Author's response to reviews

Title: Virtual CGH: an Integrative Approach to Predict Genetic Abnormalities from Gene Expression Microarray Data Applied in Lymphoma

Authors:

Huimin Geng (genghm@gmail.com)
Javeed Iqbal (jiqbal@unmc.edu)
Wing C Chan (jchan@unmc.edu)
Hesham H Ali (hesham@unomaha.edu)

Version: 2 Date: 4 January 2011

Author's response to reviews: see over
January 3rd, 2011

Editor in Chief
BMC Medical Genomics

Dear Editors,

Thank you for the review of our article “Virtual CGH: an Integrative Approach to Predict Genetic Abnormalities from Gene Expression Microarray Data Applied in Lymphoma”

We would like to thank reviewers for their insightful suggestions about the findings of the study. We have addressed the issues raised by the reviewers, as detailed in the attached point-to-point response. We have also made formatting changes in the revised manuscript as requested by the editors. We hope the revisions are satisfactory and the manuscript is now acceptable for publication.

Best Regards,

Hesham A. Ali, PhD
Professor
Department of Computer Science
University of Nebraska at Omaha
6001 Dodge Street, PKI 172B
Omaha, NE 68182
hesham@unomaha.edu
Editor’s comments

We have made the following formatting changes requested by the editors in the revised manuscript.

1. The manuscript has been proof-edited by a professional science editor at University of Nebraska at Omaha. The style of written English has been improved throughout the manuscript.

2. Abstract has been restructured with four parts: Background, Methods, Results, and Conclusions, according to the guidelines for BMC Medical journals.

3. “Competing interests” section has been added between the “Conclusions” and “Authors’ contributions”.
Reviewer #1

The authors develop a statistical approach to predict CNV from gene expression. Overall the paper seems pretty clear and reasonable. What would be useful in terms of the method is a discussion of how the states in the HMM were selected and what H, L, M mean in terms of expression being quantitative as well as +,-,o in terms of cut-offs of CNV. The discussion of individual chromosomes in terms of CNV while GEP is not in terms of chromosomes can be better discussed.

The gene expression values (continuous variable) were discretized into three distinct states: “H”, “L” or “M”, representing high, low or medium gene expression, respectively. We used 1.5-fold change as the threshold to determine high or low expression of a gene in a tumor as compared to the median expression of the gene across the tumor cohort. For CGH data, “+”, “-” or “o” represent for gain, loss or normal copy number status of a gene. The signal ratios greater than 1.25 or less than 0.75 of tumor to normal cells were considered as chromosomal gains or losses, respectively. The nine hidden states were determined by the GEP observations superimposed on the CNV status: \{H+, L+, M+, H-, L-, M-, Ho, Lo, Mo\}, where H, L and M are the GEP observations and +, - and o are the gain, loss or normal CNV status. A hidden state \(H_o\) can only emit \(H\); however an emission \(H\) could come from any of the three underlying hidden state, \(H+, H\) or \(H_o\).

We have now included the above statement in Method section under subheading “vCGH Model Structure” (page 6-7), and described in more detail under subheading “Sample Description and Data Processing” (page 11).

For the second part of the query, we added the following sentences under subheading “vCGH Training and Prediction” (page 7-8).

“For a specific cancer type, genomic aberrations often occur in a specific set of chromosomal hotspots. For example, DLBCL has frequent aberrations involving gains of 2p, 6p and 18q and loss of 6q and 17q [16], and the hallmark aberrations of MCL are gains of 3q and 8q and losses of 1p, 6q, 8p, 9p, 11q and 13q [17]. To accurately reflect the chromosomal difference, we train a separate HMM for each chromosome so that each chromosome can have a different statistical transition and emission distributions.”

The main thing this paper could really use is a clear motivation for why one would want to predict CNV from GEP. The authors state that there is a lot of GEP data and I agree but why use it for CNV imputation. This can be better motivated.

