Bioinformatics, 31(21), 2015, 3401–3405
doi: 10.1093/bioinformatics/btv370
Advance Access Publication Date: 25 June 2015
Discovery Note

OXFORD

Genome analysis

Predicting tumor purity from methylation microarray data

Naiqian Zhang[1,†], Hua-Jun Wu[2,†], Weiwei Zhang[1], Jun Wang[1], Hao Wu[3,*] and Xiaoqi Zheng[1,*]

[1]Department of Mathematics, Shanghai Normal University, Shanghai 200234, China, [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston 02215, MA, USA and [3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

*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.
Associate Editor: John Hancock

Received on April 6, 2015; revised on May 17, 2015; accepted on June 10, 2015

Abstract

Motivation: In cancer genomics research, one important problem is that the solid tissue sample obtained from clinical settings is always a mixture of cancer and normal cells. The sample mixture brings complication in data analysis and results in biased findings if not correctly accounted for. Estimating tumor purity is of great interest, and a number of methods have been developed using gene expression, copy number variation or point mutation data.
Results: We discover that in cancer samples, the distributions of data from Illumina Infinium 450 k methylation microarray are highly correlated with tumor purities. We develop a simple but effective method to estimate purities from the microarray data. Analyses of the Cancer Genome Atlas lung cancer data demonstrate favorable performance of the proposed method.
Availability and implementation: The method is implemented in InfiniumPurify, which is freely available at https://bitbucket.org/zhengxiaoqi/infiniumpurify.
Contact: xqzheng@shnu.edu.cn or hao.wu@emory.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances of high-throughput technologies have revolutionized the cancer genomics research. Tremendous efforts have been made to study the genomic characteristics of different types of cancers, for example by large consortiums like The Cancer Genome Atlas (TCGA). Results from these studies contribute significantly to the understanding of cancer etiologies and the discovery of diagnostic biomarkers and therapeutic targets. In cancer research, one important problem cannot be overlooked is that the solid tissue sample obtained from clinical settings is always a mixture of cancer and normal cells. The mixture brings complication to data analysis and results in biased and even erroneous findings if not correctly accounted for (Olshen *et al.*, 2011).

The problem of estimating 'tumor purity', or the proportion of cancer cells from a mixed tissue, has been of great interests. A number of statistical methods and software tools were developed over the last several years, based on data for gene expression (Ahn *et al.*, 2013; Clarke *et al.*, 2010; Yoshihara *et al.*, 2013) or copy number variations and point mutations (Bao *et al.*, 2014; Carter *et al.*, 2012; Roth *et al.*, 2014; Su *et al.*, 2012). In particular, ABSOLUTE (Carter *et al.*, 2012) is a statistical method for inferring tumor purity and absolute copy number from a mixed sample based on SNP array data. It takes the estimated copy ratios as input and jointly estimates purity and ploidy by maximizing the whole data likelihood. ABSOLUTE has been applied to most of the TCGA samples, and its purity estimates have become the *de facto* gold standards provided by TCGA. Nevertheless, many of the existing methods still require data from purified samples or significant prior biological knowledge, which could be difficult to obtain in clinical settings due to the heterogeneity of cancer. The limitations significantly undermine the practical usefulness of the existing methods.

In this work, we discover through extensive real data exploration that DNA methylation data from Illumina Infinium 450 k microarray are informative for tumor purities. DNA methylation is an important epigenetic modification of DNA molecule with essential role in many basic biological processes. It often shows abnormality in various types of cancers (Irizarry *et al.*, 2009). An important feature of DNA methylation is that a majority of the genome are either fully methylated or unmethylated (Lister *et al.*, 2009). From a mixed tumor tissue, genomic regions that are differentially methylated between cancer and normal will likely show mid-level methylation. So the distribution of mid-level methylation provides information for the tumor purity. Although there are several reports on the existence of intermediately methylated regions (IMR) even from pure sample (Elliott *et al.*, 2015; Landau *et al.*, 2014; Stadler *et al.*, 2011), our real data observations show that there are much more IMRs from mixed samples, so that the tumor purity can still be inferred from IMRs with carefully designed algorithm.

