OPEN ACCESS Freely available online

PLOS ONE

A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies

Emanuele Raineri*, Marc Dabad, Simon Heath

Statistical Genomics, Centro Nacional de Análisis Genómico, Barcelona, Catalonia, Spain

Abstract

We apply a known algorithm for computing exactly inequalities between Beta distributions to assess whether a given position in a genome is differentially methylated across samples. We discuss the advantages brought by the adoption of this solution with respect to two approximations (Fisher's test and Z score). The same formalism presented here can be applied in a similar way to variant calling.

Citation: Raineri E, Dabad M, Heath S (2014) A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies. PLoS ONE 9(5): e97349. doi:10.1371/journal.pone.0097349

Editor: Dajun Deng, Peking University Cancer Hospital and Institute, China

Received December 29, 2013; **Accepted** April 17, 2014; **Published** May 13, 2014

Funding: The authors would like to acknowledge the support provided by the Spanish Ministerio de Ciencia e Innovación (grant SAF2011-30391). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: emanuele.raineri@gmail.com

Introduction

Average DNA methylation at a locus can be measured by Whole Genome Bisulfite Sequencing (WGBS), which determines the fraction of DNA strands methylated at any given genomic position in a population of cells (this definition is likely to sound obvious to those who already know about WGBS and too terse to those who don't: a good introduction to this kind of measurements is contained in chapter 11 of [1]). In what follows we will call this fraction θ; when we distinguish between different samples we will write θ_1 and θ_2. WGBS experiments estimate this numbers by measuring the methylation state of a random (*i.e.* selected in some unpredictable way) set of reads sequenced from the sample. Since one can only analyze a finite number of reads per sample the value of θ will be known only up to some variability.

In this paper we propose an answer to the basic question : how does one assess whether two cell populations have different methylation levels at a genomic position? Researchers in the field have already dealt with this issue in a variety of ways: for example [2] uses a Fisher's test. In [3] Sun et al. compute a confidence interval for $\theta_1-\theta_2$ starting from some reasonable choice of a probabilistic model. Bsmooth [4] (which tackles the slightly different problem of defining differently methylated *regions* as opposed to positions) ultimately relies on a t-test. The authors of [5] use a hierarchical model to estimate the parameters needed for a Gaussian hypothesis test. Here we would like to bring attention to another possible approach, based on properties of the Beta distributions which are explained in [6], [7]. Similarly to *e.g.* [3] we do not test an hypothesis and output a p-value; rather we compute the probability distribution of the parameter of a Bayesian model.

Beta Distribution to Model Methylation Probabilities

The Beta probability distribution (over θ) with parameters a,b is defined by

$$Beta_\theta(a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

where B is the Beta function

$$B(a,b)=\frac{(a-1)!(b-1)!}{(a+b-1)!}$$

Beta_θ appears very naturally in many studies of genomic data: typically such analyses also entail the comparison between different samples, which in turn means that different Betas have to be combined. Here for concreteness we are describing the case of measuring DNA methylation differences across samples via whole genome bisulfite sequencing but the same concepts apply with almost no change to genotyping.

To appreciate how this variability can be quantified, consider a set of reads out of a WGBS experiment covering a certain genomic coordinate x with read depth d. Since not all the strands in the sample being sequenced will, in general, have the same bases methylated at the same time, this will be a collection of heterogeneous reads : some will indicate methylation at position x (these are the so called *non converted* reads), others (the *converted* reads) will correspond to molecules that are not methylated. Now, if θ were known a priori, the probability of obtaining n non converted reads would be given by a binomial distribution (which is closely related to Beta_θ):

$$P(n|\theta)=\binom{d}{n}\theta^n(1-\theta)^{(d-n)}=\frac{1}{d+1}\,\mathrm{Beta}_\theta(n+1,d-n+1)$$

If one assumes a uniform prior on θ, ($P(\theta)=1$, $\forall\theta\in[0,1]$) the expression for $P(\theta|n)$ is very similar (The factor $\frac{1}{d+1}$ cancels out when applying Bayes' theorem)

$$P(\theta|n)=\mathrm{Beta}_\theta(n+1,d-n+1)$$

Therefore, to assess whether a position is differentially methylated across two samples with non converted reads respectively n_1,n_2 and read depths d_1,d_2 one has to compute

$$P(\theta_1>\theta_2)$$

where

$$\theta_1\sim\mathrm{Beta}_\theta(n_1+1,d_1-n_1+1),\theta_2\sim\mathrm{Beta}_\theta(n_2+1,d_2-n_2+1)) \quad (1)$$

The purpose of the software we will discuss in this note is to estimate $P(\theta_1>\theta_2)$ given the result of a WGBS experiment.

