BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 23 2011, pages 3235–3241
doi:10.1093/bioinformatics/btr568

Genome analysis

Advance Access publication October 13, 2011

Sequential Monte Carlo multiple testing

Geir Kjetil Sandve[1,*,†], Egil Ferkingstad[2,†] and Ståle Nygård[3]

[1]Department of Informatics, University of Oslo, [2]Statistics for Innovation, Norwegian Computing Center and [3]Bioinformatics Core Facility, Institute of Medical Informatics, University of Oslo, Oslo University Hospital, Oslo, Norway

Associate Editor: Alex Bateman

ABSTRACT

Motivation: In molecular biology, as in many other scientific fields, the scale of analyses is ever increasing. Often, complex Monte Carlo simulation is required, sometimes within a large-scale multiple testing setting. The resulting computational costs may be prohibitively high.

Results: We here present MCFDR, a simple, novel algorithm for false discovery rate (FDR) modulated sequential Monte Carlo (MC) multiple hypothesis testing. The algorithm iterates between adding MC samples across tests and calculating intermediate FDR values for the collection of tests. MC sampling is stopped either by sequential MC or based on a threshold on FDR. An essential property of the algorithm is that it limits the total number of MC samples whatever the number of true null hypotheses. We show on both real and simulated data that the proposed algorithm provides large gains in computational efficiency.

Availability: MCFDR is implemented in the Genomic HyperBrowser (http://hyperbrowser.uio.no/mcfdr), a web-based system for genome analysis. All input data and results are available and can be reproduced through a Galaxy Pages document at: http://hyperbrowser.uio.no/mcfdr/u/sandve/p/mcfdr.

Contact: geirksa@ifi.uio.no

Received on July 1, 2011; revised on September 8, 2011; accepted on October 6, 2011

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 (Horner *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 (Pauler *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 $H0$ versus an alternative hypothesis $H1$ 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 $H0$. 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 $H0$ is true. Denote an MC sample of the test statistic as 'extreme' if it is further to the $H1$-tail of the null distribution than the observed test statistic.

When most tests are from $H0$, 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 $H0$), implying that for these tests the MC

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

sampling would have been stopped early if a sequential MC stopping criterion were adopted. In the opposite scenario, when most tests are from $H1$, 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 $H_{01},\ldots,H_{0m}$ with corresponding test statistics $T_i=t_i$, $i=1,\ldots,m$, where large values of t_i constitutes evidence against H_{0i}. 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^*_{ij} 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_{mc} (Davison and Hinkley, 1997; Phipson and Smyth, 2010) is then given by

$$P_{mc}=\Pr(T_i\geq t_i|H_{0i})=\frac{k_i+1}{n_i+1},$$

since under H_{0i}, all n_i+1 values $t_i,t^*_{i1},\ldots,t^*_{in_i}$ 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 $P_{mc}=(n_i+1)^{-1}$, and for a given significance threshold α, we need $n_i\geq\frac{1}{\alpha}-1$ 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 $n_i\geq 19$, but for $\alpha=0.001$ we need $n_i\geq 999$.

For a moderate to large number m 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 $n_i=n$ for each i, and that we use Bonferroni corrected P-values to account for multiple testing, with an overall family-wise error rate of α. Then, each test has significance threshold α/m, and $m\left(\frac{m}{\alpha}-1\right)$ samples are needed in total. For $m=1000$ and $\alpha=0.05$, this means that nearly 20 million MC samples are needed. Since each MC sample typically involves a complex reshuffling 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 $H0$. 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 according to our definition) has been obtained (at MC sample size l, say), or until some maximum number n of MC samples have been calculated. Let g be number of values exceeding t_i when this algorithm terminates. Then, the sequential P-value P_{smc} is given by

$$P_{smc}=\begin{cases}h/l & (g=h),\\ (g+1)/(n+1) & (g<h).\end{cases}$$

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 α, 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 $t_1,t_2,\ldots,t_m$, our proposed Monte Carlo False Discovery Rate (MCFDR) algorithm is as follows:

(1) Let $A=\{1,2,\ldots,m\}$ and $B=\emptyset$

(2) Until $A=\emptyset$:

  (a) For each $i\in A$, calculate/update p_i by sequential MC, using an additional MC sample. If a total of h samples exceeding t_i are then obtained, move i from A to B.