A recently developed method MethylPurify estimates tumor purity from bisulfite-sequencing (BS-seq) data (Zheng *et al.*, 2014). The method uses the methylation information from single sequence read and performs estimation through a two-component mixture model using EM algorithm. An important advantage of the method is that it does not require data from reference samples, so it has a wider application in clinical settings. However, because of the high cost of BS-seq experiment, application of the method is still rather limited.

In large-scale population level studies, for example the epigenome-wide association study (epiGWAS), microarray technology such as Illumina Infinium 450 k microarray is still widely applied. In this work, we seek to understand whether the tumor purity can be estimated from methylation microarray data. We discover that there exist probes, mostly with intermediate methylation levels, that are informative for inferring tumor purity. We develop a method to estimate tumor purity from Illumina Infinium 450 k microarray and then apply the method on the TCGA lung adenocarcinoma (LUAD) data. Results show that the method can accurately predict the tumor purity levels.

2 Methods

We first provide a formal justification of the validity of our approach. Denote the proportion of cancer cells in solid tumor be α. Because of the differential methylation between cancer and normal samples, from methylation microarray there will be a number of probes located in the differentially methylated regions (DMRs). These probes are referred to as differentially methylated probes (DMPs) hereafter.

Within DMRs genome-wide, assume the true methylation levels from the pure cancer/normal cells follow two distinct, unimodal distributions with mode μ_k, where $k = 0/1$ represent hypo- and hyper-methylation. In other words, we assume that the higher methylation levels (regardless of its sample of origin) in all DMRs follow a distribution with mode μ_1, and the lower methylation levels follow a distribution with mode μ_0. Then in the mixed sample, since the difference between μ_0 and μ_1 will reasonably large, beta values (β) from DMPs will show mid-level methylation if α is not too close to 0 or 1. In addition, they will follow a bimodal distribution with two modes located at $\alpha\mu_0 + (1-\alpha)\mu_1$ and $\alpha\mu_1 + (1-\alpha)\mu_0$, respectively. The locations of two modes, which are functions of α, can then be used to estimate α. The essence of the proposed method is to identify DMPs and then use the distribution of their beta values to estimate tumor purity. We design following estimation procedure. For illustration purpose, we use TCGA LUAD data as an example.

The algorithm is summarized in Supplementary Figure S1. The first step of the algorithm is to select DMPs. Since according to the aforementioned model, only data from DMPs provide information for tumor purity. Including non-DMPs will weaken the signal to noise ratios. We argue that a CpG site is only informative for tumor purity estimation when it exhibits stable methylation differences between pure normal and tumor cells and has relatively large variance among tumor samples due to different normal cell contaminations. So we select probes with following two characteristics as DMPs: (i) they show differential methylations among cancers and normal and (ii) their beta values have large variance in tumor samples. To do so, we first conduct a non-parametric Wilcoxon Rank-Sum test on each probe between the tumor and normal and select probes with P value less than a pre-defined threshold as candidate DMPs. Next for all candidate DMPs, we compute their variances of the beta values from all cancer samples and filter out the ones with very small variances. This step is necessary because we observe from real data that there are non-trial number (~20–30%) of candidate DMPs with very small variances cross cancers. This is perhaps caused by technical artifacts such as probe effects, since different tumor purities from cancer should result in relatively large variance for DMPs. We exclude these probes and use the rest as true DMPs for next step of the algorithm.

