



SOFTWARE TOOL ARTICLE

**REVISED** **AGA: Interactive pipeline for reproducible gene expression and DNA methylation data analyses [version 2; referees: 2 approved]**

Michael Considine<sup>1</sup>, Hilary Parker<sup>2</sup>, Yingying Wei<sup>2</sup>, Xaio Xia<sup>3</sup>, Leslie Cope<sup>1</sup>, Michael Ochs<sup>4</sup>, Elana Fertig<sup>1</sup>

<sup>1</sup>Department of Oncology Biostatistics & Bioinformatics, Johns Hopkins University School of Medicine, Baltimore, MD, 21205, USA

<sup>2</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA

<sup>3</sup>Department of Statistics and Biostatistics, Rutgers University, New Brunswick, NJ, 08901, USA

<sup>4</sup>Department of Mathematics and Statistics, The College of New Jersey, Ewing Township, NJ, 08618, USA

**v2** **First published:** 28 Jan 2015, 4:28 (doi: [10.12688/f1000research.6030.1](https://doi.org/10.12688/f1000research.6030.1))  
**Latest published:** 21 Oct 2015, 4:28 (doi: [10.12688/f1000research.6030.2](https://doi.org/10.12688/f1000research.6030.2))

**Abstract**

Automated Genomics Analysis (AGA) is an interactive program to analyze high-throughput genomic data sets on a variety of platforms. An easy to use, point and click, guided pipeline is implemented to combine, define, and compare datasets, and customize their outputs. In contrast to other automated programs, AGA enables flexible selection of sample groups for comparison from complex sample annotations. Batch correction techniques are also included to further enable the combination of datasets from diverse studies in this comparison. AGA also allows users to save plots, tables and data, and log files containing key portions of the R script run for reproducible analyses. The link between the interface and R supports collaborative research, enabling advanced R users to extend preliminary analyses generated from bioinformatics novices.



This article is included in the **RPackage** gateway.

**Open Peer Review**

**Referee Status:** ✔ ✔

	Invited Referees	1	2
<b>REVISED</b> version 2 published 21 Oct 2015			
version 1 published 28 Jan 2015		✔ report	✔ report

1 **Subha Madhavan**, Georgetown University, USA

2 **Matthew McCall** , University of Rochester, USA

**Discuss this article**

Comments (0)

**Corresponding authors:** Michael Considine ([mconsid3@jhmi.edu](mailto:mconsid3@jhmi.edu)), Elana Fertig ([ejfertig@jhmi.edu](mailto:ejfertig@jhmi.edu))

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

**How to cite this article:** Considine M, Parker H, Wei Y *et al.* **AGA: Interactive pipeline for reproducible gene expression and DNA methylation data analyses [version 2; referees: 2 approved]** *F1000Research* 2015, 4:28 (doi: [10.12688/f1000research.6030.2](https://doi.org/10.12688/f1000research.6030.2))

**Copyright:** © 2015 Considine M *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.

**Grant information:** Funding was provided from NIH/NIDCR P50DE019032, NIH/NCI P30CA006973, K25CA141053 to EJJ, and NIH/NLM R01LM011000 to MFO.

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

**First published:** 28 Jan 2015, 4:28 (doi: [10.12688/f1000research.6030.1](https://doi.org/10.12688/f1000research.6030.1))

**REVISED Amendments from Version 1**

Summary of Updates:

- We have altered the title to be more in line with the scope of the application
- Reduced the genes in [Figure 1](#) and addressed the truncation of column names
- Removed fRMA, as it was intended to be RMA
- Removed cBioPortal from the paper
- Addressed normalization procedures
- Addressed scalability
- The funder details have been updated

**See referee reports****Introduction**

While high dimensional genetic data have increased in availability at reduced cost, robust analyses remain labor intensive and costly. Numerous automated software pipelines have been developed in an effort to increase the rate and decrease the costs at which analyses can be completed, including SVAw<sup>11</sup>, Partek<sup>4</sup>, InSilicoDB<sup>18</sup>. Automated Genomics Analysis (AGA) provides a more dynamic experience, allowing the user to start with raw data and a text file containing corresponding sample annotations from either a single or multiple studies. AGA performs all necessary normalization and batch correction, and then enables the user to interactively determine the samples to contrast in the analysis based on the sample annotations. AGA is implemented in R to facilitate adaptation of state-of-the-art genomics analysis techniques. Linking R to a web browser-based interface through RStudio's shiny also facilitates collaborative analyses in research teams with diverse bioinformatics expertise.

