i \), and that we use Bonferroni corrected \( P \)-values to account for multiple testing, with an overall family-wise error rate of \( \alpha \). Then, each test has significance threshold \( \alpha/n \), and \( m(\frac{1}{n} + 1) \) samples are needed in total. For \( n = 1000 \) and \( \alpha = 0.05 \), this means that nearly 20 million MC samples are needed. Since each MC sample typically involves a complex re-shuffling of genomic elements, the computational cost is extremely high.

Besag and Clifford (1991) propose a sequential MC method for hypothesis testing. Their basic idea is to stop MC sampling at the point that it becomes clear that the null hypothesis will never be rejected. Assume that large values of the test statistic constitute evidence against \( H_0 \). Instead of using a fixed MC sample size, sampling is continued until either a pre-determined number \( h \) of values larger than the observed test statistic \( t_i \) (i.e. extreme MC samples) are drawn, or until some maximum number \( n \) of MC samples have been calculated. Let \( g \) be the number of values exceeding \( t_i \) when this algorithm terminates. Then, the sequential \( P \)-value \( P_{\text{seq}} \) is given by

\[
P_{\text{seq}} = \begin{cases} \frac{\min(g, h)}{\min(g, h) + \max(n - g, 0)} & (g \geq h), \\ \frac{\max(n - g, 0)}{\min(g, h) + \max(n - g, 0)} & (g < h). \end{cases}
\]

Besag and Clifford (1991) suggest setting \( h = 10 \) or 20. For the multiple testing setting, we propose to augment this procedure by adding a third stopping criterion, namely a \( q \)-value threshold \( \alpha \), aiming to run just as many samples as are needed to obtain an accept/reject decision for each test. For \( m \) tests with observed test statistics \( t_1, \ldots, t_m \), our proposed Monte Carlo False Discovery Rate (MC-FDR) algorithm is as follows:

1. Let \( A = \{1, 2, \ldots, m\} \) and \( B = \emptyset \).
2. (1) Let \( A = \{1, 2, \ldots, m\} \).
   (a) For each \( i \in A \), calculate/update \( p_i \) by sequential MC, using an additional MC sample. If a total of \( g \) samples exceeding \( t_i \) are then obtained, move \( i \) from \( A \) to \( B \).
   (b) Calculate \( q \)-values (see below) based on the current \( P \)-values \( P_1, P_2, \ldots, P_m \). If \( q_i < \alpha \) for all \( i \in A \), then move all \( i \in A \) to \( B \).

   Instead of adding only a single MC sample in Step 2a, a batch of \( N > 1 \) additional MC samples may be added.

In Step 2b above, we calculate \( q \)-values based on the current \( P \)-values. The \( q \)-values mainly depend on the proportion \( \pi_0 \) of true null hypotheses among the \( m \) tests. Let \( P_1 \leq P_2 \leq \cdots \leq P_m \) be the ordered, observed \( P \)-values. For a given estimate \( \hat{\pi}_0 \) of \( \pi_0 \), \( q_0 = q \)-value(\( P_{\text{seq}} \)) can be easily estimated as

\[
\hat{q}_{10} = \min\{n \in \mathbb{N} : \hat{\pi}_0 + \frac{P_{\text{seq}}}{n+1} < 1\}.
\]

Thus, the main issue is estimating \( \pi_0 \). Many methods for estimating \( \pi_0 \) have been proposed in the literature (Cézèole and Robin, 2010; Finner and Gontscharuk, 2009; Fiquet and Cauve, 2011; Huang, 2011; Jiang and Doerge, 2008; Langaas et al., 2005; Nettleton et al., 2006; Storey, 2002; Tamhane and Shi, 2009; Zhang, 2011). Most methods assume that \( P \)-values are continuous and uniformly distributed on \((0,1)\) under \( H_0 \). However, in the sequential MC case, \( P \)-values are discrete and uniformly distributed on the set

\[
H = \{1/h(1+1), \ldots, h/(n+1), h/n, h/(h-1)/\ldots, 1/n\}.
\]

Pounds and Cheng (2006) have proposed a very simple estimator of \( \hat{\pi}_0 \) due to Tamhane and Shi, 2009; Zhang, 2011. Instead of using a fixed number of extreme samples, the sequential MC case, and the MC-FDR scheme)

\[
Y = \{ y_1, y_2, \ldots, y_m \} = \{ P_{\text{seq}} \}.
\]

This estimator can be shown to give conservative estimates for both discrete and continuous \( P \)-values, and we have therefore chosen to use this estimator in our algorithm.