  (b) Calculate q-values (see below) $q_1,q_2,\ldots,q_m$ 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 π_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 π_0, $q_{(i)}=q\text{-value}(P_{(i)})$ can be easily estimated as

$$\hat{q}_{(i)}=\min_{i\leq j\leq m} m*\hat{\pi}_0*P_{(j)}/j.$$

Thus, the main issue is estimating π_0. Many methods for estimating π_0 have been proposed in the literature (Celisse and Robin, 2010; Finner and Gontscharuk, 2009; Friguet and Causeur, 2011; Hwang, 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

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

Pounds and Cheng (2006) have proposed a very simple estimator of π_0: $\hat{\pi}_0=\min(1,\frac{2}{m}\sum_{i=1}^{m}p_i)$. 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 Simulations

In order to investigate the correctness and efficiency of the proposed scheme, we perform a simulation study. The uniform (0,1) distribution is used to generate P-values from tests for which $H0$ is true, and a Beta distribution with more mass on lower values is used to generate P-values under $H1$, 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 test 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_{sim} MC samples Y_{ij}, $j=1,\ldots,n_{sim}$ 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 MCFDR 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 $m\pi_0$ samples from the Uniform distribution, and $m(1-\pi_0)$ samples from the Beta (α,β) 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 $Y\sim$ Bernoulli (p_i)

    (2) If $Y=1$, let $g=g+1$

As the main simulation, we ran 5000 tests, with $\alpha=0.25,\beta=25$, $h=20$, a maximum of 50 000 samples for standard and sequential

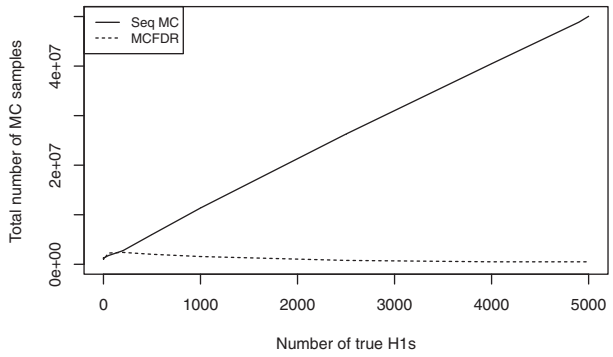

Fig. 1. Total number of samples for sequential MC and MCFDR, respectively, as a function of the number of true $H1$. When the proportion of true $H1$ is low, most tests are stopped by the sequential MC criterion, resulting in a similar number of samples for both schemes. At larger proportions of true $H1$, the multiple testing correction becomes milder, and thus fewer samples are needed to reach the FDR threshold. Thus, for the MCFDR scheme the total number of needed samples decreases with higher true $H1$ proportions. In contrast, for the sequential MC scheme, the number of needed samples increases linearly with increasing proportion of true $H1$. For standard MC, a large, constant number of samples is needed.

MC, and with π_0 varying between 0 and 1. The maximum number of samples was chosen to ensure the possibility of significant results after multiple testing correction. Figure 1 shows the resulting total number of samples for sequential MC and MCFDR, respectively, as a function of the number of true $H1$. Figure 2a shows that the number of rejected tests, as a function of the number of true $H1$, is very similar across all three schemes. When the number of true $H1$ is low, stronger multiple testing correction is needed, leading to lower power and under-rejection. When the number of true $H1$ increases, more rejections can be made, while still controlling the FDR. For example, when 4500 of the 5000 hypotheses are truly from $H1$, rejecting all 5000 null hypotheses would give an FDR of exactly $(5000-4500)/5000=0.1$. Note that these are general features of the FDR; this behavior is not specific to our approach. Figure 2b shows that the empirical FDR is also very similar across schemes, and very close to the chosen FDR threshold.