Exact Computation of Beta Differences

A method for computing

$$P(\theta_1>\theta_2)$$

which turns out to be efficient enough for our purposes is presented in full detail in [6], [7]. We will summarize its derivation here for the sake of completeness, and advise interested readers to study those papers for a more detailed discussion. We start with some preliminary definitions: let $g(a,b,c,d)\equiv P(\theta_1>\theta_2)$ where θ_1 and θ_2 are distributed respectively as $\mathrm{Beta}_\theta(a,b)$ and $\mathrm{Beta}_\theta(c,d)$. Besides, we will use the notation $I_\theta(a,b)$ for the cumulative distribution function of the Beta distribution (also known as the incomplete Beta distribution).

Now, by definition one has

$$P(\theta_1>\theta_2)=g(a,b,c,d)=\int_{-\infty}^{+\infty}\mathrm{Beta}_\theta(a,b)I_\theta(c,d)d\theta$$

But then, using the identity ([8])

$$I_\theta(c,d)=\frac{1}{cB(c,d)}\theta^c(1-\theta)^d+I_\theta(c+1,d)$$

one finds that

$$g(a,b,c,d)=\frac{1}{c}h+g(a,b,c+1,d) \quad (2)$$

where

$$h=\frac{B(a+c,b+d)}{B(a,b)B(c,d)}$$

Furthermore, one can prove that $g(a,b,c,d)$ possesses a number of symmetries. An obvious one is $g(a,b,c,d)=1-g(c,d,a,b)$. Also true are

$$\begin{aligned}g(a,b,c,d)&=g(d,c,b,a)\\ g(a,b,c,d)&=g(d,b,c,a)\end{aligned} \quad (3)$$

Using (2) and (3) one can design a nice recursive scheme

$$\begin{aligned}g(a+1,b,c,d)&=g(a,b,c,d)+h(a,b,c,d)/a\\ g(a,b+1,c,d)&=g(a,b,c,d)-h(a,b,c,d)/b\\ g(a,b,c+1,d)&=g(a,b,c,d)-h(a,b,c,d)/c\\ g(a,b,c,d+1)&=g(a,b,c,d)+h(a,b,c,d)/d\end{aligned}$$

where the base case is provided by $g(a,b,a,b)=\frac{1}{2}$ (this because if θ_1 and θ_2 have exactly the same distribution, $P(\theta_1>\theta_2)=\frac{1}{2}$).

Approximate Computation

Even if methylation data are well modelled by a Beta_θ, the comparison presented above is never (to our knowledge) used in the literature. As (hopefully fair) representatives of the methods which we have found are used instead, we will analyze the performances of the Fisher's test and that of a test based on a Gaussian approximation.

To do a Fisher's test, one builds a contingency table with the number of non converted and converted reads in the two samples (note that this kind of test breaks down when one of the rows (or columns) of the contingency table is zero). In the Gaussian approximation, one models $P(\theta)$ for each sample with a Gaussian with the same mean and variance of Beta_θ; and then uses the two Gaussians to test for differences between θ_1 and θ_2. In both cases we will consider one tailed tests.

Results and Discussion

Comparison with Approximate Results

We organized the comparison between the exact and approximate solution in two steps. First, we looked at the behaviour of the two tests on a pair of real samples (see below for instructions on how to access the data we used).

The results are shown in figure 1. On the x axis we plotted $P(\theta_1>\theta_2)$, on the y axis we plotted the corresponding p-value obtained by approximating the Beta respectively with a Fisher's test (on the left) and with a Gaussian (on the right). We did the comparison over 100000 positions : the plot is in fact a two dimensional histogram, in which different shades of blue indicate how many times the two values fall into a certain region of the plane. There is not much to comment there except to note that, as expected, there is a broad correspondence between the different methods. Also, at such a scale the Beta probabilities seem more similar to the Z score test than to the Fisher's p-values (the right hand side plot looks more like a diagonal).