3 RESULTS

3.1 Simulation

In order to investigate the correctness and efficiency of the proposed method, we perform a simulation study. The uniform \((0,1)\) distribution is used to generate \( P \)-values from tests for which \( H_0 \) is true, and a Beta distribution with more mass on lower values is used to generate \( P \)-values under \( H_1 \), giving a Uniform-Beta mixture distribution of underlying \( P \)-values for tests (Pounds and Morris, 2003). By definition, the probability that a test statistic randomly sampled under \( H_0 \) is more extreme than the observed test statistic is given by the \( P \)-value. Let \( P_i \) be the \( P \)-value corresponding to \( t_i \). Then, instead of drawing test statistics directly, we may draw Bernoulli variables with parameter \( P_i \) as indicators of whether a randomly drawn test statistic would be more extreme than the observed test statistic. Accordingly, simulation for a single test \( i \) may be performed by drawing \( n_{\text{sim}} \) MC samples \( Y_{ij}, j = 1, \ldots, n_{\text{sim}} \) of a Bernoulli variable with parameter \( P_i \), with \( Y_{ij} = 1 \) corresponding to an extreme sample.

For standard non-sequential MC we draw a fixed number of samples for each test. For sequential MC, samples are drawn until a given number of extreme samples is observed or until a maximum number of samples is reached. For the MC-FDR scheme, samples are drawn either until a given number of extreme samples is observed or until the estimated FDR-value falls below a given threshold. The simulation procedure is summarized below:

1. Draw \( n_{\text{MC}} \) samples from the Uniform distribution, and \( m(1 - \pi_0) \) samples from the Beta \((\alpha, \beta)\) distribution.
2. (For each test \( i, i = 1, \ldots, m \))
   (a) Set \( g = 0 \)
   (b) While \( g \) is smaller than a limit specified by the stopping criterion (differently defined in the basic scheme, the sequential MC scheme, and the MC-FDR scheme)
      (1) Generate \( Y \sim \text{Bernoulli}(P_i) \)
      (2) If \( Y = 1 \), let \( g = g + 1 \)

As the main simulation, we ran 5000 tests, with \( \alpha = 0.25 \), \( \beta = 0.25 \), \( h = 20 \), a maximum of 50,000 samples for standard and sequential
For standard MC, a large, constant number of samples is needed.

At the left side of the plot, corresponding to very low P-values, tests sorted in the same order as in Figure 3a. A few tests (seen at the left side of the plot) correspond to very low P-values. Only a few of the P-values are large enough to stop early by the sequential MC criterion. However, the mild multiple testing correction (due to the relatively low \( \pi_0 \)) means that a limited number of samples is needed to reach \( q \)-values below the chosen threshold (0.1).

In order to further investigate the generalizability of the simulation results, we performed additional simulations with varying \( h \) (detailed plots are provided in the accompanying Galaxy Pages document). As expected, both the number of samples and the precision of estimated \( P \)-values increased as a function of increasing \( h \), for both sequential MC and the MCFDR scheme. Apart from this, the behavior and relation between the schemes were as for the main simulations (using \( h=20 \)). We have also tried simulations with a range of different settings for various other parameters, without observing any unexpected behavior. Thus, we are not aware of any setting for which the MCFDR scheme would fail to work as intended.

### 3.2 Real data

The regulatory role of epigenetic modifications is gaining increasing attention, to a large degree driven by the increased availability of high resolution, genome-wide data on such modifications (Barski et al., 2007). A recent study by Pekowska et al. (2010) investigated how profiles of H3K4me2-modifications in T-Cells (Wang et al., 2008) are distributed within genes. Based on a clustering of H3K4me2-profiles, five classes of genes are distinguished. A main distinction between these classes is whether H3K4me2 is localized around the transcription start site or to a larger degree spread throughout the gene body. Pekowska et al. (2010) proceed to discuss implications of these patterns for expression and tissue specificity.

The distributive aspect of H3K4me2-modifications is in Pekowska et al. (2010) mainly considered on a per class basis, where clustering allows patterns to be seen across a large number of genes. An alternative is to ask, individually for each gene, whether H3K4me2-modifications appear significantly more at the upstream end of the gene. This requires a precise hypothesis to be evaluated statistically for each gene. A natural test statistic is the average relative positioning of modifications within the gene. As histone modifications are connected to nucleosomes, which favor certain inter-spacings along DNA, the empirical inter-point length distribution should be preserved in a null hypothesis (Sandve et al., 2010). This requires a Monte Carlo based hypothesis test, where H3K4me2-modifications are permuted while preserving inter-point distances, and where the test statistic is the average relative position within a gene.