In order to inspect the behavior of the schemes more closely, it is necessary to look at the number of samples and estimated P-values at individual settings of true $H1$. Figure 3a shows underlying and estimated P-values in a collection where 10% of tests come from $H1$. Figure 3b shows the number of needed samples per test, with tests sorted in the same order as in Figure 3a. A few tests (seen at the left side of the plot) corresponding to very low P-values, need a large number of samples to become significant since they are subject to strong multiple testing correction (due to the relatively high π_0). However, the remaining (majority of) tests are stopped early by the sequential MC threshold. The other end of the spectrum, with a large proportion of true $H1$, is shown in Figure 4a and Figure 4b. Figure 4a shows underlying and estimated P-values in a simulation where 80% of tests come from $H1$. Here, most P-values are small, and estimated with reasonable accuracy. Figure 4b shows the corresponding number of needed samples. 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 π_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) investigates 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 appears 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

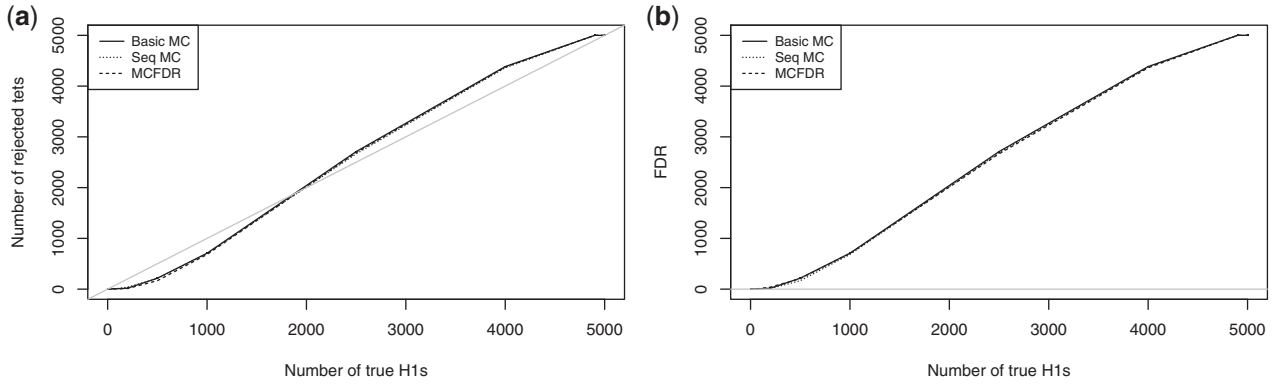

Fig. 2. Behavior of the sequential MC and MCFDR schemes as a function of the number of true $H1$. (**a**) Number of rejected tests as a function of the number of true $H1$. (**b**) Empirical FDR on test collections as a function of the number of true $H1$.

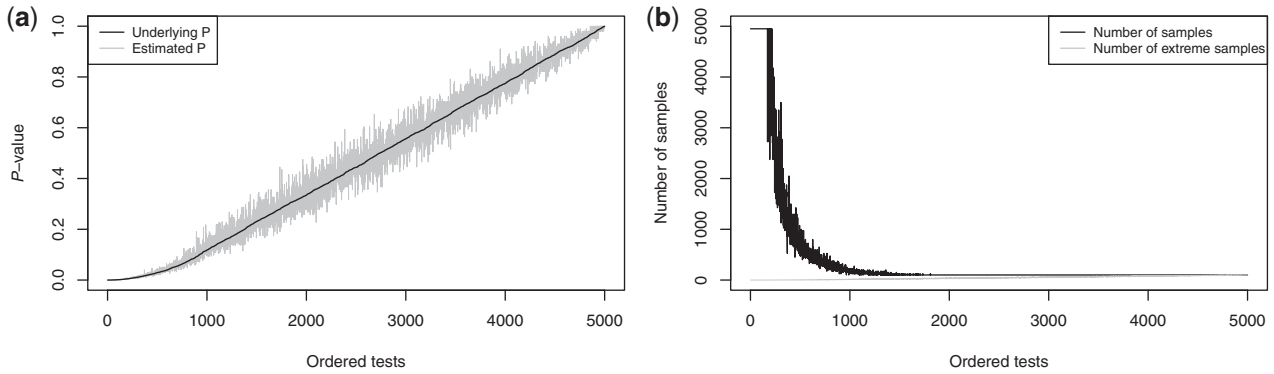

Fig. 3. Test collection at $\pi_0 = 0.9$. (**a**) Underlying and estimated P-values (sorted by underlying P-value). The small P-values, mostly from $H1$, are accurately estimated. Larger P-values, mainly from $H0$, are less accurately estimated, as for sequential MC. (**b**) Number of samples drawn per test, as well as the number of extreme samples among these, with tests sorted in the same order as in panel (a).

