**Dear PLOS ONE editor.**

**We would like to thank the reviewers for their helpful comments, which have helped improve the paper. Major changes in the text are highlighted in yellow, while individual comments are replied to in the answer below. Furthermore, a new author, Rasmus F Brøndum, has been added since he has worked on the implementation and analysis of the calculation of RLE values for one-by-one normalized arrays.**

**On behalf of the authors**

**Martin Bøgsted**

Response to Reviewers

Reviewer #1: The manuscript by Falgreen et al. is well motivated and describes a research tool that could be of great value to the community. The website and analytical tools available are well developed. However, important issues regarding technological and cohort bias are given little attention in the manuscript, and completely absent from the web tool. The combination of ease of use and lack of attention to analytical detail will only create confusion as erroneous calls will proliferate.   
  
The current manuscript provides no method to assess whether the underlying assumptions are correct, and thus that the classifications are accurate. This is true for all the classifiers that are implemented. The authors must clearly state assumptions, suggest methods for assessment of technological and biological variation, and demonstrating the importance of these factors in both the manuscript and on the website.   
  
Primary criticisms:  
  
1. The one-by-one and reference based methods both carry assumptions which are not considered in detail by the manuscript. Calls in one-by-one normalization can be heavily biased by batch effects. Batch effects are typically probe specific, and thus the sample normalization will not remove them. The authors acknowledge this by proposing the reference normalization approach, and show that it is clearly superior with their reference set. However, if no reference set is available, then how does a researcher assess whether the assigned classification is accurate? A simple proposal would be to plot results of PCA to understand if the training and test cases are similar. Another possibility would be to calculate the RLE mean and IQR and consider invalidating a one-by-one classification if these statistics are too extreme.

**AU: We have implemented calculation of the RLE into HemaClass and tested it’s validity for distinguishing between a “proper” and “improper” user reference. We find that an RLE IQR of 0.6 serves as a reasonable threshold for this purpose, although it is not foolproof. This is now mentioned in the discussion and more detail is given in supplementary section S6. Furthermore, we now make a more modest evaluation of the results obtained with an ExLab one-by-one RMA reference and more clearly emphasize that users should supply their own reference for valid results.**  
  
2. My criticism of the reference based method is similar - there is no way to assess whether a reference set adequately represents the training set. This is especially important since the method performs scaling of the probeset estimates to have similar mean and variance between the training set and the reference set. This method is very reasonable, but the authors should acknowledge the underlying assumption that the training and reference sets are random samples of the same population. When this assumption is not true, then the reference set will greatly bias the normalization and result in erroneous classification. The authors should clearly state this and any other assumptions of the reference based scaling that is performed, should demonstrate the importance of this assumption experimentally, and implement methods to assess the reliability of a given reference set.

**AU: We acknowledge that this is a reasonable question to ask. However, in these types of analyses it is not possible to expect that the validation data (i.e. GEP from patient data) represents the training data (i.e. GEP from sorted normal tissue) as the training data is based on very different tissue as well as subpopulations sorted and profiled under very well controlled conditions. One should recall that the process is that we find the subpopulation which looks most similar to the patient data. When that is done, one investigates whether the stratification method has prognostic potential. The rationale behind this is then validated by biological analyses both *in silico* as well as in the wetlab. If “the classification” turns out to be reasonable one “forgets” about the training data and we only need to validate against the cohort based results. This procedure is well known from other similar cell of origin approaches, see e.g. Rosenwald et al. (2002) New England Journal of Medicine 25:1937—1947 for an ABC/GCB example and  Figure 1C of Rapin et al. (2013)Blood 123:894—904 for an example in subclassification of AML.**

**In these types of analyses one has to trust the right tissue has been extracted and handled correctly through all the steps in the laboratory ending up with a reference array data set, which has the sufficient quality. The latter could be ensured by having a central tissue bank with officially approved data by e.g an international consortia for the specific diseases. Our reference data have for instance been controlled by looking at the frequency of ABC/GCBs, BAGS classes and their survival curves as well as tissue control by experienced pathologists.**

**As long as the reference set is trustworthy one could use the RLE procedures outlined above to assess the quality in a one-by-one manner. These assumptions have now been clearly stated in the paper.**

3. Table 3 measures one-by-one and reference based assignments against the cohort based assignment. I understand the nature of this comparison, but these measures should also be compared to the gold standard assignment as in Table 2.