To focus on genes where there should be enough data to support conclusions to be drawn, we consider the 3466 Ensembl genes that include 10 or more histone modifications. We find that 2747 (79%) of the considered genes have significantly more H3K4me2-modifications at the upstream end of the gene, confirming that the H3K4me2-modifications preferentially localize close to the transcription start site.

For any particular gene, we may also ask the opposite question: does H3K4me2 localize preferentially at the downstream end of the gene? We find four Ensembl genes with significantly more H3K4me2 modifications downstream in the gene. Although the preferential localization downstream in these genes could represent a distinct regulatory signal targeting this gene set, a more plausible explanation is that the genes in question are overlapping other genes (or gene variants) that drive the association to H3K4me2 modifications. Manual inspection of these particular regions supports this latter explanation. Two of the genes overlap...
Fig. 2. Behavior of the sequential MC and MCFDR schemes as a function of the number of true $H_1$. (a) Number of rejected tests as a function of the number of true $H_1$. (b) Empirical FDR on test collections as a function of the number of true $H_1$.

Fig. 3. Test collection at $\pi_0=0.9$. (a) Underlying and estimated $P$-values (sorted by underlying $P$-value). The small $P$-values, mostly from $H_1$, are accurately estimated. Larger $P$-values, mainly from $H_0$, are less accurately estimated, as for sequential MC. (b) Number of samples drawn per test, as well as the number of extreme samples among these, with tests sorted in the same order as in panel (a).

Fig. 4. Test collection at $\pi_0=0.2$. (a) Underlying and estimated $P$-values (sorted by underlying $P$-value). (b) Number of samples drawn per test, as well as the number of extreme samples among these, with tests sorted in the same order as in panel (a).

with an alternative gene variant, where the methylations display a typical pattern in reference to this alternative variant. A third gene is overlapping with another (Ensembl) gene at the opposite strand (Fig. 5).

When asking whether H3K4me2-modifications appears significantly more at the upstream end of the gene, there is a high proportion of very low $P$-values. Therefore, the sequential MC threshold on number of extreme samples does not apply early...
for the majority of tests. An unlimited sequential MC run was still running after 2 weeks, having surpassed 1.5 million samples for some tests. We therefore imposed a ceiling on the maximum number of samples for sequential MC, set to 50,000 to ensure the possibility of significant results after multiple testing correction (indeed, using a maximum of 10,000 samples for sequential MC missed all significant results when asking about overrepresentation of H3K4me2 downstream in genes). Many of the tests hit the imposed ceiling on maximum number of samples, with a total number of >30 million samples across tests. In the MCFDR scheme, using a FDR-threshold of 0.1, the total number of samples was <350,000, almost a factor 100 less than with sequential MC (running time were for MCFDR <5 min, as compared to >9 h for sequential MC). As a large proportion of the null hypotheses end rejected, the number of samples needed to reach \( q \)-values below the chosen threshold is limited.

When investigating whether H3K4me2 modifications are preferentially located downstream in genes, there is a very low proportion of rejected tests. Most of the tests stop early by the sequential MC threshold, making the two schemes behave very similarly. The few rejected tests are subject to a very strong multiple testing correction, and as discussed above needs >10,000 samples to at all allow any test to beat the FDR threshold. When applying a ceiling of 50,000 samples to both schemes, the total number of samples are essentially similar. If the ceiling is removed, the total number of samples increases only slightly for MCFDR, while it increases substantially for sequential MC (surpassing one million samples for some of the tests). Details are provided in the Galaxy Pages document.

4 DISCUSSION AND CONCLUSIONS

We have provided a simple and efficient method for MC-based multiple testing. The method is freely available as part of the Genomic HyperBrowser, an open source, generic web-based system for statistical analysis of genomic annotation data. The method has been shown to work well on simulated data, and also to be highly useful for a realistic example, with computation times reduced by a factor of nearly one hundred.

MC-based hypothesis testing is often needed for genomic data. In our example of H3K4me2-modifications, the simplifying assumptions that would be needed to do analytic tests would be highly unreasonable. As discussed in Ewan Birney et al. (2007), assuming Poisson distributed positions of H3K4me2-modifications (as would be needed for an analytic test) gives an unrealistically small variance of the null distribution, and hence leads to false positives. Indeed, the analytic version of the test gave 112 significant findings for the downstream positioning test, as opposed to only four findings when preserving inter-point distances in the MC version.