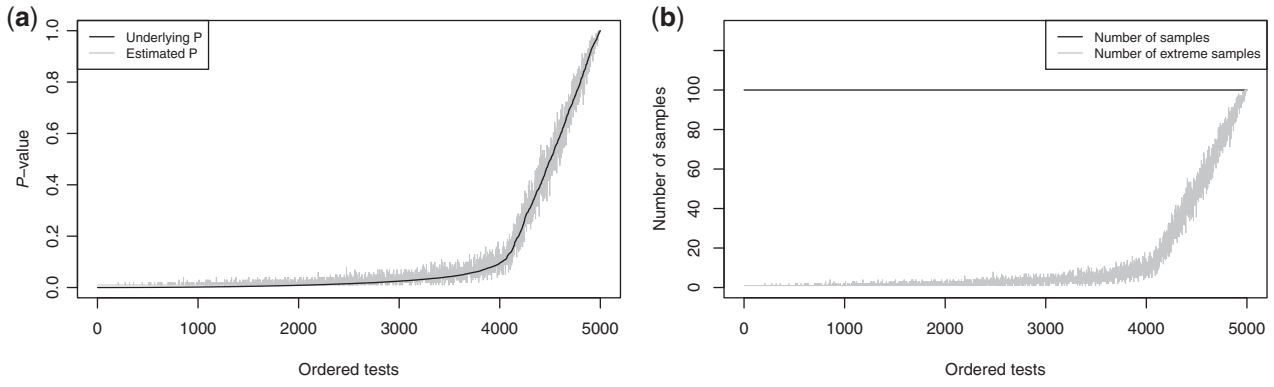

Fig. 4. Test collection at $\pi_0 = 0.2$. (**a**) Underlying and estimated P-values (sorted by underlying P-value). (**b**) Number of samples drawn per test, as well as the number of extreme samples among these, with tests sorted in the same order as in panel (a).

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

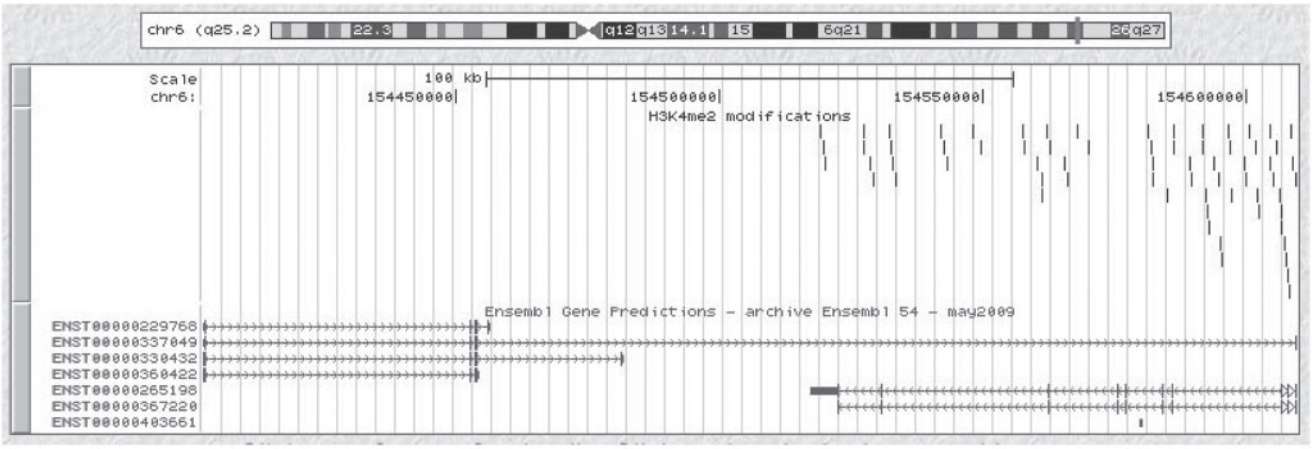

Fig. 5. H3K4me2 modifications and Ensembl genes occurring in gene region corresponding to Ensembl gene ID ENSG00000112038 (chr6:154,402,136-154,609,693), visualized by the UCSC Genome Browser. In reference to the above-mentioned gene (corresponding to Ensembl transcript ID ENST00000337049 in the figure), H3K4me2 modifications occur significantly more downstream in the gene. However, in reference to the gene corresponding to ID ENST00000367220, which is shorter and at the opposite strand, the H3K4me2 modifications are preferentially located close to TSS, occur gradually less frequently throughout the gene body and stop appearing after the gene body.