With DMPs available, the beta values from DMPs within a dataset follow a bimodal distribution and the location of the modes will be used to estimate purity. To increase the signal to noise ratio in data, we use the following procedure to convert the bimodal into a unimodal distribution. We first determine the hypo-/hyper-methylation status for all selected DMPs through comparing their mean beta values of DMPs from two groups. DMPs with higher mean beta values in cancer groups are deemed hyper-methylated in cancer and vice versa. Next, we transform all beta values for DMPs according to their methylation status. The beta values for hypo-methylated probes will be changed to $1-\beta$, and there will be no change for hyper-methylated probe. We assume that approximately $\mu_0 + \mu_1 = 1$. This assumption is valid based on real data observation from both microarray and BS-seq data. Under this assumption, the transformed beta values for all DMPs will follow a unimodal distribution with mode located at $\alpha\mu_1 + (1-\alpha)\mu_0$, denote such value as m_c. To estimate m_c, we compute the probability density of all transformed DMP beta values using kernel density estimation. The mode of the distribution is deemed estimate of m_c.

The estimated m_c is related but not exactly equal to the tumor purity. Consider tumor sample i, we have $m_{ci} = \alpha_i\mu_1 + (1-\alpha_i)\mu_0 = \mu_0 + (\mu_1 - \mu_0)\alpha_i$. Here we implicitly assume that μ_0 and μ_1 take the same values from all samples, which is reasonable. Given the size of the genome, it is conceivable that the modes of methylation level distributions from cancer/normal within DMRs are very similar from different patients. The model suggests that there is a linear relationship between m_{ci} and α_i, and the coefficients depend on μ_0 and μ_1. Instead of making assumptions on μ_0 and μ_1, we take a supervised learning approach to estimate α_i. To be specific, we obtain the purity estimates from ABSOLUTE for the LUAD samples, use them as true α_i and then fit following linear regression model: $\alpha_i = b_0 + b_1 m_{ci} + \epsilon_i$. The estimated model (regression coefficients) can then be used to convert m_{ci} to α_i.

The method is implemented in software InfiniumPurify, which is freely available at https://bitbucket.org/zhengxiaoqi/infiniumpurify. InfiniumPurify provides excellent computational performance. Processing and estimating the purity of all 466 samples in LUAD data only takes 24 min on a single PC.

3 Results

We obtain LUAD data from TCGA, which profiled 466 resected LUADs and 32 matched normal samples using messenger RNA, microRNA and DNA sequencing integrated with copy number, methylation and proteomic analyses (Cancer Genome Atlas Research, 2014). The tumor purity estimates from ABSOLUTE are readily available for 197 samples.

We first look at the distribution of beta values from tumor tissues and normal controls. Figure 1A and B shows the beta value distributions for cancer and normal samples, respectively. One can observe a clear difference between tumor and normal samples in their global methylation distributions, in particular the number of intermediate methylated probes. Even though there are non-trivial numbers of probes with intermediate methylation levels from both samples, the cancer samples have much more such probes, evidenced by the bump around the middle methylation levels. Figure 1C compares the percentage of probes with intermediate methylation levels in all tumor and normal samples. Clearly the percentages of such probes are much greater for tumor samples. These results validate our claim that there are more intermediate methylation regions in cancer samples and demonstrate the possibility to use some probes from these regions to infer tumor purity.

We next apply the proposed procedure to select DMPs between tumor and normal samples. We use a rather stringent cutoff (P value = 1e-19) as threshold to select candidate DMPs because results show that only a small number of most informative DMPs are enough to accurately determine the tumor purity. We also tried other P-value cutoffs and final results from those are shown in Supplementary Table S1. There are a total of 1257 probes selected after this step. Variance of methylation levels of these candidate DMPs in tumor samples are much greater than those in normal samples in general (Fig. 1D). However, there are still some probes with small variances, likely caused by artifacts. We only keep probes with variance greater than 0.005 in tumor samples as the final DMPs. After this step, a total 957 probes are selected as DMPs (699 hypermethylation and 258 hypo-methylation), which accounts for 0.197% of all probes. Figure 1E shows the distribution of untransformed beta values of DMPs from one randomly selected tumor samples. The beta values are mostly mid-level because there are no spikes at close to 0 or 1, unlike the distributions shown in Figure 1A or B. On average, 87.0% of the selected DMPs are from intermediate methylation regrions (methylation levels between 0.2 and 0.8) for all tumor samples. This validate our claim that some probes from intermediate methylation level regions are informative for purity. Furthermore, the distribution clearly show two modes. The bimodal distributions demonstrate our theory the distributions of beta values from the tumor samples are affected by the sample purity, which furture validate our approach of using such distributions to estimate purities.