Next, we simulated a pair of samples whose counts are generated by the same underlying binomial process (*i.e.* $\theta_1=\theta_2=0.5$) at different coverages. These constitute a negative control, in the sense that none of the methods should report a significant difference between the samples. Furthermore, we generated a pair of samples such that their underlying binomial probabilities are markedly different $\theta_1=0.9,\theta_2=0.5$; those are the true positives, *i.e.* cases for which the tests should detect that $\theta_1>\theta_2$. We then compare the receiver operating characteristic

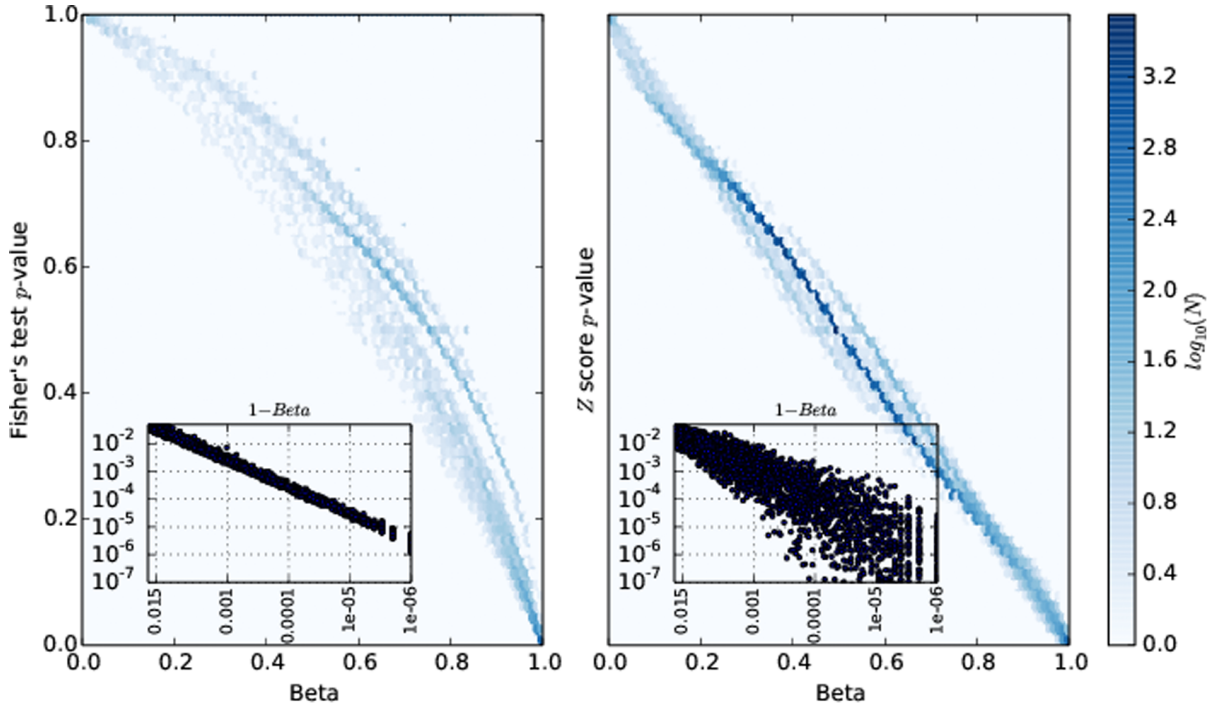

Figure 1. Comparing beta distribution with Fisher's test and Z score test. Each plot contains an enlarged version around p-value ~ 0.05. Notice that the in these magnified plots the x axis is $1-\text{Beta}_\theta$, for exact powers of 10 take less space in the labels then string of 9 s.
doi:10.1371/journal.pone.0097349.g001

(ROC) curves of the three methods for different values of the samples' coverages, d_1,d_2. The results are depicted in figure 2. That plot justifies the usage of the Beta_θ distribution: the number of false negatives accumulated by the other two methods considered stops them from reaching an high enough true positive rate (even when the threshold for computing it is very permissive). Note, for example, that the blue line is not even visible in the leftmost panels of figure 2. This effect is also shown in figure 3 where we depict the distribution of the outputs for the three methods at read depth $d_1 = 10, d_2 = 10$.