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 up 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 $H1$, while still giving correct results if few or no tests are truly $H1$. 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 $q_i < \alpha$. 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 Spjøtvoll, 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 bioinformaticians toolbox in the future.

ACKNOWLEDGEMENTS

We thank Arnoldo Frigessi for very helpful comments and discussions.

Funding: EMBIO at University of Oslo (to G.K.S.); 'Statistics for Innovation (sfi)^2' Norwegian Centre for Research-Based Innovation (to E.F.); Norwegian Functional Genomics program (FUGE) (to S.N.).

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* , **57**, 289–300.

Besag,J. and Clifford,P. (1991) Sequential Monte Carlo p-values. *Biometrika*, **78**, 301.

Celisse,A. and Robin,S. (2010) A cross-validation based estimation of the proportion of true null hypotheses. *J. Stat. Plan. Inf.*, **140**, 3132–3147.

Davison,A. and Hinkley,D. (1997) *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.*, **99**, 96–104.

Ewan Birney,J. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, **447**, 799–816.

Finner,H. and Gontscharuk,V. (2009) Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *J. R. Stat. Soc. Ser. B*, **71**, 1031–1048.

Friguet,C. and Causeur,D. (2011) Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Comput. Stat. Data Anal.*, **55**, 2665–2676.

Gandy,A. (2009) Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *J. Am. Stat. Assoc.*, **104**, 1504–1511.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Hammersley,J. and Morton,K. (1954) Poor man's Monte Carlo. *J. R. Stat. Soc. Ser. B*, **16**, 23–38.

Hope,A. (1968) A simplified Monte Carlo significance test procedure. *J. R. Stat. Soc. Ser. B* , **30**, 582–598.

Horner,D.S. *et al.* (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief. Bioinformatics*, **11**, 181–197.

Hwang,Y. (2011) Comparisons of estimators of the number of true null hypotheses and adaptive FDR procedures in multiplicity testing. *J. Stat. Comput. Simul.*, **81**, 207–220.

Jiang,H. and Doerge,R. (2008) Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Informat.*, **6**, 25–32.

Kustra,R. *et al.* (2008) Efficient p-value estimation in massively parallel testing problems. *Biostatistics*, **9**, 601.

Langaas,M. *et al.* (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B*, **67**, 555–572.

Lin,D. (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, **21**, 781–787.

McPherson,J. (2009) Next-generation gap. *Nat. Methods*, **6**, S2–S5.

Nettleton,D. *et al.* (2006) Estimating the number of true null hypotheses from a histogram of p values. *J. Agri. Biol. Environ. Stat.*, **11**, 337–356.

North,B. *et al.* (2002) A note on the calculation of empirical p values from Monte Carlo procedures. *Am. J. Hum. Genet.*, **71**, 439.

Owen,A. (2005) Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B*, **67**, 411–426.

Pauler,F.M. *et al.* (2009) H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Research*, **19**, 221–233.

Pekowska,A. *et al.* (2010) A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.*, **20**, 1493–1502.

Phipson,B. and Smyth,G. (2010) Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, **9**, 39.

Pounds,S. and Cheng,C. (2006) Robust estimation of the false discovery rate. *Bioinformatics*, **22**, 1979–1987.

Pounds,S. and Morris,S. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Sandve,G.K. *et al.* (2010) The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol.*, **11**, R121.

Schweder,T. and Spjøtvoll,E. (1982) Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.

Seaman,S. and Müller-Myhsok,B. (2005) Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.*, **76**, 399–408.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Storey,J. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* , **64**, 479–498.

Tamhane,A.C. and Shi,J. (2009) Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of FDR. *Lect. Notes Monograph Ser.*, **57**, 304–325.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Zhang,S. (2011) Towards accurate estimation of the proportion of true null hypotheses in multiple testing. *PLoS One*, **6**, e18874.