AGA bridges the gap between interactive and reproducible analyses for several platforms, including expression arrays, methylation arrays, and processed RNAseq data. Through the interface, the user determines the size and scope of the analyses. AGA first performs data normalization, including the ComBat<sup>7</sup> and SVA<sup>9</sup> batch correction algorithms to enable comparison across multiple datasets for non-methylation platforms. The software then performs differential analysis<sup>16</sup>, and gene set analyses<sup>2,17</sup> based upon defined sample groups. Users obtain standard visualization of genomics data, including hierarchical clustering, boxplots and heatmaps as part of the default analysis. Plots and tables summarizing the results from each analysis are customizable through the interface. The figures and tables in AGA are interactive and customizable. In contrast to other point and click software, AGA logs the R code, and exports the workspace with each figure and table, ensuring that each analysis can be reproduced and further customized. The runtime of analyses will depend largely on the desktop hardware, but also on the data platform and optional analyses selected. On a Mac Pro workstation, containing a 3.2 GHz Quad-Core Intel Xeon processor and 10Gb 1066 MHz DDR3 RAM, analyses containing under 100 samples were completed in under 30 minutes.

**Methods**

The AGA application is run through R and interactive through web browsers. AGA is implemented with RStudio's shiny<sup>13</sup>, integrating the R code used in the analysis with HTML and JavaScript, for the interactive user interface. Usage requires R version 3.0.1 or higher, and either Mozilla Firefox or Google Chrome, and R packages described in the AGA User's Manual. The program is divided into seven tabs. Clicking the respective Update button generates the results to be displayed in each tab and clicking the Download buttons save the plots and data.

**Data platforms**

AGA supports analyses of DNA methylation and gene expression data. Currently, AGA supports DNA methylation analysis on Illumina 450k arrays. It also supports gene expression analysis of any human Affymetrix expression platform, including exon arrays, and normalized gene counts from RNAseq data. Notably, the flexible format for normalized RNAseq data may be adapted to analyze normalized data from other platforms measuring continuous data, many of which we plan to incorporate in future versions of AGA.

**Initiation**

Users of AGA select to load annotation files and high throughput genomic data from files in a specified directory. AGA accepts raw CEL files and iDat files for Affymetrix and DNA methylation arrays, respectively where background correction is performed, as well as quantile normalization for the expression arrays. For gene expression microarrays, AGA performs RMA normalization implemented in the Bioconductor package affy<sup>5</sup>. Probe-level estimates of DNA methylation are computed from iDat files using Illumina standards with the minfi package<sup>1</sup>. RNAseq data are formatted as individual text files for each sample, assumed to contain gene names and normalized counts for each sample. It is assumed that normalized RNAseq data are formatted as individual text files for each sample, containing gene names and normalized counts for each sample. More details about the format for each data type are provided in the User's manual. Sample annotations are specified in a CSV file, whose first column matches the names of the data files. By default, it is assumed the annotation file defines the sample batch; however, this can be updated by editing the annotation files to contain a 'Batch' column with unique identifiers for each respective batch within the dataset. Further details about the sample annotations are also provided in the User's manual.

**Sample selection for differential analysis**

After loading in the annotation files, AGA users select categories from the annotation for differential expression analysis. AGA automatically groups samples with common levels in each category as groups for differential analysis. Samples may be further subset from the complete dataset from the criteria selected for each group. When selected, AGA updates the display to output the sample size for each group. Samples are set for analysis by clicking the "Run the Analysis!" button. In cases for which samples span multiple batches, the analysis automatically performs ComBat and SVA batch correction protecting for the biological groups in the annotation selected by the user. Help boxes are available to clarify each input field with further details in the User's manual.

### Interactive plots and tables

The Dendrogram Plot tab in displays unsupervised hierarchical clustering based upon the complete correlation between values of genes (rows) and samples (columns). The Heatmap Plot tab provides an interactive Javascript heatmap of the genomic data, allowing users to customize genes plotted and color rows by sample annotations. For both Dendrograms and Heatmaps, an option is available to view the pre-batch corrected data to show the effects of batch on and efficacy of correction of the data. The Gene Box Plot tab creates boxplots to summarize values of a user-selected gene in the selected groups.

