Author's response to reviews

Title: A novel SNP analysis method to detect copy number alterations with an unbiased reference signal directly from tumor samples

Authors:

Alex Lisovich (all67@pitt.edu)
Uma R. Chandran (chandranur@upmc.edu)
Maureen A. Lyons-Weiler (lyonsweilerma@upmc.edu)
William A. LaFrambiose (laframboisewa@upmc.edu)
Ashley R. Brown (arg39@pitt.edu)
Regina I. Jakacki (regina.jakacki@chp.edu)
Ian F. Pollack (ian.pollack@chp.edu)
Robert W. Sobol (rws9@pitt.edu)

Version: 2 Date: 24 November 2010

Author's response to reviews: see over
Dear Drs. Pujana and Ghosh:

Please find enclosed our revised manuscript (MS: 2639767364203720) entitled, “A novel SNP analysis method to detect copy number alterations with an unbiased reference signal directly from tumor samples” for publication in BMC Medical Genomics by Alex Lisovich, Uma R. Chandran, Maureen A Lyons-Weiler, William A. LaFramboise, Ashley R. Brown, Regina I. Jakacki, Ian F. Pollack and Robert W. Sobol.

In this study, we describe a new and novel algorithm allowing generation of an internal reference signal for the detection of genome copy number alterations (CNA) directly from the test samples of a given high-density (250K) SNP analysis, thus eliminating the need for acquiring a normal reference set. This new approach, referred to as Virtual Normals (VN), allows for the construction of an unbiased reference signal, free of CNA.

In response to your Editorial comments and Reviewer concerns, we have revised the manuscript accordingly and have provided a detailed, point-by-point response to those comments (to follow). We think you will find we were able to effectively respond to each comment and concern, making corrections, changes and additions as requested. We look forward to hearing from you with regard to publication.

Sincerely,

Robert W. Sobol, Jr., PhD
Assistant Professor
Departments of Pharmacology & Chemical Biology and Human Genetics
Co-Director, UPCI Vector Core Facility
Director, UPCI Lentiviral Facility
University of Pittsburgh Cancer Institute
Response to Editorial comments and Reviewer concerns

We are pleased that the reviewers suggested our manuscript is “an article whose findings are important to those with closely related research interests”.

In response to the Editorial comments and Reviewer concerns, we have revised the manuscript accordingly and have provided a detailed, point-by-point response to those comments below.

MS: 2639767364203720

Title: A novel SNP analysis method to detect copy number alterations with an unbiased reference signal directly from tumor samples

Authors: Alex Lisovich, Uma R. Chandran, Maureen A Lyons-Weiler, William LaFramboise, Ashley R. Brown, Regina I. Jakacki, Ian F. Pollack and Robert W. Sobol

Editorial comments

1) Please structure your abstract according to our guidelines for medical journals within the BMC series: Background: This should place the study into the context of the current knowledge in its field and list the purpose of the work; in other words, the authors should summarise why they carried out their research. Methods: This section should summarize how the study was performed and mention the different techniques employed. Results: This section should describe the main findings of the study. Conclusions: A brief summary of the content of the manuscript and the potential implications of its results.

Response: Thank you for bringing this to our attention. The manuscript has been modified accordingly.

2) Informed consent must be documented. Manuscripts may be rejected if the editorial office considers that the research has not been carried out within an ethical framework, e.g. if the severity of the experimental procedure is not justified by the value of the knowledge gained.

Response: Thank you for bringing this to our attention. The manuscript has been modified accordingly (See Data Set, lines 3 and 4).

