SOFTWARE TOOL ARTICLE

# shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R [version 1; referees: 2 approved]

Jean-Philippe Fortin[1], Elana Fertig[2], Kasper Hansen[1,3]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA
[2]Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA
[3]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA

**v1** **First published:** 30 Jul 2014, **3**:175 (doi: 10.12688/f1000research.4680.1)
**Latest published:** 19 Sep 2014, **3**:175 (doi: 10.12688/f1000research.4680.2)

## Abstract

We present shinyMethyl, a Bioconductor package for interactive quality control of DNA methylation data from Illumina 450k arrays. The package summarizes 450k experiments into small exportable R objects from which an interactive interface is launched. Reactive plots allow fast and intuitive quality control assessment of the samples. In addition, exploration of the phenotypic associations is possible through coloring and principal component analysis. Altogether, the package makes it easy to perform quality assessment of large-scale methylation datasets, such as epigenome-wide association studies or the datasets available through The Cancer Genome Atlas portal. The shinyMethyl package is implemented in R and available via Bioconductor. Its development repository is at https://github.com/jfortin1/shinyMethyl.

This article is included in the RPackage channel.

This article is included in the Bioconductor channel.

**Open Peer Review**

**Referee Status:** ✓ ✓

| | Invited Referees 1 | 2 |
|---|---|---|
| REVISED version 2 published 19 Sep 2014 | ✓ report | |
| version 1 published 30 Jul 2014 | ✓ report | ✓ report |

1 **Timothy Triche**, University of Southern California USA

2 **Tiffany Morris**, University College London UK

**Discuss this article**

Comments (0)

**Corresponding authors:** Jean-Philippe Fortin (jfortin@jhsph.edu), Kasper Hansen (khansen@jhsph.edu)

**How to cite this article:** Fortin JP, Fertig E and Hansen K. **shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R [version 1; referees: 2 approved]** *F1000Research* 2014, **3**:175 (doi: 10.12688/f1000research.4680.1)

**Copyright:** © 2014 Fortin JP *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

**Grant information:** JPF was partially supported by the Natural Sciences and Engineering Research Council of Canada and by les Fonds de recherche Nature et technologies du Québec as well as under the Johns Hopkins Head and Neck Cancer SPORE awarded to EJF.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Competing interests:** No competing interests were disclosed.

**First published:** 30 Jul 2014, **3**:175 (doi: 10.12688/f1000research.4680.1)

## Introduction

The recent release of the R package *shiny*[1] has substantially lowered the barriers to interactive visualization in R, opening the door to interactive exploration of high-dimensional genomic data.

DNA methylation is an epigenetic mark, and changes in DNA methylation have been associated with various diseases, such as cancer[2]. For DNA methylation data, thousands of samples from the state-of-the-art Illumina 450k methylation array[3] have been generated and are accessible online from The Cancer Genome Atlas (TCGA) and through the Gene Expression Omnibus (GEO). This array has a series of probes used to measure a methylation and an unmethylation signal for a series of loci. Probes are designed using two main chemistries resulting in a challenging array design, essentially a mix of a two color and a one color array discussed in Bibikova *et al.*[3]. Analysis of data from this array requires careful quality control and pre-processing that account for these distinct chemistries. The assessment of these steps could benefit from an interactive visualization tool.

Our solution is *shinyMethyl*, an interactive visualization package for 450k arrays, based on the packages *minfi*[4] and *shiny*[1]. The goal of shinyMethyl is two-fold; (1) to help with quality assessment and (2) to help with assessing the effect of pre-processing. We use pre-computation to enable interactive visualization of thousands of samples to circumvent computational bottlenecks during data exploration. The pre-computation can happen on a large computing server and the resulting data object can be used for interactive visualization on a laptop. Quality control and pre-processing large 450k datasets become easy and intuitive with *shinyMethyl*.

## Methods

### shinyMethyl workflow

The first step of *shinyMethyl* is pre-computation of various summaries of the 450k array data, using the function `shinySummarize`. This pre-computation is run on raw (not pre-processed) data and – optionally – pre-processed data, resulting in either one or two summary objects, as described below. These summary objects, called `shinyMethylSet`, are saved in a platform-independent format. The interactive interface is then launched via the function `runShinyMethyl`. The function requires a `shinyMethylSet` containing the summary data from the raw data. In addition, the function accepts as a second argument a `shinyMethylSet` that contains summaries from pre-processed data, in which case both raw and pre-processed data will be displayed in the interactive interface. Figure 1 illustrates the *shinyMethyl* workflow.