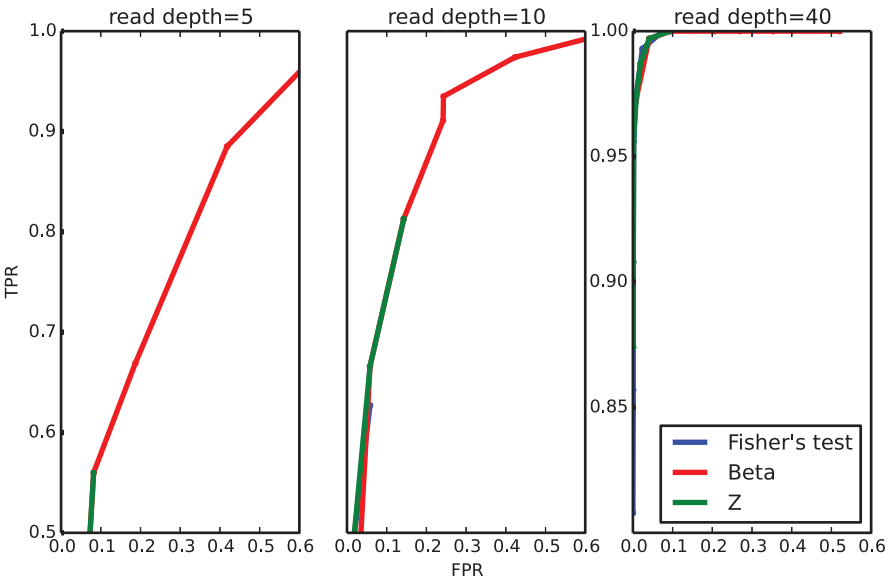

Figure 2. ROC curves for the three methods under comparison. Each point in the ROC curve is obtained by choosing a different threshold for calling differential methylation. For the Z score test and the Fisher's test the p-values are: 10^{-1},5×10^{-2},2×10^{-2},10^{-2},5×10^{-3},2×10^{-3},10^{-3}. For the Beta distributions the threshold probabilities are: 0.5,0.6,0.7,0.8,0.9,0.95,0.99. TPR means true positive rate; FPR means false positive rate.
doi:10.1371/journal.pone.0097349.g002