3) Competing interests - Please include a 'Competing interests' section between the Conclusions and Authors' contributions. If there are none to declare, please write 'The authors declare that they have no competing interests'. The questions that are asked of authors are:
   Financial competing interests
   - In the past five years have you received reimbursements, fees, funding, or salary from an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future? Is such an organization financing this manuscript (including the article-processing charge)? If so, please specify.
   - Do you hold any stocks or shares in an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future? If so, please specify.
   - Do you hold or are you currently applying for any patents relating to the content of the manuscript? Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript? If so, please specify.
Do you have any other financial competing interests? If so, please specify.
Non-financial competing interests
Are there any non-financial competing interests (political, personal, religious, academic, ideological, intellectual, commercial or any other) to declare in relation to this manuscript? If so, please specify.

Response: Thank you for bringing this to our attention. The manuscript has been modified accordingly.

4) Data availability and deposition - Nucleic acid sequences, protein sequences, and atomic coordinates should be deposited in an appropriate database in time for the accession number to be included in the published article. In computational studies where the sequence information is unacceptable for inclusion in databases because of lack of experimental validation, the sequences must be published as an additional file with the article. Where appropriate, authors should adhere to the standards proposed by the Microarray Gene Expression Data Society (http://www.mged.org) and must deposit microarray data in one of the public repositories, such as ArrayExpress (http://www.ebi.ac.uk/arrayexpress), Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/projects/geo/) or the Center for Information Biology Gene Expression Database (CIBEX; http://cibex.nig.ac.jp).

Response: Thank you for bringing this to our attention. The manuscript has been modified accordingly (See Data Set, last paragraph) and the data is being deposited with GEO. We will forward a link once the deposit is finalized.
Lisovich et al. described a novel algorithm called “Virtual Normal” for constructing a reference baseline for chromosomal copy number variation detection, using mainly the Affymetrix 250K arrays as examples. The algorithm can potentially be used to fix the “batch effect” commonly seen with such data, and to address the practical limitations of not having the perfect reference samples or enough number of them. This is a typical problem faced by small laboratories with limited resources or cohorts with limited sample size. They demonstrated the effectiveness of their method with convincing results. The paper is well written and algorithm is clearly explained and justified. BMC Medical Genomics is a proper journal for its publication, which may benefit the biomedical research community and is recommended by the reviewer. The reviewer though does have a couple of minor comments, suggestions, or concerns.

Minor Essential Revisions
1. The noise reduction process described can potentially eliminate two types of variations: 1) batch effect, and 2) common variations among tumor samples. It is not 100% clear as to how or whether or not they are differentially treated. It would be contradictory to the goals of many projects if the method will not allow detecting these common variants. Again, it is likely that it is described somewhere or the reviewer did not fully understand it. But it might argue for a better or different way to explain this.

Response: The noise reduction process described in our manuscript can eliminate the high frequency batch component. However, it is specifically designed to preserve the common variations among tumor samples thus treating these two types of variations differentially as discussed in detail in the Methods (Methodological Considerations) section and illustrated in Fig. 1 and Fig. 2 (revised). In summary, the purpose of the algorithm is to generate a synthetic reference signal with two key properties:

a) Its high frequency component is highly correlated with the one common for all tumor samples in the given batch and therefore the noise in (log2) ratio between a given tumor raw CN and the reference signal is reduced.

b) The reference signal does not contain the CN aberrations common for the part or the whole tumor sample set, which means that when each individual tumor sample is compared to the reference signal, the aberration will be picked up for this particular tumor.

Specifically (see Methods, Algorithm description section), the algorithm starts from creating the reference signal from tumor samples (see eq. 1a). Indeed, if directly used as a reference, such a signal would eliminate the common variations among tumor samples. However, the next step in the algorithm includes detection of common variations among tumor samples present in this reference signal (eq. 2 and 3) and then modifying it in such a way that the common variations in the resulting synthetic reference signal are eliminated (eq. 4). The resulting raw CN for each tumor signal is then computed as the (log2) ratio between the given tumor intensities and the synthetic (VN) reference signal described above (see eq. 5). Here the reference signal (denominator) does not contain common variations along tumor samples, and therefore the raw CN signal for each tumor sample will preserve the sample specific CN variations.