Other than the wealth of GEP data already available to predict copy number alterations (CNAs), the major advantage of utilizing GEP to predict CNAs is to identify the functionally critical genetic loci important in the disease pathogenesis. Many of the common CNAs are pathogenetically significant and provide additional information on a tumor which may not be immediately evident from CGH data. GEP additionally provide functionally relevant CNA regions that might be more important than the CNA regions identified by CGH, which in principle defines only the chromosomal structural changes,
but the functional effects of CNAs can be reflected only by the changes of gene expression. The information is important in cancer research aiming at identifying the target genes in regions of CNAs and the biological effect of the CNAs. For example, if loss of heterozygosity in loci is accompanied by the loss of the gene expression, those genes are likely to play critical role in the pathogenesis of disease, rather than genes whose expression remained unchanged. This principle has been used to define tumor suppressor genes in regions of losses or oncogenes in region of gains. The gene expression status in abnormal loci can give an indirect indication about epigenetic status (DNA methylation or histone modification). Additionally, with increasing evidence of polymorphic genomic variation in genome it is more important to critically look at structural changes and its influence on gene expression status.

The above statements have been briefly added in the Abstract (page 2), Background (page 6, lines 3-10) and Results and Discussion (page 17 last paragraph, page 18 first paragraph).
- Major Compulsory Revisions
1. The source code and the program need to be made available to the research community.

We now made the source code and supplementary data publically available at: [http://vcgh.sourceforge.net](http://vcgh.sourceforge.net). This statement has also been included at the end of the Method section (page 13, last sentence before the Results and Discussion section).

- Minor Essential Revisions
2. Current analysis showed very high concordance between CGH and vCGH at the cytoband level. The authors mentioned about gene level analysis. As GEP is regulated by many other mechanisms besides genomic gain or loss, it is expected to see much less concordance at the gene level. It will be interesting to see this data.

We agree with the reviewer’s comments that we would expect less concordance between the CGH and GEP at the gene level as compared to the cytoband level in vCGH. As we showed in Table 1, the rGEP method, which is a gene-level direct mapping between GEP and CGH, did show less concordance with CGH in terms of sensitivity, specificity and accuracy as compared to the sGEP and vCGH methods, which are cytoband-/segment-level mapping between GEP and CGH. We have discussed this data in page 14, paragraph 2.

Appreciated from the model design, with intra- and inter- Markov sub-chain transitions, vCGH can identify alternative gain, loss or normal DNA segments automatically. vCGH is basically a segment-level prediction tool, but can also be evaluated at gene level when considering genes as the unit in a segment (all genes in a segment have the same CNA status). By vCGH, we would not expect gene-level and cytoband-level prediction differ much in terms of concordance with the CGH. This data was newly added in the Supplemental Table 2 (in Additional file 2), where we compared the vCGH performance on gene level and on cytoband level using the criteria of sensitivity, specificity and accuracy on the DLBCL and MCL datasets, and as expected, we observed similar prediction sensitivity, specificity and accuracy by the gene-level vCGH and the cytoband level vCGH. The above results and discussion were added in the Results and Discussion section in page 16-17. The model design of vCGH was described under subheading “vCGH Model Structure” (page 7, paragraph 2).

3. Genomic gain and loss derived from vCGH are presumably functional subsets. This is, however, not readily perceivable from current results.

We agree with the reviewer’s comments that the genomic gain and loss derived from vCGH are functional subsets of abnormalities that may be relevant to the pathogenesis of disease. Since these changes are reflected in their gene expression status, we have added such genetic abnormalities in Figure 7 under subheading “Additional Recurrent Gains...”
and Losses Predicted by vCGH”, (page 17). We investigated the functions of these genetic abnormalities with the patient outcomes (Figure 8), and identified that some of the genetic abnormalities are associated with poorer outcome. Additional genetic abnormalities may act as “passenger abnormalities” and may not have any influence on the outcome of the disease (page 17).

In addition, we have also discussed that the expression data can help identify target tumor suppressor genes in loss or oncogenes in gain CNA regions and cited the references [2, 15-25] under the Background section (page 4 line 1-3).

- Discretionary Revisions

4. It seems that the performance of rGEP and sGEP is much inferior to that of vCGH. However, there are not too many samples with paired GEP and CGH data out there that will allow vCGH to work, which was the original impetus for the development of this algorithm. Therefore, sGEP and rGEP will most likely be used in reality. However, they do not seem to perform well enough. As an alternative approach, it may be interesting to examine whether the parameters in vCGH can be trained in one small dataset (which has paired GEP and CGH data) and be used to predict vCGH in another GEP dataset of the same sample type. Cross-data validation will enhance the application of this method.