### Raw data summarization

Summarizing the raw data uses the *minfi*[4] and *illuminaio*[5] R packages to parse Illumina IDAT files into a *minfi* object called `RGChannelSet`. `shinySummarize` operates on this `RGChannelSet` and the summarization object created by this function is 35x smaller than the full data representation in *minfi*; 1,000 samples use 205 MB. Specifically, the summarized data contain the quantile distributions of the raw intensities for the unmethylated (U) and methylated (M) channels, copy numbers (CN = M + U), Beta values (Beta) and M values (M-Val). The object contains also the raw control probes intensities and the results of the principal component

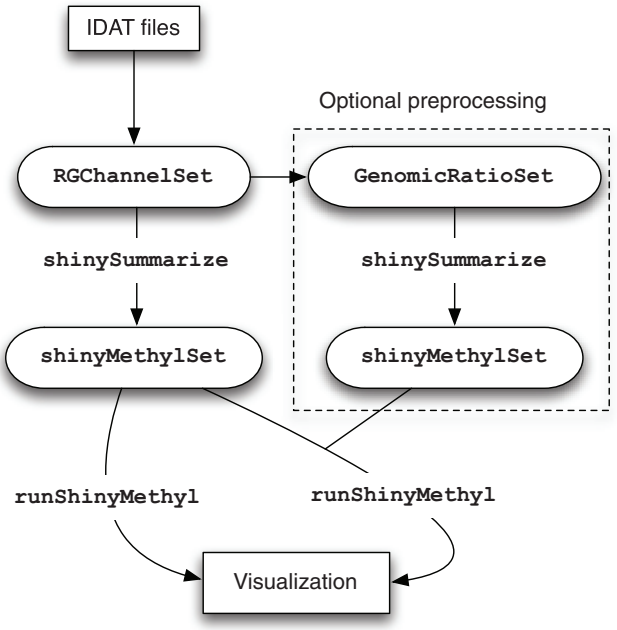


**Figure 1. The workflow of *shinyMethyl*.** IDAT files are parsed using *minfi* and *illuminaio* into a `RGChannelSet`. This object is summarized using `shinySummarize`. Optionally, the data are pre-processed and the pre-processed data are summarized. For visualization, `runShinyMethyl` is used on either one or two sets of summarized data.

analysis performed on the autosomal Beta values. The function also extracts the phenotype variables stored in the `RGChannelSet`. The summarization is done separately by probe types (I and II, see Bibikova *et al.*[3]) and for sex chromosomes. An S4 class, called `shinyMethylSet`, is used to represent the data in R, and this object is independent of the operating system. The *shinyMethyl* interface is launched by passing the `shinyMethylSet` to the function `runShinyMethyl`. An example of the interface is shown in Figure 2.

### Pre-processed data summarization (optional)

Summarizing pre-processed data in *shinyMethyl* operates on an S4 object in *minfi* termed `GenomicRatioSet`. The summaries of the pre-process data are stored in an additional `shinyMethylSet`. Again, the summarized data object is substantially smaller than the full data representation in *minfi*. If this `shinyMethylSet` is also included in the `runShinyMethyl` command, the summaries of the pre-processed data are automatically added to the *shinyMethyl* interface. This option represents a powerful diagnostic tool to assess the global performance of a normalization method, such as plate effect correction (Figure 2), or preservation of the expected biological differences between different tissues or conditions (Figure 3).

### Quality control assessment

Once the DNA methylation data have been summarized, *shinyMethyl* offers three interactive plots for quality control. These plots react conjointly to the user mouse: (1) a density plot of the M/Beta