The process of generating this synthetic signal is also illustrated on Fig. 1 where Fig. 1b shows the aberrations in the synthetic reference signal which are common among tumor samples as detected by the first stage of the algorithm (eq. 1a and eq. 2) while Fig. 1c shows the results of the final stage of the synthetic reference signal generation where these common variations were eliminated. Fig. 1d then shows the final results of processing a particular sample where CN variations specific for a particular sample were preserved.

For factual evidence of the ability of the VN algorithm to detect CN aberrations common for most samples, see Fig. 5 as well as similar plots for chromosomes 1 – 22 in supplemental materials.
(Figure S3 and S6), where we compare the segmentation results obtained using the VN algorithm with the results utilizing an in-lab FFPE normal reference set acquired recently during the validation stage. We have rewritten the “Methodological Considerations” section to make the description of the overall procedure more clear.

2. The data used for the study was generated using a rather quite outdated platform, the Affymetrix 250K genotyping arrays. Conceptually, the algorithm described can be likely expanded and applied to other similar types of data, such as Affymetrix 500K and Illumina 550K HumanHap, or data from the latest genotyping arrays. However, it would be nicer if the author can demonstrate that. It shouldn’t be hard to obtain such data sets given the large number of CNV publications in the last a couple of years, many of which used the newer technologies.

**Response:** We were motivated to develop the algorithm for this particular study using the Affymetrix250K_Sty micro array chip, one of two chip types (the second being Affymetrix250K_Nsp) constituting the Affymetrix 500K platform. We agree that using the words “Affymetrix 250K” without the proper suffix could potentially lead to confusion and have added the suffix “_Sty” to the “Affymetrix 250K” throughout the text. This algorithm can be used on each of the 250K chips alone or the 500K data. We have not formally tested the algorithm on Illumina. However, it was successfully tested on the Affymetrix Genome-Wide 6.0 platform. We agree that it would be beneficial to mention these results and have expanded the “Discussion” section describing these processing results and provided the corresponding supplemental data (Figure S7 A, B and C).

3. As the authors recognized, the biggest advantage of genotyping arrays over conventional arrayCGH platform, for copy number detection, is that they offer another type of data or metric that can be used for improved accuracy with CNV calls, namely the genotyping data or allelic ratios. However, it is not explained or made clear as to how this type of data can be incorporated into their method. While the authors stated that it will be their future plan, it is hard to imagine that it will indeed happen given the likelihood of the genotyping technologies will soon be totally phased out given the increased capacity and throughput with the high-throughput sequencing technologies.

**Response:** The VN algorithm itself does not utilize the genotyping call results and allele specific CN estimates. We have clarified in the text (in Conclusions) that the VN is applicable for allele specific CN estimates and that using the genotyping call results can aid in allele specific CN detection. We emphasize that the allele specific CN estimates can benefit from VN. We mentioned our future plans to use VN for allele specific estimates and to use genotyping call results to expand the aroma.affymetrix package. We realize that high-throughput sequencing technologies have great potential, however we believe that genotyping technologies will be around for a while at least until high-throughput sequencing reaches maturity and long after that taking into account the vast amount of data accumulated by the scientific community using genotyping technologies. In Conclusion, we emphasize that the described algorithm is general enough to be applied to any processing pipeline which utilizes a reference signal derived from a normal reference set.
Reviewer's report #2

The authors present a novel method for calculating a reference signal to remove the experimental bias from microarray probe hybridization intensities with the purpose of detecting copy number on cancer cells. This is important, because obtaining a set of normal samples with the same experimental biases is not always possible. The method proposed by the authors in the paper offers a very original and relatively simple solution to the problem that should be very useful for other researchers. In particular, the authors demonstrate that it should be especially useful for dealing with FFPE samples in older repositories which are much more noisier and usually do not have paired normal samples, or available reference normal samples with a similar bias. These FFPE samples are important in cancer types that are rare and it is very difficult to obtain new fresh samples. In any case, the proposed method also appears to be also useful for newer samples obtained from frozen/fresh tissues. The idea behind the approach could also be imported in different copy number calling pipelines. I have some minor change suggestions that I think can improve the paper.

