OXFORD JOURNALS
Bioinformatics

Submit a Revision

Review the information below for correctness and make changes as needed. **After reviewing the manuscript proofs at the foot of this page, you MUST CLICK 'SUBMIT' to complete your submission.**

1 View and Respond to Decision Letter
2 Type, Title, & Abstract
3 Attributes
4 Authors & Institutions
5 Details & Comments
6 File Upload
7 Review & Submit

Save and Go Back Submit

My Manuscript Information

Step 1:View and Respond to Decision Letter

Edit

Response to Decision Letter:

Dear all,

we wish to thank the editors and especially reviewers for your valuable feedback. We have updated the manuscript accordingly. All modifications are highlighted in red in the revised manuscript/PDF. In addition, there are now three separate Supplementary Notes.

Next, we will address each of the comments below. We have labelled the comments from Reviewer #1 by 'R1:', the ones from Reviewer #2 by 'R2:' and so on, and all our answers by 'A:'. At the very end we have added a small list of references.

- -
Reviewer: 1
- -
Comments to the Author

This article describes a method for pre-processing Affymetrix genotyping platforms and deriving locus-specific estimates of copy number (CN), extending previous work on this subject (Bengtsson et al., 2008).

General comments:

R1: Locus-specific CN estimates are important for segmentation and hidden Markov model algorithms that identify alterations in copy number spanning multiple-loci (CN variants). The main innovation in this article is an algorithmic change to aroma.affymetrix that allows arrays to be processed one sample at a time instead of in batches. Overall, the preprocessing methodology is similar to Bengtsson et al. 2008, with additional adjustments for co-hybridization of fragments digested by multiple enzymes in the latest Affymetrix 6.0 platform. The authors demonstrate that their approach is as effective or superior to the Affymetrix CN5 tool and dChip for detecting hemizygous and homozygous deletions.

A: Clarification: We wish to emphasize that the method is called CRMA v2 and there exist an implementation in the aroma.affymetrix framework. Neither the method nor the algorithm should not be referred to as 'aroma.affymetrix'.

Specific comments (Major):

R1: Identifying CN deletions are only half the problem. Amplifications (CN >= 3 for autosomes) are equally important and likely to be more challenging as the signal to noise ratio diminishes for larger CN estimates.

A: We agree that CN gains are equally important as CN losses (and CN neutral regions). We also agree that it is easier to separate CN=1 from CN=2 than, say, CN=3 form CN=2, for the same reason as it is easier to separate CN=0 from CN=1 compared with CN=1 from CN=2. The latter we comment on in the beginning of Section 3.2.

A: There are a few reasons why we did not *compare* methods for other changes than CN=2 v. CN=1 and CN=1 v. CN=0, namely:

(i) Given any data set with males and females, CN=0, CN=1 and CN=2 are the only CN states we know exists in *any* data set.

(ii) Assuming we are *within the dynamic range* of the technology, it is what *we* choose to be the reference that defines the (relative) copy neutral state. For instance, we choose to use CN=2 to be the neutral state for the CN=2 to CN=1 evaluation, and we choose CN=1 to be the neutral state for the CN=1 to CN=0 evaluation. We could equally well choose CN=1 to be the neutral state in the first evaluation, and then the change from CN=1 to CN=2 would be interpreted as a "gain". This would correspond to calling "females" among "males" instead of, as now, "males" among "females". The ROC results for the latter two are nearly symmetric to each other.

(iii) We did not explicitly write it, but the ROC performance of a method strongly depends on the signal-to-noise ratio (SNR) of the raw data, that is, the change in mean levels relative to the noise of the two levels. Skewness of the distribution of raw CNs can also play a role. We "assume" (read have observed but not formally shown) that the SNR/skewness changes gradually and smoothly with the change in CN change. This is something we intend to present in an upcoming manuscript on generic methods for evaluating CN data.

(iv) Because of the above paragraph, and *interpolating* that these properties hold also as long as we are well within the dynamic range of the technology, we expect for *any given method* that the ROC performance for, say, a change from CN=1.8 to CN=0.9 (in case of sample heterogeneity) is somewhere in between the method's ROC performance for CN=2 to CN=1, and CN=1 to CN=0. Similarly, we *extrapolate* the performance of a change from for instance CN=1.8 to CN=0.4, or from CN=2 to CN=0 (homozygote deletion).

(v) In previous paragraph(s) we argue that one can interpolate/extrapolate the ROC performance for a particular method as long as we are within the dynamic range, and that this performance change is smooth. From this we argue that we expect the ROC performance of *methods relative to each other* to behave smoothly. In other words, we do not expect abrupt differences in relative performances as we go to other CN levels.

(vi) Using similar arguments, we expect to see similar relative performances for, say, CN=4 to CN=2 (or almost equivalently CN=2 to CN=4) ROC performances.

(vii) Finally, one of the initial goals of this new method was to extend the existing CRMA (v1) to work also for the newer chip types (GWS5 and GWS6). There were quite a few intrinsic modeling as well as technical problems to solve in order to achieve this, which some are mentioned in the paper. Considering this work to be such an *extension*, we have then *added* an additional test compared with the accepted and published CRMA (v1) paper (ChrY data did/does not exist for those older chip types so it was not possible to evaluate CN=1 to CN=0 changes). Thus, we believe that the evaluation presented here is an improved version of an already scientifically accepted evaluation method. This does of course not answering the question any reader may have on the actual ROC performance of gains.