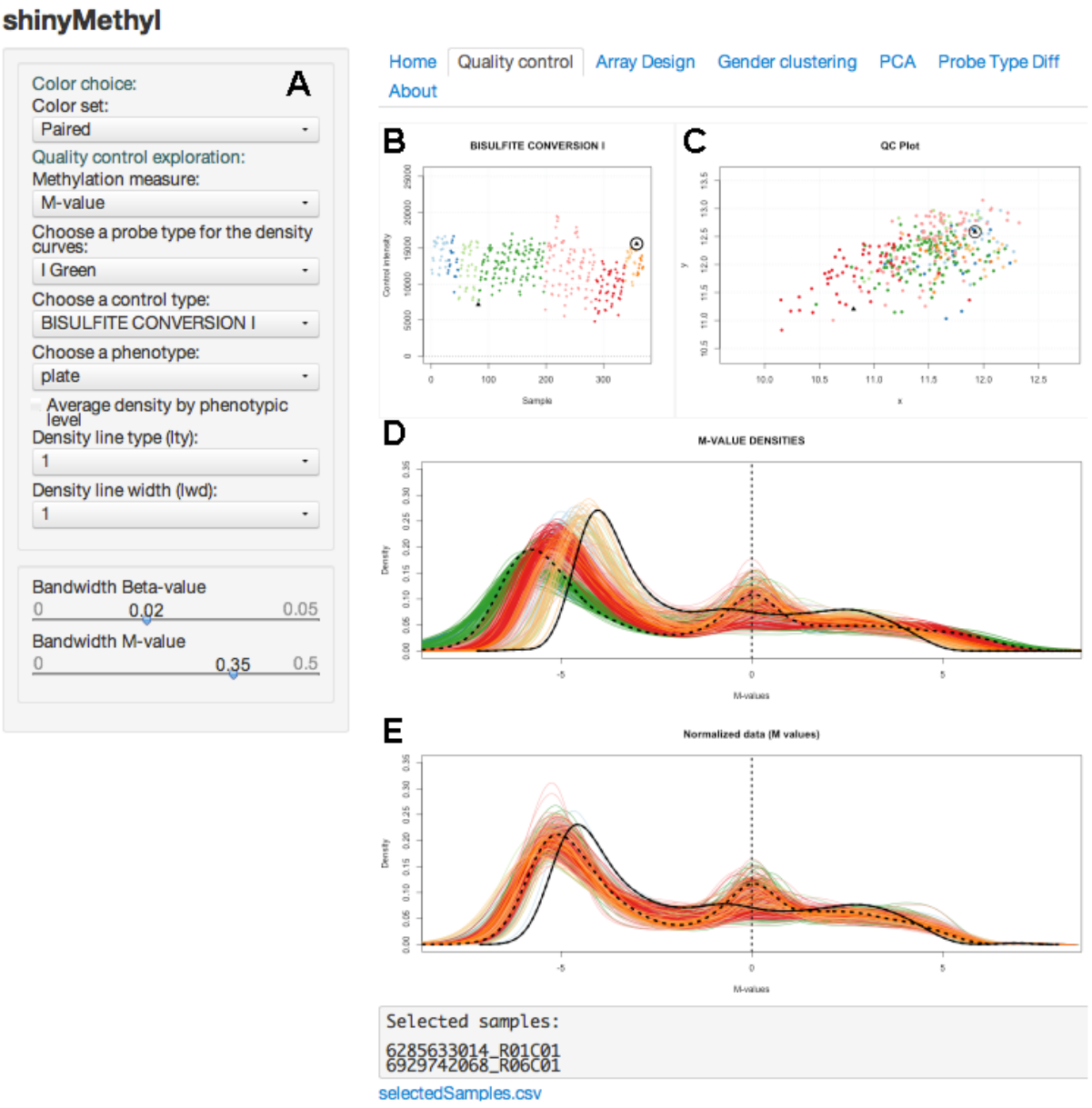


**Figure 2. The shinyMethyl user interface for quality control.** The interface shows an example interactive visualization of batch effects and quality control (TCGA head and neck squamous cell carcinoma, HNSCC dataset). The interface is divided into a user menu and a plotting area. (**a**) A menu containing a number of user-settable visualization parameters. The "phenotype" is set to "plate" which makes the color scheme reflect batch. The four plots (**b**–**e**) are interactive and react simultaneously to the user mouse clicks, so that samples selected on one plot are immediately highlighted on the additional plots. The solid lines in black represents the sample(s) currently selected by the user and match the dot circled in black on (**b**,**c**). The dashed lines in black represents another sample, previously selected by the user and match the black dot without the circle. (**b**) Average negative control probes intensities; (**c**) the median intensity of the M channel against the median intensity of the U channel; (**d**–**e**) M-value densities for Infinium I probes before and after functional normalization.