Minor Revisions
1) Section “Methodological Considerations” and “Unbiased Reference Signal Restoration” would be easier to follow if along the text was making more references to the relevant parts of Figure 1 and 2 which highlight the overall procedure.

Response: We have rewritten the “Methodological considerations” section, inserting the references to Fig. 1 to make the description of the overall procedure more clear.

2) Figure 5, the meaning of the red and green lines should be on the caption.

Response: We have modified the caption of Figure 5 as follows:
Figure 5 - Comparison of detection results on Chr. 7. The red and green segments represent detected amplifications and deletions, correspondingly.
A: Using FFPE in-lab normals.
B: Using the VN reference set. The 7p13 region was detected using VN algorithm only.

3) Figure 6, the “08_” portion should be added on the plot, for more clarity.

Response: Thank you for bringing this to our attention. Figure 6 has been modified accordingly.

4) Table 1, the column headings with line breaks are confusing.

Response: Thank you for bringing this to our attention. We have reconfigured the headings of the Table 1 to make it more readable and changed the caption to address the reviewer’s revision 8 (see below).

5a) Equation (1b), ‘i’ should probably be ‘j’

Response: Thank you for bringing this to our attention. We have corrected this discrepancy and the manuscript has been modified accordingly.

5b) Variable names and equations inside the main text should appear in italics for more clarity.

Response: Thank you for bringing this to our attention. We have followed the reviewer’s advice and changed the fonts for variable names and equations to italics.
6) Equation (4), super-index ‘vn’ seems to be later changed to sub-index ‘v’. I think all are referring to the same thing.

**Response:** Thank you for bringing this to our attention. We have corrected this discrepancy and made the notation consistent among all equations.

7) Is the alteration on Chr 4 chosen for PCR validation only detected using the VN approach?

**Response:** It was detected by all approaches.

8) How is the signal-to-noise ratio calculated exactly?

**Response:** We have modified the caption of the Table 1 to include the definition of the signal to noise ratio as follows:

Table 1 - Comparison of a standard deviation (SD) and signal to noise (S/N) ratio for a raw CN signal obtained using different types of references. Here SD is an average of all SDs obtained for each raw CN segmented region in log2 scale, and S/N is computed as a ratio of the difference between the raw CN values corresponding to CN=3 and CN=2 in log2 scale and the SD value defined above.

Discretionary Revisions
8) I wonder how the VN method would perform if you use a different template for normal instead of the HapMap reference panel. For example, for the FFPE samples what happens if you use the FFPE normal as the template for the VN.

**Response:** We performed a few such experiments using various combinations of cell culture and preservation protocols utilizing the filtering based version of the VN algorithm. When the results of VN on FFPE tumors are compared for two different templates, one of which is HapMap and the other is an in-lab FFPE based sample, the signal to noise ratio stays roughly the same, but the smoothing window size used to generate the second VN reference can be reduced which indicates that spatial resolution for regions common for a larger part of a tumor sample set is increased for an in-lab based template.

9) I think the noise reduction and the signal-to-noise ratio are already good indicators of improvement of performance, but it may be interesting to complement the results with more validation. For example: i) find if there are more PCR validated alterations for the VN approach than for the other approaches, or ii) instead of PCR validations one could simulate alterations by sampling portions of the ‘X’ chromosome for males (or from some other known alterations).

**Response:** We agree that other validation approaches may provide additional information. We selected the PCR validation and the targets as “proof of concept”. In continued investigation of this dataset, other regions will also be selected for validation. Simulations of the X regions were also performed and we planned initially to describe the results in a way similar to Bengtsson et al.,” Estimation and assessment of raw copy numbers at the single locus level”, but have decided that such validation on a very large region would be largely equal to validation of a general principle that signal to noise ratio reduces the quantization error considering CN=0,1,2... as quantization levels. We’ve made the decision that description of these results is beyond the scope of this manuscript.