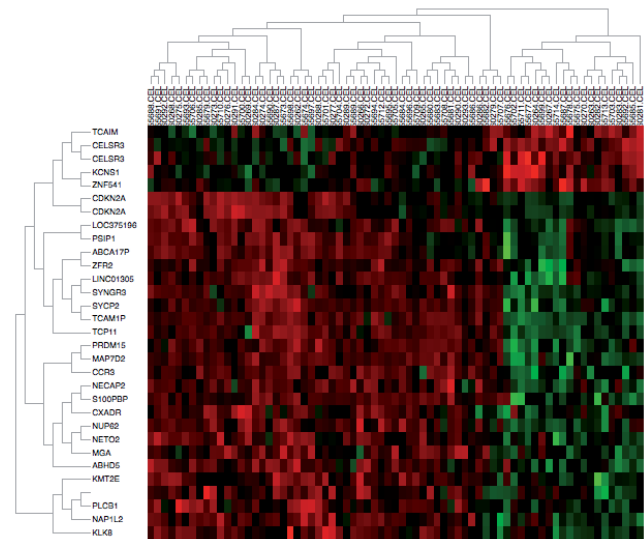
The Differential Results tab displays the results from the differential analysis using empirical Bayes moderated t-statistics with the Bio-conductor Package limma<sup>16</sup>. Statistics are computed on data that have been batch corrected by combining ComBat with SVA, protecting for the biological groups selected for comparison<sup>10</sup>. The p-values are adjusted utilizing the Benjamini-Hotchberg method for multiple hypothesis testing<sup>8</sup>. Optionally, gene set statistics can be performed for each gene set defined in Biocarta and Gene Ontology using a Wilcoxon rank-sum test comparing the t-statistics from the most differentially expressed probe for genes in the set to similarly selected t-statistics for genes outside of the set. If selected, results from gene set analysis are displayed in the GSA Results tab.

### Example

As an example, we perform analysis on sample datasets containing gene expression of primary head and neck squamous cell carcinoma (HNSCC) tumors. We downloaded measurements from a combination of frozen tumor samples from two distinct studies in GEO available under accession numbers GSE10300<sup>3</sup> and GSE6791<sup>12</sup>, representing two distinct batches. Raw CEL files and annotation csv files were obtained as described in the User's manual. We initialize AGA by selecting the directory containing these data. Once loaded, we check the HPV and Tumor.Source.Type columns to group the samples into primary HPV-positive and HPV-negative tumors for differential expression analysis. We then click "Run the Analysis" to normalize the CEL files with RMA<sup>6</sup>, batch correct the data with ComBat and SVA, and perform differential expression analysis. The plot in the Dendrogram Plot tab confirms that the batch effects are apparent between these datasets but removed after batch. The heatmap generated in the Heatmap Plot tab (Figure 1) demonstrates that the batch correction nonetheless preserves gene expression difference between HPV-positive and HPV-negative tumors. Moreover, performing differential expression analysis comparing HPV-positive and HPV-negative HNSCC in the "Differential Analysis" tab confirms the well-established overexpression ( $p=8.74e-9$ ) of CDKN2A (p16) in HPV-positive HNSCC<sup>14,15</sup>.

### Discussion

AGA provides an interface to enable users who may be unfamiliar with R to perform reproducible genomics class comparison analysis. Unlike other automated pipelines, experienced R users can reproduce, extend or modify preliminary analyses. Thus, AGA facilitates collaborations between novice and expert R users for genomics



**Figure 1.** Heatmap displaying the relative expression of the probes with p values below 0.0001 from the example analysis, including CDKN2A. We note that sample names are truncated in the heatmap, but users can reduce the lengths of sample names or ensure that sample identity can be determined by the final characters in the name to associate specific samples with the heatmap.

analysis. Future work will extend the AGA pipeline to encode normalization routines to DNA methylation, and analysis routines for other genomics platforms, including copy number data.

### Software availability

#### Latest source code

<https://gist.github.com/78f566e1a51d745fac3b>

#### Source code as at the time of publication

<https://gist.github.com/F1000Research/9d2acc6aca8ba2d1cc76>

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

<http://dx.doi.org/10.5281/zenodo.1405619>

### License

GNU GPL V2

### Author contributions

MFO and EJJ conceived the software and EJJ and MC designed the web interface. MC designed and coded implemented the software application, and prepared the manuscript. HSP researched and composed cross-study normalization techniques. XXX standardized annotation files for the two example data sets. YW and LC assisted by providing the initial coding for alternative analyses. All authors helped prepare the manuscript.

## Competing interests

No competing interests were disclosed.

## Grant information

Funding was provided from NIH/NIDCR P50DE019032, NIH/NCI P30CA006973, K25CA141053 to EJJ, and NIH/NLM R01LM011000 to MFO.