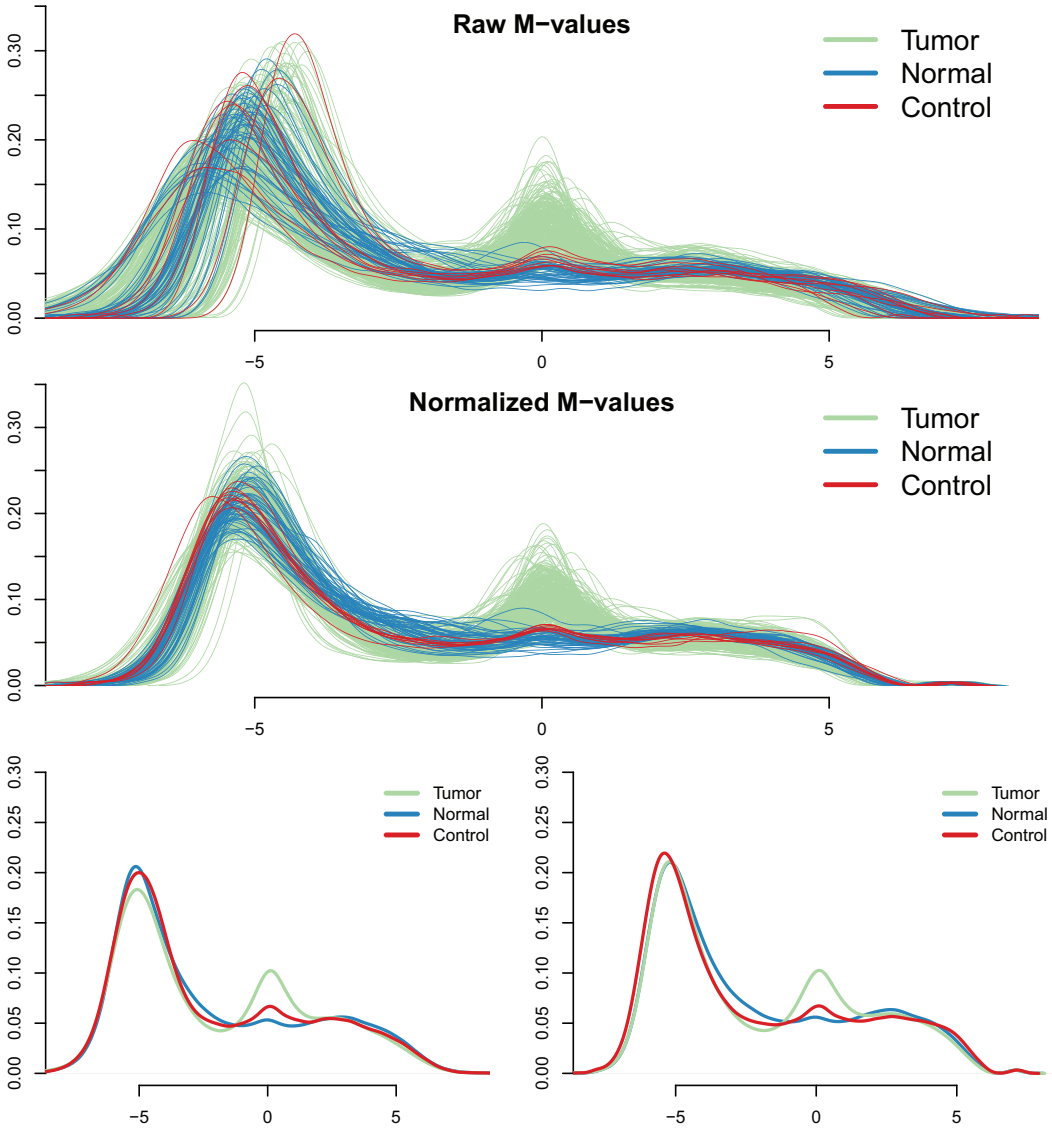


**Figure 3. Visualization of cancer/normal differences in the TCGA dataset, before and after normalization.** In the first two plots are shown the densities of the M-values for Type I green probes before (**a**) and after (**b**) functional normalization as presented in the shinyMethyl interactive interface. Green and blue densities represent tumor and normal samples respectively, and red densities represent 9 technical replicates of a control cell line. The last two plots show the average density for each sample group before and after normalization. Functional normalization preserves the expected marginal differences between normal and cancer, while reducing the variation between the technical controls (red lines).

values, (2) a QC plot proposed in *minfi* and (3) a plot of control probes intensities. The samples are colored by a phenotype variable selected by the user. The three plots together allow the user to select aberrant samples, whose array identifiers are saved into a csv file for exclusion in subsequent analyses (outside of *shinyMethyl*). An example of quality control panel is presented in Figure 2 in which summaries from the TCGA head and neck squamous cell carcinoma (HNSCC) samples are colored by batch; *shinyMethyl* allows to observe significant batch effects, a source of obscure variation that has critical consequences in downstream analysis[6].