Reviewer's report #3
This article gives a way of normalizing tumor samples using normal samples not produced in the same lab. This is applicable when a lab only process tumor samples, and does not have a normal reference to provide the baseline measurement for each probe (or probe-set). The idea is that the unique conditions in each lab produce lab-specific probe effects, which needs to be accounted for in copy number estimation. If a normal sample from that lab were available, then it could be used as a baseline of comparison for tumor samples.

1) The paper suggests that when such a normal control is unavailable, the tumor can be first segmented, the median across many tumors samples computed and normalized (using an external reference sample), and that be used as a reference instead of a “normal control”.

Response: We feel that the method described in the paper differs substantially from the reviewer’s interpretation. Therefore, we would like to clarify further. The expanded description follows:

The set of tumor samples and the set (not a single one) of normal samples obtained from another batch, lab or using the different preservation protocol (frozen vs. FFPE as an example) are preprocessed and normalized using standard methods available in the aroma.affymetrix R package as outlined in the Methods, generating the VN reference signal.

Further, the conventional reference signal is computed by taking the median intensity across all normal arrays for each SNP position (see eq. 1b). In conventional processing, the next step would be to compute the (log2) ratio between given tumor sample intensities and the abovementioned reference signal (which gives what is known as raw CN signal) followed by segmentation of such raw CN signal for each tumor sample.

We observe, however, that the reference signal obtained from the set of normal samples, carries the additional high frequency component uncorrelated to the given set of tumor samples due to the unknown batch effect and/or preservation protocol differences which leads to an increased noise even when standard methods of normalization are applied during the preprocessing step and all known batch effects are compensated for.

Further, we observe that if the average (median) of all tumor samples is used, in place of a standard reference signal (eq. 1a), this could reduce the high frequency component due to the high correlation between such reference and each of the tumor sample signals. However, such a signal would carry the CN aberrations common for the part or the whole tumor sample set and therefore if used unmodified as a reference, such common CN variations would not be detected.

To overcome this problem, the proposed algorithm computes the ratio between the tumor based reference signal and normal sample set based reference signal (eq. 2), detects the aberrations in the resulting signal (eq. 3) and then modifies the tumor-based reference signal in such a way that these common aberrations are eliminated relatively to the normal set based reference signal (eq. 4). The resulting synthetic signal is used to generate the raw CN signals for each tumor sample (eq. 5). This synthetic reference signal has two key properties:

i. It’s high frequency component is highly correlated with the one common for all tumor samples in the given batch, hence the noise in (log2) ratio between a given tumor raw CN and the reference signal is reduced

ii. The reference signal does not contain the CN aberrations common for the part or the whole tumor sample set, which means that when each individual tumor sample is compared to the reference signal, the aberration will be picked up for this particular tumor.
It should be emphasized that in our paper we present two forms of eq. 3 which describes the process of common CN variation detection: the one which is using the mean preserving segmentation to detect these common variations in synthetic reference signal and the second one based on the filtering which is presented right after the eq. 5. In Methods, Unbiased Reference Signal Restoration and in the end of Algorithm Description sections we discuss the tradeoffs presented by these two approaches as well as conditions under which using one or another would be preferable.

To address these concerns: We have expanded the Methods (Methodological Considerations) section to make the description of the overall procedure more clear and added an additional block to Figure 2 (Algorithm block-scheme) which depicts the preprocessing and normalization step performed before the VN is applied.

2) I feel that the method in the paper simply does not work, except in special cases delineated by their examples. The “median” is not a robust estimator of the normal state among tumor samples. For example, “recurrent” changes usually have biological significance. But if a position were deleted in more than half of the tumors, then the normalized samples would lose the deletion, and this signal would be lost!