Increasingly, we see applications where the calculation of each single MC sample is quite computationally expensive, and where the problem is further compounded by the need to do thousands of hypothesis tests (Ewan Birney et al., 2007; Sandve et al., 2010). In general, we must consider both statistical and computational efficiency. The FDR was introduced with the aim to improve statistical efficiency (as compared with e.g. Bonferroni correction): reject as many null hypotheses as possible, while controlling a reasonable error rate. As we have shown, taking the FDR into account during MC sampling can also greatly improve computational efficiency when we need to do MC-based multiple tests.

As shown by both our simulation study and our real data example, our method is particularly useful in the case where many tests are truly \( H_1 \), while still giving correct results if few or no tests are truly \( H_1 \). Much work on multiple hypothesis testing and FDR has (implicitly or explicitly) assumed that the proportion of true null hypotheses is close to one (Efron, 2004). While this may be a natural assumption in the oft-studied case of testing for differential gene expression, we see no reason why it should be made in general. In fact, our study of H3K4me2-modifications provides an example
where the assumption is clearly wrong. The case with relatively few true null hypotheses leads to the largest computational burden using current algorithms such as sequential MC. We speculate that such cases will become increasingly common in the future, with the advent of more elaborate study designs and research questions—in particular when doing local analysis of a generic question, such as investigating a specific relation between genomic features in bins along the genome.

Our method assumes that the user is mainly interested in an FDR-based accept/reject decision for each test, and aims to stop MC sampling when further sampling would not change the decision. In such a setting, the concept of resampling risk due to Gandy (2009) is relevant. Attempting to control resampling risk for \( q \)-values would be an interesting undertaking, though not straightforward as it would require taking uncertainty (variance) of the FDR into account (Owen, 2005). Kustra et al. (2008) aims to improve computational efficiency of \( P \)-value estimation in specific MC settings. Although beyond the scope of the present article, it would be interesting to study the adoption of our FDR-based stopping criterion also to the methods of Gandy (2009) and Kustra et al. (2008).

We have here considered the MC-based \( P \)-values as the values of direct interest, as these indeed satisfy the criteria for valid \( P \)-values (Davison and Hinkley, 1997; Phipson and Smyth, 2010). An alternative view is to think of the MC computed \( P \)-values as estimates of an underlying true \( P \)-value. Then, the MC \( P \)-value estimate follows a binomial distribution around the underlying \( P \)-value (North et al., 2002). As new samples are added, this estimated \( P \)-value will change, although it will still be highly dependent on the previous estimate. A possible variation of our MCFDR algorithm would be to stop sampling individually as each test reaches the specified \( q \)-value threshold: at Step 2b in the MCFDR algorithm as described in Section 2, for each \( i \in A \), move \( i \) from \( A \) to \( B \) if \( q_i < a \). Although even less computationally demanding, this could introduce a bias toward stopping sampling at estimates lower than the underlying \( P \)-value. The reason for this is the tendency to stop sampling at ‘opportune’ times, when the estimate happens to be at left-hand side of the binomial distribution around the underlying \( P \)-value. This is less likely to happen when using the global criterion, as it would need to happen for several estimates simultaneously. Simulations (described in the Galaxy Pages document) confirm this empirically.

Both simulations and applications of the MCFDR algorithm is available through the Genomic HyperBrowser. A simple web tool allows anyone to run simulations at different parameter settings, providing detailed inspection of the properties of the algorithm. MCFDR is integrated into the main analysis engine of the Genomic HyperBrowser, allowing anyone to make use of the algorithm for analyses on their own data, or reproduce our biological findings (see the Galaxy Pages document referred to in the abstract). Furthermore, due to the inherent simplicity of the algorithm, it is easy to apply to any computational investigation that involves MC and multiple testing.

Although modern technologies for data generation and computation is neither a necessity for MC estimation (Hammersley and Morton, 1954) nor for multiple testing (Schweder and Spjotvoll, 1982), it seems clear that their adoption has been driven by increased computer power and data generation technologies such as microarrays. In the same way, although the ideas here presented on MC in multiple testing settings are general, we believe their relevance will increase strongly along with the future developments in e.g. next-generation sequencing, making them an important part of a bioinformatics toolbox in the future.

ACKNOWLEDGEMENTS

We thank Arnaldo Frigessi for very helpful comments and discussions.

Funding: EMBIO at University of Oslo (to G.K.S.); ‘Statistics for Innovation (sifi)’ Norwegian Centre for Research-Based Innovation (to E.F.); Norwegian Functional Genomics program (FUGE) (to S.N.).

Conflict of Interest: none declared.

REFERENCES