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

## Acknowledgements

We would like to thank Joe Cheng and Winston Chang of RStudio for their support with shiny. Alla Guseynova, Michael Fox and Louis Franceschi are very much appreciated their technical support and implementations of various iterations of the project. We thank Thomas Considine for his assistance in proofreading this manuscript; and Bahman Afsari and Thomas Considine for testing the application and User Manual. Finally, we also thank Luigi Marchionni and Jean-Philippe Fortin for collaborative efforts.

## Supplementary material

AGA User's Manual: [available here](#).

## References

- 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](#)
- Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet.* 2000; **25**(1): 25–29.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cohen EE, Zhu H, Lingen MV, *et al.*: **A feed-forward loop involving protein kinase Calpha and microRNAs regulates tumor cell cycle.** *Cancer Res.* 2009; **69**(1): 65–74.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Downey T: **Analysis of a multifactor microarray study using Partek genomics solution.** *Methods Enzymol.* 2006; **411**: 256–270.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gautier L, Cope L, Bolstad BM, *et al.*: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics.* 2004; **20**(3): 307–315.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Irizarry RA, Hobbs B, Collin F, *et al.*: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics.* 2003; **4**(2): 249–264.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics.* 2007; **8**(1): 118–127.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Klipper-Aurbach Y, Wasserman M, Braunsiegel-Weintrob N, *et al.*: **Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus.** *Med Hypotheses.* 1995; **45**(5): 486–490.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Leek JT, Johnson WE, Parker HS, *et al.*: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics.* 2012; **28**(6): 882–883.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Parker H: **Practical Statistpractical Statistical Issues in Translational Genomical** *Issues in Translational Genomics (doctoral dissertation).* Johns Hopkins University, Baltimore. 2013.
- Pirooznia M, Seifuddin F, Goes FS, *et al.*: **SVAW - a web-based application tool for automated surrogate variable analysis of gene expression studies.** *Source Code Biol Med.* 2013; **8**(1): 8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pyeon D, Newton MA, Lambert PF, *et al.*: **Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers.** *Cancer Res.* 2007; **67**(10): 4605–4619.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- RStudio and Inc: **shiny: Web Application for R.** R package version 0.7.0. 2012.  
[Reference Source](#)
- Robinson M, Sloan P, Shaw R: **Refining the diagnosis of oropharyngeal squamous cell carcinoma using human papillomavirus testing.** *Oral Oncol.* 2010; **46**(7): 492–6.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Smeets SJ, Hesselink AT, Speel EJ, *et al.*: **A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen.** *Int J Cancer.* 2007; **121**(11): 2465–72.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol.* 2012; **3**: 1. Article3.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Subramanian A, Tamayo P, Mootha VK, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A.* 2005; **102**(43): 15545–15550.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Taminau J, Steenhoff D, Coletta A, *et al.*: **inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO.** *Bioinformatics.* 2011; **27**(22): 3204–3205.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Considine M, Parker HS, Wei Y, *et al.*: **Automated Genomics Analysis.** *Zenodo.* 2015.  
[Data Source](#)

# Open Peer Review

Current Referee Status:  

## Version 1

Referee Report 22 June 2015

doi:10.5256/f1000research.6456.r8835



**Matthew McCall** 

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

The authors describe a software package for interactive (via a shiny webapp) genomic analysis. By running R behind the scenes, this software addresses a common challenge in genomic data analysis -- the transition from simple initial analyses (typically performed by a novice user) and more complex later analyses (typically performed by an advanced user). When the initial analyses are not easily examined / reproduced, the advanced user often must start from scratch. The AGA software will hopefully address this issue.

The title of the article is currently too broad -- the software is only able to handle Affymetrix expression arrays, Illumina 450k methylation arrays, and normalized RNA-seq gene counts. However, I trust that the authors will expand the functionality of the software to handle many other platforms and types of genomic data.

My primary criticism of this work is that I was unable to successfully use the package. The package depends on a large number of other packages (shinyIncubator, googleVis, and heatmap among others). In particular, I was unable to install the heatmap package. Additional instructions on how to obtain / install all of the required dependencies should be added to the user manual.

I also have a few minor criticisms of the article:

1. In Figure 1, the gene names are not legible due to over-plotting, and the column names are truncated.
2. The citation for the fRMA method is incorrect. The correct citation is:

McCall MN, Bolstad BM, and Irizarry RA (2010). Frozen Robust Multi-Array Analysis (fRMA), *Biostatistics*, 11(2):242-253.