According to the reviewer’s suggestions, we performed CNAs prediction using vCGH on an independent cohort of 176 DLBCLs (Monti et al. 2005). The following is added to the Results and Discussion section (page 18-19) and a figure is also generated in Supplemental Figure 1 in Additional file 2.

“vCGH Prediction on an independent dataset of 176 DLBCLs

We applied vCGH which is trained using the paired GEP and CGH data on our 190 DLBCLs to an independent dataset of 176 DLBCLs using the GEP data [40]. The GEP data of the 176 DLBCLs were generated by Shipp Lab and downloaded at http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=102 [40]. The CGH data of the independent cohort of 176 DLBCLs is not available and therefore we compared the predicted CNAs by vCGH on the 176 DLBCLs with the CNAs identified by the CGH on the 190 DLBCLs since a specific tumor type would feature certain specific genetic abnormalities even in different cohorts. Supplemental Figure 1 (in Additional file 2) showed the prediction results on the 176 DLBCLs and the results were compared with the CGH data on our 190 DLBCLs. In Supplemental Figure 1, the top panel showed the chromosomal gains and bottom panel showed the chromosomal losses, where the 190 DLBCL data were shown above and the 176 DLBCL data were shown below the X-axis. Since the two cohorts are completely independent, we observed some differences between the two cohorts, especially in the losses. However we do observe overall similarity between the two cohorts, such as gains of 1q, 2p14-p16, chr3, chr5, 6p, chr7 and chr9, and losses of chr4, 6q, 13q and 17p. Those recurrent regions have also been reported in another independent aCGH study on 99 DLBCLs [41].”
Reviewer #3
Reviewer’s report:
This submission describes an algorithm to determine copy number from gene expression data. Overall, the approach is reasonable and may be useful in some limited circumstances. As the authors state, there is an overwhelming availability of gene expression data, but the availability of array CGH or copy number datasets is limited. The HMM approach is interesting and the paper is well-written. Critiques are listed below based on review criteria.

Major compulsory revisions
1. The background is limited, and the authors are missing key publications in the field (for example, Hertzberg et al, Genes Chromosomes Cancer 2007 an;46(1):75-86). There should be a discussion of why their approach is stronger, or how it differs from previously published approaches to address this problem.

According to the reviewer’s suggestion, we did a reference update and added the appropriate references (ref 31-34 in the revised manuscript) into the Background section. We have also discussed in detail the differences in our approach and previously published ones in pages 5-6. Briefly,

“HMMs have recently been applied in aCGH for segmentation, a procedure to divide the signal ratios of each clone on the array into states, where all of the clones in a state have the same underlying copy number [31, 32]. In this paper, HMM was first time used for an integrative analysis of GEP-to-CGH prediction which intended to capture two primary sources of uncertainty embedded in genomic data: (1) the significant but subtle correlations between GEP and CGH; (2) the sequential transitions of DNA CNAs along a chromosome. Hertzberg et al. has developed a method for predicting chromosomal aneuploidy from GEP data using fold change and chromosomal relative expression calculation for each chromosome [33]. The major limitation with this approach is that it can only call whole chromosome gain or loss. Nilsson et al. proposed a method that employed total variance minimization techniques for chromosomal segmentation based on altered gene expression pattern [34]. Our proposed vCGH method differs from the previous methods in two important respects. First, the proposed vCGH is based on HMMs, which are classical pattern recognition methods with a rich set of existing estimation and inference algorithms for sequential observations. Second, the vCGH is specifically designed to train paired CGH and GEP datasets and predict CNAs using GEP data only. The special requirement of vCGH is to ensure specificity of copy number calling from GEP data.”

We also added ref 24-25, along with the other citations in the first submission, to show the correlation between GEP and CGH (page 4 in the Background section).