Response: Our method does not use the median as an estimator of the normal state among tumor samples (see the first response). Instead, it assumes the normalization is already done using standard methods (which ensure that the median over all SNP positions is nearly identical for all samples and represents the normal state) and uses the (log2) ratio between the median of the tumor sample signals (synthetic reference signal, see eq. 1a) and the median of the normal samples (the latter being just the conventional reference signal, see eq. 1b) to detect the aberrations common for a group of samples and further transforms the synthetic reference signal in such a way that these common aberrations are removed from synthetic reference. Consequently, each particular tumor sample for which the (log2) ratio against this synthetic reference is computed will keep all individual CN aberrations intact even if this particular aberration is common for a large group of samples. To demonstrate the ability of the VN algorithm to detect CN aberrations common for the large part of samples, see Fig. 5 as well as similar plots for chromosomes 1 – 22 in supplemental materials (Figure S3 and S6), where we compare the segmentation results obtained using the VN algorithm with the results utilizing in-lab FFPE normal reference set acquired recently during the validation stage.

3) Also, the method does not correct for batch effects, which usually contributes to a significant portion of the systematic experimental noise.

Response: The method is applied after standard normalization is done which corrects for known batch effects including systematic experimental noise (see the first response). The method corrects for a high frequency pattern present across a given batch not removed by standard normalization methods. As an example, please compare Fig. 1a showing the raw CN and segmentation results obtained after standard normalization was performed across all arrays (including the normals) using the HapMap normal set as a reference, and Fig. 1d where the same normalized data were processed using the VN-based synthetic reference signal. For normalization we used the standard algorithms available in the aroma.affymetrix R package, but we have also processed data using CNAG and Affymetrix SNP analysis software and obtained results similar to those shown on Fig. 1a.
4) Finally, the method relies on applying existing segmentation tools to un-normalized data as the first heuristic step to identify the normal regions. Most segmentation tools are very unstable when applied to un-normalized data, and this would corrupt the downstream results.

**Response:** We emphasize that we apply our method to the already normalized data, and perform segmentation as a very final step of processing as in the traditional approach, but instead of using the reference signal derived from the normal reference set we use a synthetic reference signal generated by the VN algorithm. The segmentation step described in the paper is used to detect and remove the common CN variations in the synthetic reference signal and we present two alternative forms of eq. 3 as emphasized at the end of the Response 1. Eq. 4, which describes the back shift procedure, does not rely on segmentation and can use any type of filtering. We’ve implemented, tested and described two different versions of the algorithm, one using the segmentation if noise is low and a second using the filtering when noise is high and discuss the trade-offs between the higher resolution of the first approach compared to the reduced false positive rate for the second. During the processing of our data (tumor tissue) we used the filter based version of the algorithm (segmentation was not used at all), and we are using the segmentation when processing samples with relatively low noise and the same preservation protocol (blood samples as an example). Other methods of region detection like a wavelet based approach could also be applied.

5) Although it is not common to have matched normal and tumor samples, it is routine (and not terribly expensive) for a lab to process one normal reference sample along with their tumor samples. In that case, the methods in this paper would not apply.

**Response:** We feel that processing a single normal reference sample along with tumor samples would not be adequate and would have substantial limitations. In fact, the minimal number of samples considered to be necessary to generate the reference in SNP analysis is preferred to be at least 8 (CNAG, dChip etc) which may become prohibitively expensive if part of every study the use of more than 40. Any CN aberration due to the normal mutation in a single normal sample would lead to false positives in all tumor samples processed using such a reference. The ‘at least N’ condition from above ensures that the effect of such a random aberration gets reduced proportionally to the square root of the number of normal samples, i.e. would not impart the normal CNV detection. In our method the synthetic reference signal is generated using the number of samples equal to the tumor sample set size (definitely more than one), and therefore is large enough to reduce the random noise the same way the standard normal reference set does. As was emphasized above, the synthetic reference signal does not contain the aberrations common for the part or the whole tumor reference set and therefore has the same properties in this regard as the standard reference signal.