**AU: We have added comparisons between Wright ABC/GCB classifications on cohort normalized data and ABC/GCB classifications obtained by applying the elastic net classifier on one-by-one normalized data.**  
Secondary:  
  
1. The one-by-one RMA procedure was previously described as fRMA BY McCall et al. (Biostatistics, 2010) and later described by Piccolo et al., (Genomics, 2012). One or both of these should be cited as appropriate

**AU: We have added references to previous approaches, and discussion of an application of the fRMA approach.**  
2. The manuscript has numerous typos, missing words, changes in tense, etc - too many to list as this point. One of note is a lack of reference to table 2 in the text that describes these results.

**AU: We have added a reference to table 2 in the text, and proofread the manuscript to clear up typos and grammatical errors.**  
  
3. How many genes are used in the REGS and BAGS classifiers?

**AU: We have added a table (S5) to the supplementary material with the number of probes and corresponding gene annotations for each of the classifiers used in the manuscript.**

4. The conclusion states that the method is for 'microarrays', but the current implementation only applies to a few specific Affymetrix chips. More specificity (or broadening of acceptable inputs to the tool) is needed here.

**AU: The conclusion in the text does not mention microarrays, while the suggestion in the abstract-conclusion that the proposed one-by-one pre-processing method is relevant for all researchers using microarrays is valid. Using the functions implemented in HemaClass it is indeed possible to build and use a user reference for other arrays than the u133+2, but you are correct that the classifiers only works with one specific array.**  
  
5. If an Affymetrix array doesn't have one of the classifier genes, then is the model re-trainined without that gene? Or is it imputed from the training set? or can a value be missing? How does this affect classification performance?

**AU: If expression levels for a given probe is missing it is set to zero (median). We have not tested how this affects classification performance**  
  
6. Figure 1 does not seem to accurately reflect the description of the workflow in the text. For instance, does the 'Reference' label refer to the reference set one would input, or to the training sets?

**AU: We have altered figure 1 to give a more intuitive representation of the workflow on hemaclass.org**  
  
7. Figure 3 appears to be mislabeled

**AU: We have replaced the letter labelling in the figure**

**Reviewer #2:**

1. This is an extended work from the authors' earlier publications; It provides a useful and user-friendly tool for non-bioinformatician or non-biostatistician to assess hematological patient data from microarray in terms of drug sensitivity and subtype classification;

2. I have tested the usefulness of some functionalities of the online tool. It appears easy to use and user friendly. In terms of classification, it is accurate for class ABC/GCB, but not NC. Though this limitation is adequately discussed in the manuscript.

3. The term ‘Agreement’ across the manuscript should be replaced by ‘Accuracy’ to avoid confusion with the Cohen’s weighted k, which is another measurement of ‘agreement’;

**AU: We have replaced the term.**

4. Fig. S1B shows log(λ) = -7.29 but it is stated as -7.41 in the main text.

**AU: We have replaced the value in the text**

5. The resolution for Figure 2 and 3 is poor. The labeling of figure 2 and 3 are not readable. The letter labeling (A, B, C…) of sub-figure in Figure 3 is incorrect.

**AU: We have redone all figures in higher resolution and corrected the letter labelling in Figure 3.**

6. The figure and table numbers are not labeled carefully in the supplemental figures and tables;

**AU: We have corrected the labelling in the supplementary.**

7. The limitations were adequately discussed. However, it is well known that RNAseq is gaining ground against microarray. It is recommended that the authors should discuss the possibility of extension of this work to RNAseq data, which is different type of data from microarray;

**AU: We have added a paragraph about RNA-seq and possibilities for adapting HemaClass to this technology in the discussion.**