2. Following (1), these data in the submitted manuscript have been presented by the authors in at least three published conference proceedings in 2007, 2008 and 2009 with the same figures and conclusions. How does this report extend the original findings? These reports are not referenced.
We have been working on the vCGH methodology since last four years and have answered several important questions during the development of the method, which resulted in several conference proceedings. For example, in 2007 conference (Geng et al. HICSS 2007), we described a multinomial probabilistic model for chromosomal segmentation from GEP data. This method was referred as “smoothing algorithm” or sGEP method in this submitted manuscript. In 2008 conference (Geng et al. ISBRA 2008), we proposed using the Viterbi decoding in the hidden Markov model and applied the method on DLBCL dataset, and in 2009 conference (Geng et al. HICSS 2009), we further proposed the posterior decoding for vCGH, and illustrated its performance using the MCL dataset as an example. We have appropriately cited and discussed these conference proceedings as ref 35-37 in pages 9 lines 1-3 and page 10 line 3 after subheading “Smoothing Algorithm”.

In the current study, we refined and evaluated the method and integrated all components for previous developments in two large datasets of DLBCL and MCL. Both these data sets were accompanied by aCGH data for validation as well and we have made the datasets along with the source code publically available at http://vcgh.sourceforge.net. We have refined and extended the previous versions by comparing the Viterbi and Posterior decoding algorithms and generated complete results on the DLBCL and MCL datasets using both decoding algorithms. Second, we compared the sGEP method with vCGH method using both DLBCL and MCL datasets. Third, importantly we investigate the functional relevance of genetic abnormalities and developed a permutation test on the association of the genetic abnormalities with gene expression (Fig 7). Fourth, we performed a survival analysis of the association of the novel genetic abnormalities with the survival time of our DLBCL cohort to further elicit the functions of genetic abnormalities (Fig 8).

3. The HMM model is appropriate and interesting, but the authors should specifically address why they limited the number of states. For example, high amplification, one copy gain, no change, heterozygous loss, and homozygous loss.

The main reason that we limit the number of states to three for GEP (L, M, H) and three for CGH (-, o, +) is the model complexity. Suppose we have five levels for CGH (-, -, o, +, ++) and GEP (LL, L, M, H, HH) respectively, then the HMM model will have 5x5 = 25 hidden states (i.e., the five GEP observations superimposed on the five CNA levels) and the transition matrix would have 25x25 = 625 parameters which is much more than the current 9x9 = 81 parameter model. Since we generally have a limited number of training samples, the five-level model will not be feasible in the current framework. But we totally agreed with the reviewer that a 5-level model would be more desirable in practice.

We have discussed the above under subheading “vCGH Model Structure” (page 7, paragraph 1).
4. The authors should provide the actual number of predicted regions, rather than simply showing the percentage. How many regions did they predict from vCGH. For example, for 90% accuracy, did the model predict 90 of 100 genes correctly, or 9 out of 10? From Figure 4, for chr 1 the output from vCGH has a sensitivity of 1.0. How many genes were predicted?

According to the reviewer’s suggestion, we have provided a table (Supplemental Table 1 in Additional File 2) indicating the actual numbers, in addition to the percentage, for the sensitivity, specificity and accuracy of vCGH for each chromosome.

The reviewer mentioned that in Figure 4, vCGH has a sensitivity of 1.0 on chr 1. Actually in Fig 4, chr1 has a sensitivity of 0 and a specificity of 1.0. As discussed in the manuscript (page 14), the low sensitivity is due to the insufficient number of cases harboring a gain or loss on certain chromosomes. We added the following sentence in on page 15.

“We noticed that vCGH did not predict well on some chromosomes, such as gain on chromosome 4 and loss on chromosome 11 for DLBCL (Figure 3) and gain on chromosomes 1, 6, 9, 10 and 13 and loss on chromosomes 4, 5, 15 and 18 for MCL (Figure 4). This is due to infrequent aberrations and hence insufficient training data for the gains or losses on those chromosomes. For example, in 190 DLBCLs, cases with chr4 gain (n=7) and with chr11 loss (n=1); in 64 MCLs, cases with gains on chr1 (n=1), chr6 (n=3), chr9 (n=1), chr10 (n=2) and chr13 (n=1), and with losses on chr4 (n=2), chr5 (n=1), chr15 (n=1) and chr18 (n=2).”