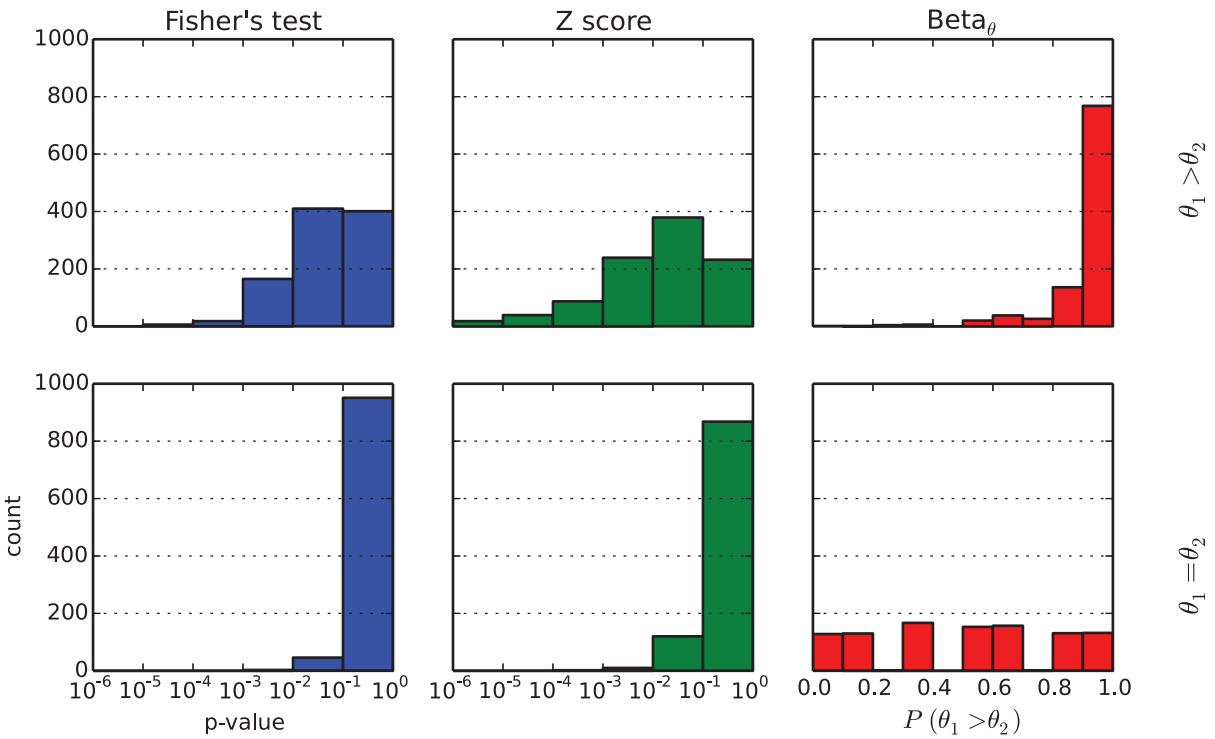

Figure 3. Distribution of p-values (for the hypothesis tests discussed) and of $P(\theta_1 > \theta_2)$ computed with the Beta$_\theta$ model. The first row depicts the truly different samples ($\theta_1 = 0.9, \theta_2 = 0.5$). The bottom row refers to the control samples. For all the plots $d_1 = 10, d_2 = 10$.
doi:10.1371/journal.pone.0097349.g003

Differentially Methylated Regions and Effects of Coverage

Using the above concepts, we can compute differentially methylated regions (DMR) along the genome : these are uninterrupted blocks of nucleotides where the two samples have different methylation. One possible technique to find such blocks is to conjoin a number of adjacent nucleotides in a DMR, disregarding their exact methylation probabilities, and to assign hard boundaries. This usually implies that a number of *ad hoc* rules must be established to control the minimum distance between 2 neighbouring DMRs, the minimum length of a DMR, how to exactly count the intersection of DMRs with annotated regions, and so on and so forth. Using our method, though, one can simply assign to each nucleotide the probability computed by the algorithm presented here; any further analysis can be conducted without imposing arbitrary threshold or boundaries. For instance one can ask what is the average value of this probability over some specific regions (introns, enhancers) with respect to randomly chosen regions of the genome. Often it is not clear a priori what is the correct scale to use when looking at methylation : if this is the case, one can smooth the probability per nucleotide by computing a kernel density estimation at various bandwidths, or simply clump together the values of a number of nearby bases in a single (average) value. Note that smoothing is justified by the fact that methylation levels are correlated in space (the strength and persistence of the correlation is different from sample to sample, reflecting technical and biological variability); in fact as hinted at in [4], analyzing together nearby positions could provide a way of correcting measurement errors.

We would also like to comment on the fact that the different coverage of the samples does have an effect on the estimation of differential methylation. The main idea to understand here is that low coverage means uncertainty: and uncertainty can give rise to results which, while correct, are slightly counterintuitive. For example in figure 4 we show that a sample with low methylation and low coverage can be (maybe, one cannot say for sure) more methylated than a sample with high, certain methylation. The right panel of the same figure suggests that a good way of filtering for certainty is to select positions with low estimated variability (rather than to select based on read counts): this is because the same read depth can correspond to different variances depending on how many reads are non converted or converted.

Finally, once one has the estimates for θ_1 and θ_2 (as obtained via the ratio of unconverted reads over the coverage) and $P(\theta_1 - \theta_2 > 0)$ (*i.e.* the output of the algorithm expalined in this paper) one can take an informed decision on a locus, keeping into account both the size of the difference in methylation and its variability.