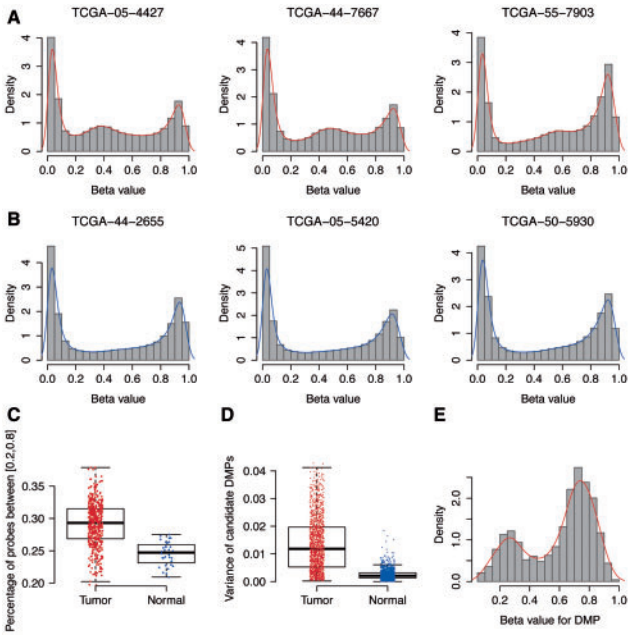

Fig. 1. (**A**) Distribution of beta values from tumor samples. (**B**) Distribution of beta values from normal samples. (**C**) Boxplot comparing percentages of probes with beta values between 0.2 and 0.8, for all tumor and normal samples. (**D**) Variances in beta values of candidate DMPs in tumor and normal samples. (**E**) Distributions of beta values of DMPs, from a tumor sample

Next we apply the proposed procedure to estimate purity from all tumor samples. Figure 2A shows the scatter plot of the estimated m_c and the purity estimates from ABSOLUTE. The Pearson correlation is very high at 0.831. In spite of the high correlation, one can see the scales of these estimates are different. This indicates that the m_c values cannot be directly used as the purity estimates. The reasons for different scales can be complicated, involving the variances of beta value distribution, microarray measurement errors from different artifacts, or different stages of tumorigenesis in genetics (ABSOLUTE) or epigenetics (InfiniumPurify). It will be very difficult to figure out and correct all the artifact. That is the reason why we design the last step of the algorithm, which is to use a supervised training approach to convert the m_c values to purity.

We use a 10-fold cross validation to verify the approach. To be specific, we divide all 466 tumor samples into 10 subsets. Nine of them are used as training data and then the result is validated in the 10th one. In each training step, the algorithm is applied to obtain the estimated linear regression coefficients b_0 and b_1. We then estimate m_c values from the test data and apply the linear model to transform them into purity levels. Figure 2B shows the results from cross validation, e.g. the predicted purity levels versus the ABSOLUTE estimates from the test data. Again, the correlation is very high at 0.829. These results strongly support our proposed approach and show that using methylation microarray data for estimating tumor purity provides very accurate results, and the accuracy is comparable to those from ABSOLUTE.

To evaluate the effect of DMP selection on final results, we try using different P-value cutoffs for Wilcoxon Rank-Sum test. Results are shown in Supplementary Table S1. We find that using a looser cutoff and including many probes DMPs will hurt the final results.