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

**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.**

Author Response 18 Sep 2015

**Michael Considine**, Johns Hopkins University, USA

1) *The title of the article is currently too broad -- the software is only able to handle Affymetrix expression arrays, Illumina 450k methylation arrays, and normalized RNA-seq gene counts. However, I trust that the authors will expand the functionality of the software to handle many other platforms and types of genomic data.*

The title has been revised to "AGA: Interactive pipeline for reproducible gene expression and DNA methylation data analyses"

2) *My primary criticism of this work is that I was unable to successfully use the package. The package depends on a large number of other packages (shinyIncubator, googleVis, and heatmap among others). In particular, I was unable to install the heatmap package. Additional instructions on how to obtain / install all of the required dependencies should be added to the user manual.*

The number of packages is necessary for the diverse functionality and formatting, and to make the best possible application. We have added instructions to manually install the required packages on page 4 of the revised user's manual.

"In the event of difficulties installing libraries in R, copy lines 144 to 200 in the global.r script and enter them into your R console. Afterwards, run the command: sporelibs()"

*I also have a few minor criticisms of the article:*

3) *In Figure 1, the gene names are not legible due to over-plotting, and the column names are truncated*

We have revised Figure 1 to reduce the number of genes so that the gene names are legible. The column names are set by the sample identifiers and must be truncated to facilitate visualization. We note in the revised caption to Figure 1 that: "We note that sample names are truncated in the heatmap, but users can reduce the lengths of sample names or ensure that sample identity can be determined by the final characters in the name to associate specific samples with the heatmap".

4) *The citation for the fRMA method is incorrect. The correct citation is:*

By default, AGA implements RMA instead of fRMA. We have removed this citation from the revised manuscript and clarified this choice in by adding this second sentence of Initiation subsection of the revised manuscript: "For gene expression microarrays, AGA performs RMA normalization implemented in the Bioconductor package affy" and revising the fourth sentence of the Example subsection of the revised manuscript "We then click "Run the Analysis" to normalize the CEL files with RMA"

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

Referee Report 02 February 2015

doi:10.5256/f1000research.6456.r7514



**Subha Madhavan**

Innovation Center for Biomedical Informatics, Georgetown University, Washington, DC, USA

The authors have described Automated Genomics Analysis (AGA), an interactive program to analyze high-throughput genomic data sets on a variety of platforms.

The software is implemented in R with web app using Shiny.

Specific comments are noted below:

1. cBIOPortal is listed as an example for reducing cost of genomic analysis using AGA. cBioPortal's purpose is to help researchers mine analyzed results and it is available for free for non-commercial use. cBIOPortal is not in the same class of software as AGA and is an inappropriate comparison.
2. Title needs to be changed - current AGA software supports expression and methylation analysis only. The title is very broad, especially given that there is not support for genomic variant analysis in the software.
3. Describe any quality checks performed on CEL and idat files briefly beyond batch correction. How does the software deal with missing values?
4. Address scalability. How does the software scale for large studies?

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

**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.**

Author Response 18 Sep 2015

**Michael Considine**, Johns Hopkins University, USA

*1) cBIOPortal is listed as an example for reducing cost of genomic analysis using AGA. cBioPortal's purpose is to help researchers mine analyzed results and it is available for free for non-commercial use. cBIOPortal is not in the same class of software as AGA and is an inappropriate comparison.*

We have removed reference to cBIOPortal in the revised manuscript.

*2) Title needs to be changed - current AGA software supports expression and methylation analysis only. The title is very broad, especially given that there is not support for genomic variant analysis in the software.*

The title has been revised to "AGA: Interactive pipeline for reproducible gene expression and DNA methylation data analyses"

*3) Describe any quality checks performed on CEL and idat files briefly beyond batch correction. How does the software deal with missing values?*

We have revised the Initiation subsection of the Methods section to note that: "For gene expression microarrays, AGA performs RMA normalization implemented in the Bioconductor package *affy*<sup>5</sup>. Probe-level estimates of DNA methylation are computed from iDat files using Illumina standards with the *minfi* package<sup>1</sup>. RNAseq data are formatted as individual text files for each sample, assumed to contain gene names and normalized counts for each sample.



4) *Address scalability. How does the software scale for large studies?*

We have revised the Introduction section to include more information on the length of analyses, “The runtime of analyses will depend largely on the desktop hardware, but also on the data platform and optional analyses selected. On a Mac Pro workstation, containing a 3.2 GHz Quad-Core Intel Xeon processor and 10Gb 1066 MHz DDR3 RAM, analyses containing under 100 samples were completed in under 30 minutes.”

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

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

F1000Research