Implementation and Data Availability

The algorithm described above is implemented in a C program, called methyl_diff, available from the Github page of one of the authors : http://emanueleraineri.github.io/. The program takes as input (from stdin) four integers, *i.e.* the number of non converted and converted reads for the first and the second sample respectively, and prints $P(\theta_1 > \theta_2)$ on the stdout. It takes $3.3s$ to process 10^5 lines on off-the-shelf hardware (MacBookPro with Intel i7@2.66 GHz). Note that the data used to produce figure 1 are publicly available (they were generated for BLUEPRINT, a consortium, studying epigenetic marks in immune system cells.) in at least two ways (also corresponding to two different formats):

1. First of all, they can be downloaded from the same web page where the source code of our implementation is stored. The file G199.G202 contains the methylation levels of 100000 random positions from the chromosome 1 of samples G199 and G202

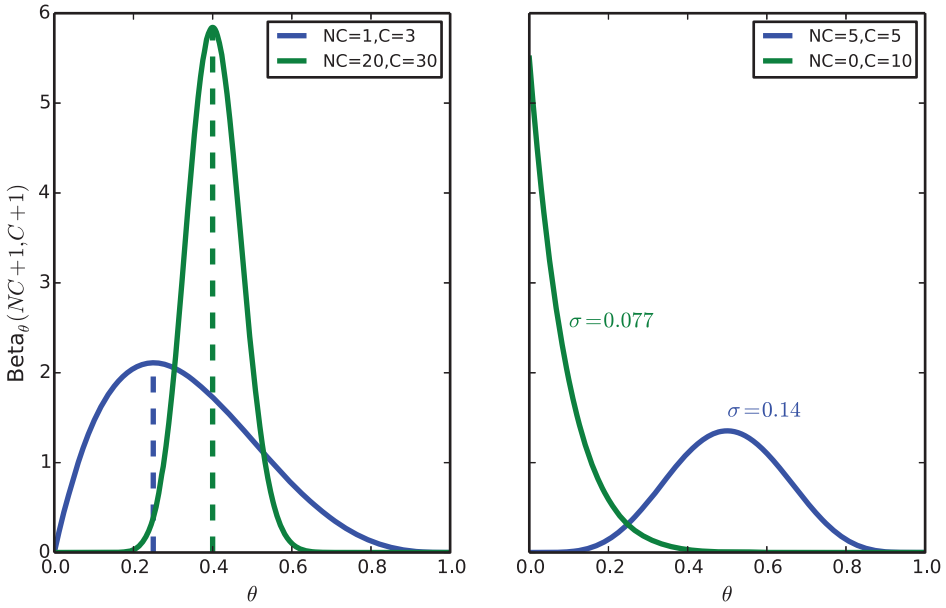

Figure 4. Effects of coverage on $P(\theta_1 > \theta_2)$. In the left panel we show that a sample for which the methylation is estimated (with high uncertainty) to be low can be (with some probability) more methylated than a sample for which the methylation level is higher, and certain. In the right panel : even if the total coverage is the same, the uncertainty over θ varies according to the count of non converted (NC) and converted (C) reads.
doi:10.1371/journal.pone.0097349.g004

(first we determined which positions had been sequenced in both samples; then we extracted a random subset of those). One can feed columns 6,7,10,11 directly to the methyl_diff executable (those columns are the unconverted, converted reads from the two samples).

2. Secondly, they can be downloaded from the BLUEPRINT project ftp site ftp.ebi.ac.uk/pub/databases/blueprint/data/homo_sapiens/.

Author Contributions

Conceived and designed the experiments: ER SH. Performed the experiments: ER MD. Analyzed the data: ER SH. Contributed reagents/materials/analysis tools: ER. Wrote the paper: ER.

References

1. Calladine CR, Drew H, Luisi B, Travers A (2004) Understanding DNA: The Molecule and How it Works. Academic Press.
2. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human dna methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.
3. Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, et al. (2014) Moabs: model based analysis of bisulfite sequencing data. Genome Biology 15: R38.
4. Hansen KD, Langmead B, Irizarry RA (2012) Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 13: R83.
5. Feng H, Conneely KN, Wu H (2014) A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic acids research: gku154.
6. Cook JD (2008) Numerical computation of stochastic inequality probabilities. Technical report, UT MD Anderson Cancer Center Department of Biostatistics.
7. Cook JD (2005) Exact calculation of beta inequalities. Technical report, UT MD Anderson Cancer Center Department of Biostatistics.
8. Abramowitz M, Stegun I (1965) Handbook of Mathematical Functions. Dover Publications Inc.