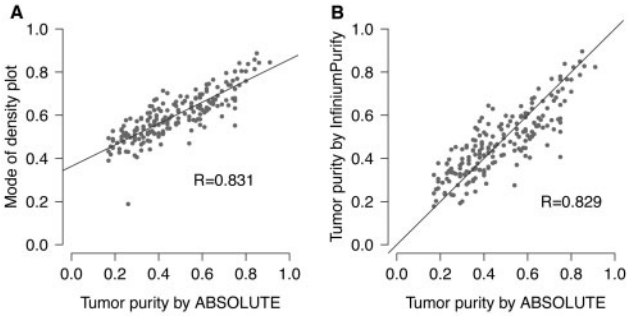

Fig. 2. Tumor purity estimation results from TCGA LUAD data. (**A**) Scatter plot of m_c versus ABSOLUTE estimations. (**B**) Estimation from InfiniumPurify versus ABSOLUTE, from 10-fold cross validation

For example using 1e-10 as cutoff chooses close to 45 000 DMPs and the final purity estimations has a correlation of 0.14 with ABSOLUTE results. On the other hand, using 1e-20 as cutoff and a little more than 100 DMPs produces a correlation of 0.80. These results suggest that it is better to be more conservative on DMPs selection. Overall, using 1000 or so DMPs gives the best results and using two to four thousands provide slightly worse but comparable results.

To test the robustness of our method, we further apply InfiniumPurify to several other tumor types from TCGA, including lung squamous cell carcinoma, colorectal adenocarcinoma, breast cancer, head and neck squamous cell carcinoma and glioblastoma multiforme. Results from these samples are consistent with that from LUAD data and show strong correlation (over 0.8 on average) with ABSOLUTE estimates. These further demonstrate that InfiniumPurify provides consistent results and can serve as an alternate for ABSOLUTE in estimating tumor purity (Supplementary Fig. S2). Moreover, regression coefficients (b_0 and b_1) are quite similar across different cancer types, implying that InfiniumPurify could potentially infer purity of cancer types without ABSOLUTE estimates by borrowing existing regression coefficients.

We also compare our method with two other exising tumor purify estimation tools based on copy number variance data, i.e. AbsCN-seq (Bao *et al.*, 2014) and THetE2 (Oesper *et al.*, 2014). However, the published purity estimates from AbsCN-seq and THetE2 are rather limited, and we are only about to find a few samples with both AbsCN-seq/THetE2 estimates and 450 K array (which are necessary for our method): two samples for AbsCN-seq and four samples for THetE2. The purity estimates for these samples are provided in Supplementary Tables S2 and S3. In general, results show that the estimates from InfiniumPurify have strong correlations with both AbsCN-seq and THetE2, further demonstrating that InfiniumPurify is in good agreement with copy number variation (CNV)-based methods even though it uses completely different biological infromation and data type.

4 Conclusion

In this work, we discover that the methylation microarray data from tumor samples contain important information for tumor purity. In particular, the shape of the beta values distribution is strongly influenced by the purity. We explained the phenomenon using a statistical model and show that the distributions can be utilized to estimate tumor purity. In essence, beta values of informative probes (DMPs) follow a bimodal distribution, and the locations of the modes are related to the tumor purity. By obtaining the modes of these beta values, tumor purities can be estimated from a linear model. We design an algorithm InfiniumPurify for purity estimation and show that it provides good results from several sets of TCGA data.

There are several advantages of InfiniumPurify compared with existing tumor purity estimation methods. First it does not require data from reference panels, which could be difficult to obtain in clinical settings due to the heterogeneity of cancer. Second, it is very cost effective compared with base resolution methylation data such as whole-genome BS-seq. Moreover, since the DNA samples are much easier to collect compared with mRNA, the purity estimation based on methylation data will provide a more practical mean in clinical settings. From large cancer-related epiGWAS studies, the purity estimation can be obtained as a by-product from the methylation microarrays. Finally, ABSOLUTE usually provides several solutions corresponding to local maxima of the likelihood. Prediction results from InfiniumPurify can help researches make a final decision.