5. There should be more detail with respect to how aCGH was mapped with gene expression data. What were the criteria for one to one mapping? Were multiple probes within the same gene averaged? How were promoter, coding region and 3'UTR regions treated?

According to the reviewer’s suggestion, we added the following description in page 12, under subheading “Sample Description and Data Processing”.

“We used the probeset level data from GEP without averaging multiple probesets within the same gene. A probeset from GEP data would be marked with “+” or “-” if its chromosomal locations were covered by the start and the end of a gain or a loss region from the CGH data. The chromosomal locations of genes and cytobands were obtained by Affymetrix probesets alignments and NCBI Human Genome database Build 36.1, respectively. The vCGH model is based on hidden Markov Models that consider expression probesets as a sequence of hidden states without considering the distance between probesets. The vast majority of the expression probesets were near the 3’ end of coding region and probesets located at other regions were equally treated. With the vCGH model, the genomic gain and loss regions of variable length (spanning a number of probesets) can be called automatically.”

6. Figure 8: HMM based vCGH predicted new regions not detected in array CGH
– is it possible the Vysis CGH kits did not correctly identify the regions of aberration? The company recommends 1.25 and 0.75 cutoffs to define amplification and deletion, which may not be an optimal approach.

We agree with the reviewer comments that Vysis CGH kits have technical limitations especially in small aberration detection, and may not identify all the CNA regions in the tumor cells. CGH is a technically challenging hybridization-based technique and gives a gross impression of the genomic map of gains and deletions in tumor. Other aCGH platforms (such as NimbleGen 385K tilting arrays or Affymetrix SNP6.0 arrays) would further validate those newly identified CNA region by vCGH.

We agree with the reviewer that the optimal cut off may vary for calling a “gain” or “deletion”. The company recommends a uniform approach for a large series of cases; however it might not always be true given the background noise in hybridization and the contribution by the normal cells in stromal or other reactive elements in the tumor microenvironment for a particular sample. For some of the critical regions, FISH is recommended to accurately quantify the copy number changes in specific loci.

We have added the above discussion in page 18, paragraph 1.

Discretionary Revisions
1. This report would be stronger if the algorithm were applied to a third dataset. Ideally, this would proceed a biologic validation of candidate regions. However, another approach would be to apply this algorithm to a third dataset at another cancer site where the array CGH data are available and test the sensitivity and specificity. This would be a nice addition, as mentioned above the results in this manuscript have already been presented multiple times.

Following the reviewers’ great suggestions, we performed CNAs prediction using vCGH on an independent cohort of 176 DLBCLs (Monti et al. 2005). The following is added to the Results and Discussion section (page 18-19) and a figure is also generated in Supplemental Figure 1 in Additional file 2.

“vCGH Prediction on an independent dataset of 176 DLBCLs
We applied vCGH which is trained using the paired GEP and CGH data on our 190 DLBCLs to an independent dataset of 176 DLBCLs using the GEP data [40]. The GEP data of the 176 DLBCLs were generated by Shipp Lab and downloaded at http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=102 [40]. The CGH data of the independent cohort of 176 DLBCLs is not available and therefore we compared the predicted CNAs by vCGH on the 176 DLBCLs with the CNAs identified by the CGH on the 190 DLBCLs since a specific tumor type would feature certain specific genetic abnormalities even in different cohorts. Supplemental Figure 1 (in Additional file 2) showed the prediction results on the 176 DLBCLs and the results were compared with the CGH data on our 190 DLBCLs. In Supplemental Figure 1, the top panel showed the chromosomal gains and bottom panel showed the chromosomal losses, where the 190
DLBCL data were shown above and the 176 DLBCL data were shown below the X-axis. Since the two cohorts are completely independent, we observed some differences between the two cohorts, especially in the losses. However we do observe overall similarity between the two cohorts, such as gains of 1q, 2p14-p16, chr3, chr5, 6p, chr7 and chr9, and losses of chr4, 6q, 13q and 17p. Those recurrent regions have also been reported in another independent aCGH study on 99 DLBCLs [41].”