## Sex prediction

The sex of the samples can be accurately predicted by using the intensities of the probes mapping to the sex chromosomes in the

M and U channels[4]. *shinyMethyl* implements this prediction algorithm and allows the user to interactively specify a cutoff to cluster samples by sex.

The array identifiers of the samples for which the predicted sex does not agree with the user-provided sex phenotype are displayed within the interface and can be saved into a csv file for further analysis. From the HNSCC TCGA dataset (described in Example data), one sample shows discrepancy, indicating possible mislabeling (Figure 4).

## PCA analysis and design confounding

*shinyMethyl* also performs a principal component analysis (PCA) on the 20,000 most variable autosomal probes. This analysis enables the observation of associations between phenotype and methylation levels. An additional panel displays the physical arrays colored by phenotype. This coloring allows the user to discern potential confounding between phenotype and study design.

## Example data

The data package *shinyMethylData* contains the summarized data for 369 HNSCC cancer samples from TCGA. It is available from the Bioconductor project (http://www.bioconductor.org). All analyses were performed on raw IDAT intensity files available from Level I data in the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga). Both raw intensities and normalized methylation values obtained by functional normalization using control probes and a slide covariate[7] are included. The `shinyMethylSet` objects

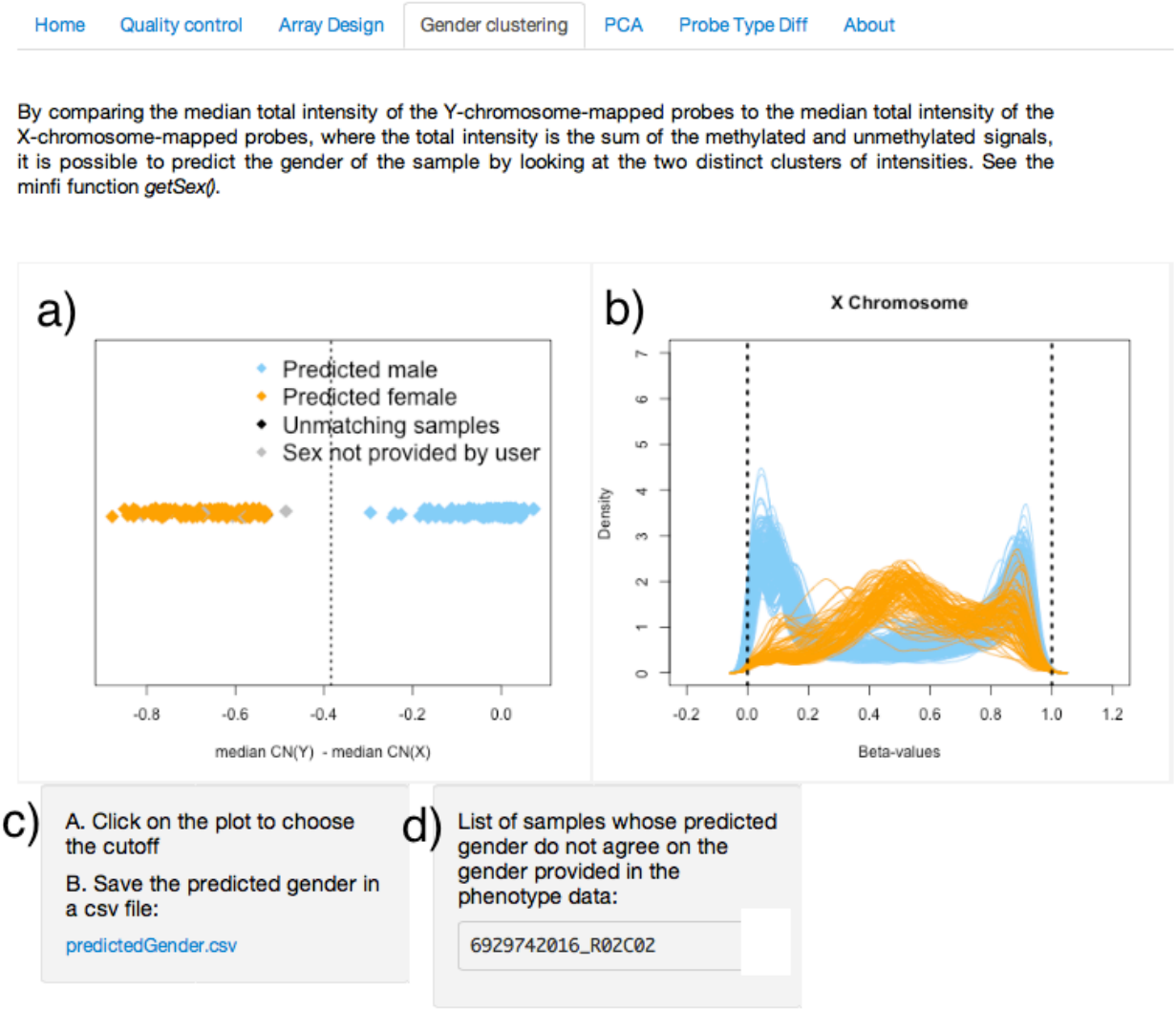


**Figure 4. Sex prediction interface.** The difference of the median copy number intensity for the Y chromosome and the median copy number intensity for the X chromosome can be used to separate males and females. In **a**), the user can select the vertical cutoff (dashed line) manually with the mouse to separate the two clusters (orange for females, blue for males). Corresponding Beta-value densities appear in **b**) for further validation. The predicted sex can be downloaded in a csv file in **c**), and samples for which the predicted sex differs from the sex provided in the phenotype will appear in **d**).

containing respectively the raw and normalized data can be accessed by `summary.tcga.raw` and `summmary.tcga.norm`.

## Discussion

*shinyMethyl* makes the quality control and pre-processing of 450k methylation array data fast and intuitive through an interactive application in R. We also show, by example, how to use *shiny* to develop interactive visualization interfaces. Our example will facilitate future developments of interactive visualization tools for the processing of high-dimensional genomic data in subsequent Bioconductor[8] packages.