We use ABSOLUTE estimates as benchmark for model training and results comparison because of the following reasons. First, ABSOLUTE is one of the earliest and arguably the most influential tools for tumor purity estimations. It is widely used as a gold standard to evaluate new purity estimation methods, for example AbsCN-seq and THetE2. Second, it provides tumor purity estimates for several types of tumors and thousands of samples, which can facilitate large-scale comparison. Third, InfiniumPurify is designed and tested using data from TCGA samples, and ABSOLUTE is the official tumor purity predictor for TCGA consortium. The fact that InfiniumPurify estimates are highly correlated with ABSOLUTE estimates, albeit using a very different data type and computational model, further justifies the accuracy of ABSOLUTE estimates independently.

We compared our results with several CNV-based methods and show good agreements in purity estimation. However, although both CNV and methylation contain information for cancer purity, they do not have to be correlated in the raw data level (Feber *et al.*, 2014; Houseman *et al.*, 2009). Both types of methods use genome-wide distributions of quantities (CNV or intermediate methylated probes) to estimate purity, the CNV and methylation could be completely unrelated at finer scales such as within a few hundred base pairs.

It is important to note that InfiniumPurify does not take ploidy information of tumor cells into consideration, which could slightly bias our prediction. However, since the prediction is based on relatively large number of DMPs across the genome and a majority of them will not have copy number variation, we expect the aberrant copy numbers would not significantly change the overall result.

The selection of DMPs plays an important role in the method. In its current form, InfiniumPurify only works when the number samples is reasonably large. Fortunately, in TCGA, most important cancer types have large sample size. It will be interesting and useful to explore whether it is possible to combine data from different cancers and construct 'universal' DMPs. Moreover, InfiniumPurify is only focused on estimating tumor purity. Decoupling signals from mixed sample to estimate the methylation levels from cancer/normal is our research plan in the near future.

InfiniumPurify is specifically designed for data from Illumina Infinium 450 k arrays, which is the most widely used platform for DNA methylation. It is conceivable that the same principle and methods can be applied to data from other platforms.

In this work, we present a simple but effective method to estimate tumor purity from methylation microarray data. The method can serve as an alternative for existing methods with similar goals. Although the method is focused on tumor tissues, it can potentially be applied to other highly heterogeneous samples such as blood or brain. Furthermore, it will be very interesting to combine methylation with other genetic and genomic data such as gene expression and genetic variants and build a joint model for tumor purity estimation. Such model will have potential to significantly improve the estimation accuracy.

Acknowledgements

The authors thank Dr Xiaole S. Liu for helpful suggestions and comments. The results here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Funding

This work was supported by the National Natural Science Foundations of China (31100953 to X.Z.; 11171224 to J.W.).

Conflict of Interest: none declared.

References

Ahn,J. *et al.* (2013) DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, **29**, 1865–1871.

Bao,L. *et al.* (2014) AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics*, **30**, 1056–1063.

Cancer Genome Atlas Research. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Clarke,J. *et al.* (2010) Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, **26**, 1043–1049.

Elliott,G. *et al.* (2015) Intermediate DNA methylation is a conserved signature of genome regulation. *Nature communications*, **6**, 6363.

Feber,A. *et al.* (2014) Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol.*, **15**, R30.

Houseman,E.A. *et al.* (2009) Copy number variation has little impact on bead-array-based measures of DNA methylation. *Bioinformatics*, **25**, 1999–2005.

Houseman,E.A. *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.

Irizarry,R.A. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.

Landau,D.A. *et al.* (2014) Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*, **26**, 813–825.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Oesper,L. *et al.* (2014) Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, **30**, 3532–3540.

Olshen,A.B. *et al.* (2011) Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics*, **27**, 2038–2046.

Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–398.

Stadler,M.B. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.

Su,X. *et al.* (2012) PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, **28**, 2265–2266.

Yoshihara,K. *et al.* (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.

Zheng,X. *et al.* (2014) MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol.*, **15**, 419.