## Software availability

### Software access

*shinyMethyl* is an R package available from the Bioconductor project (http://www.bioconductor.org).

### Latest source code

https://github.com/jfortin1/shinyMethyl

### Source code as at the time of publication

https://github.com/F1000Research/shinyMethyl/releases/tag/v1.0

### Archived source code as at the time of publication

http://dx.doi.org/10.5281/zenodo.10748[10]

### Software license

Artistic-2.0

### Author contributions

JFP conceived and developed the shinyMethyl package, supervised by EJF and KDH. All authors wrote and approved the final manuscript.

### Competing interests

No competing interests were disclosed.

### Grant information

JPF was partially supported by the Natural Sciences and Engineering Research Council of Canada and by les Fonds de recherche Nature et technologies du Québec as well as under the Johns Hopkins Head and Neck Cancer SPORE awarded to EJF.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## References

1. R Studio and Inc. **shiny: Web Application Framework for R**. R package version 0.10.0. 2014.
Reference Source
2. Feinberg AP, Vogelstein B: **Hypomethylation distinguishes genes of some human cancers from their normal counterparts.** *Nature.* 1983; **301**(5895): 89–92.
PubMed Abstract | Publisher Full Text
3. Bibikova M, Barnes B, Tsan C, *et al.*: **High density DNA methylation array with single CpG site resolution.** *Genomics.* 2011; **98**(4): 288–95.
PubMed Abstract | Publisher Full Text
4. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al.*: **Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays.** *Bioinformatics.* 2014; **30**(10): 1363–1369.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Smith ML, Baggerly KA, Bengtsson H, *et al.*: **illuminaio: An open source IDAT parsing tool for Illumina microarrays.** *F1000Res.* 2013; **2**: 264.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Leek JT, Scharpf RB, Bravo HC, *et al.*: **Tackling the widespread and critical impact of batch effects in high-throughput data.** *Nat Rev Genet* 2010; **11**(10): 733–739.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Fortin JP, Labbe A, Lemire M, *et al.*: **Functional normalization of 450k methylation array data improves replication in large cancer studies.** *bioRxiv.* 2014.
Publisher Full Text
8. Gentleman RC, Carey VJ, Bates DM, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol.* 2004; **5**(10): R80.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Feinberg AP, Vogelstein B: **Hypomethylation distinguishes genes of some human cancers from their normal counterparts.** *Nature.* 1983; **301**(5895): 89–92.
PubMed Abstract | Publisher Full Text
10. Fortin JP, Hansen KD: **F1000Research/shinyMhethyl.** *ZENODO.* 2014.
Data Source

# Open Peer Review

## Current Referee Status: ✔ ✔

**Version 1**

Referee Report 03 September 2014

**doi:**10.5256/f1000research.5001.r5620

✔ **Tiffany Morris**
UCL Cancer Institute, University College London, London, UK

This manuscript written by Fortin *et al.* presents an R package to assess the quality control of 450k methylation array data through an interactive interface. A unique aspect of *shinyMethyl* is the use of the shiny package to allow the interactive visualisation. *shinyMethyl* is easy to use and therefore suitable for a novice data analyst with limited programming experience and also experienced bioinformaticians that would like to quickly assess large datasets. Additionally the package utilises the established and widely used R package *minfi* for the data analysis.

*shinyMethyl* is available for download from Bioconductor or github. In this manuscript, the authors clearly describe the use of the package through a workflow and various screenshots and plots. They discuss the example dataset that is available through TCGA. Data interpretation is only explained briefly as the purpose of this manuscript is to present the interface. In depth background information on analysis methods (terminology, normalisations etc) and details of data interpretation can be found in the *minfi* publication or the Bioconductor vignette.

The package does require the raw IDAT files. Some publicly available datasets may not provide this information and therefore *shinyMethyl* would not be applicable. However, it is becoming mainstream to provide this data and with more packages utilising this raw data it is expected that investigators make it available.

The authors have clearly presented their package that is well designed, easy to use and a useful addition to the options for 450k data analysis especially as it is the platform of choice for large epigenome-wide association studies.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

***Competing Interests:*** No competing interests were disclosed.

Author Response 15 Sep 2014

**Kasper Daniel Hansen**, Johns Hopkins University, USA

Thank you very much for taking the time to review our work and for the kind referee report.

***Competing Interests:*** No competing interests were disclosed.

Referee Report 19 August 2014

**doi:**10.5256/f1000research.5001.r5619

**Timothy Triche**
Center for Personalized Medicine, University of Southern California, Los Angeles, CA, USA

The title of the report is entirely appropriate.

The abstract would benefit from mention of the interactive demo on spark.rstudio.com, as the interface is trivially easy to grasp while performing the various quality control checks.

The methods not directly treated in this report (functional normalization, Shiny internals, 450k chip design) have been discussed elsewhere at length.The discussion of design and implementation decisions for shinyMethyl is sufficient (again an interactive/exploratory session will provide the interested user with most if not all of the same information).

The conclusion (shinyMethyl eases 450k quality control for large datasets with potential technical artifacts) is justified in the report, the included example for the package, and the interactive demo hosted by Rstudio.

The data for the example are available as raw binary IDAT files from the TCGA data portal; reproducibility via minfi is trivial, provided the user has sufficient compute resources to process and normalize a large dataset. (In my experience, datasets of 1000+ samples can be read with illuminaio and normalized via funnorm() on a typical Linux server with 48-64GB of RAM)

This report serves a useful purpose (beyond the announcement of the software) by providing a usable baseline for 450k quality control. Epigenome-wide association studies are prone to a staggering number of potential confounding factors, not least of which are technical and experimental design artifacts. Investigators who have an interest in reproducible conduct of such studies now have further incentive to use and deposit raw binary IDAT files for their experiments. Investigators who choose not to do so may be viewed with skepticism, especially when the data has been processed for analysis via minfi (or, for that matter, methylumi), and verification of straightforward quality controls is made simpler by the interactive shinyMethyl package.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

***Competing Interests:*** No competing interests were disclosed.

Author Response 15 Sep 2014

**Kasper Daniel Hansen**, Johns Hopkins University, USA

We thank the referee for taking the time to review our work, and thank for the kind review.

We have submitted a revision with an added sentence about the demo at spark.rstudio.com under "Software access", but caution that this website at times is slow to respond to input. This is a hosting service freely available to the community.

***Competing Interests:*** No competing interests were disclosed.