A: However, we have introduced an additional approach to evaluate also gains. This was possible after another public GWS6 data set was recently published. Contrary to the existing assessment, this alternative approach assess the performance using a single change point in a single sample. The idea is to identify a region containing a single change point. After locating the change points with some accuracy we exclude loci in a safety region surrounding the change point. The remaining data points are used as the "truth" where one side belongs to one CN state and the ones on the other side another CN state. Although we don't know the true CN levels, we can assume they are different and therefor use the region for ROC analysis similar to what we did in the approach based on ChrX and ChrY across multiple samples. We used a published tumor-normal data set for this. This confirms that our previous results on the performance of CRMA v2 relative to the two other methods is also valid for this evaluation method and for both gains and loss. We have added Supplementary Notes #3 with all details and results. We intend to publish a more thorough description of different CN evaluation methods elsewhere. We have added these results also the manuscript in the Results section (and updated the abstract accordingly).

R1: Assuming batch/lab effects are very important for copy number methods, is the single sample approach to preprocessing likely to be useful?

A: It is well known that lab effects exist and are very hard to control for. We have not performed any study on how well the proposed method and other methods can control for lab/batch effects (we made a note on this in the Discussion section). Furthermore, it has often been suggest that one can use HapMap data as a gold-standard reference in other studies. We believe this is suboptimal and we do discourage people from doing this. It is much better to use an "in-house" data set (normal or not) [Bengtsson H et al. 2007; RECOMB 2007 poster]. Others reports they observe the same (see various aroma.affymetrix threads online). To the best of our knowledge, there are no methods that "magically" removes lab effects. This includes the well known/used quantile normalization method [Bengtsson H et al. 2007; RECOMB 2007 poster].

A: We wish to note that the term "single-array" is mainly referring to the fact that the model does not *need to* estimate parameters across arrays as "multi-array" methods does. Because a method estimate parameters based on multiple arrays, does not mean it is controlling for batch/lab effect. For instance, there is nothing particular with the quantile normalization method that makes it control for batch/lab effects. In other words, just because our method is applied to each array independently, it does not mean that is better or worse at controlling for batch/lab effects compared with existing multi-array methods. One related comment, our method does control for differences across arrays in the sense that it rescales them to have the same (predefined) median signal (the 50% quantile vs all quantiles).

A: Furthermore, there is nothing in quantile or the fragment-length/GC-content normalization methods that requires them to be multi-array methods. It is only how we choose the target distribution/effect that makes it a single- or a multi-array method. For fragment-length/GC-content normalization, we argue in the paper, by comparing with the models in Bengtsson et al. 2008, that the target distribution makes no difference. Hence, it can be made/is a single-array method. For similar reasons, if you would use quantile normalization, you can choose to use a predefined target distribution, instead of estimating it from the existing data set. We wish the point out that the dChip software has for a long time use this strategy; by using one of the arrays as the "baseline array" the quantile normalization method can be applied to each array independently and is hence also bounded in memory. Similarly, in the Affymetrix GTC software one have the option to use aprecalculated distribution based on the HapMap data as the target reference. As long as one choose a reasonable smooth distribution, this should have very little impact on the final results.

R1: Are estimates of the standard errors for the locus-level CN available? Standard errors will help downstream approaches that identify variants spanning multiple loci.

A: To estimate standard errors one either has to borrow from other loci on the same array or across arrays for the same locus. For a single-array preprocessing method, the latter has to be done after processing each array, more a multi-array method it can naturally be done while fitting the probe-level model for each locus. However, if the multi-array method is followed by downstream normalization methods, it is likely better calculate standard errors on the summarized data (as done with the single-array approach). Forpractical purposes, there are methods in the aroma.affymetrix framework for calculating the (robust or non-robust) mean and the variance across arrays.

Specific comments (Minor):

R1: Are the multi- and single-sample (v1 versus v2) [CRMA] [was: aroma.affymetrix] also comparable in terms of FPR/TPR?

A: This "minor comment" is actually three different comments/questions in one:

(i) The main purpose for developing the CRMA v2 was to provide a method for the GWS5/GWS6 platforms, because CRMA v1 is not applicable to those. So, from the perspective of GWS5/GWS6, it does not makes sense/it is not possible to compare CRMA v1 with CRMA v2.

(ii) In order to compare CRMA v1 and CRMA v2, we have to utilize 10K, 100K, or 500K SNP data sets. Hence, any results from such a comparison would only gain these

types of data sets. Since the GWS5/GWS6 chip types were release in early 2007, we think that providing a formal comparison based on these older chip types is of less interest.

(iii) The question may also be interpreted as to what extent the choice of not modelling probe affinities affects the performance. For the 10-500K chip types one can indeed argue that modelling probe affinities makes a difference and that therefore a log-additive model (or similar should be used), whereas for the GWS5/GWS6 chip types this is no longer necessary because all SNP probes are technical replicates (with identical affinities) and therefore the mean/median is reasonable estimator. In our studies, if we model the probe affinities for GWS5/GWS6 data, we observe no difference in the performance of the raw CNs (not shown). However, the reviewers comment did make us aware that we should make it clear to the reader that it still may be worthwhile modeling probe affinities for earlier chip types. The reason why this was left out is that, as notes above, the main purpose/focus of the CRMA v2 methods are the newer chip types. We have clarified this by adding a paragraph to Section 2.2.3. As stated in the update, we have not conducted a formal study on this. One reason is that the outcome will depend on the number of arrays included in the multi-array approach.

R1: For the evaluation section, restate the method used to call CN=1 and CN=2 from the raw copy numbers

A: We have added a paragraph to Section 2.5.1 'Differentiating CN=1 and CN=2 (ChrX)' restating the evaluation method in Bengtsson et al. (2008).

R1: The affine transformation uses SNPs that are likely to be diploid. How are these SNPs selected...were regions excluded because of CNV identified by array-based methods? How many SNPs were used?

A: The main purpose of writing "SNPs that are likely to be diploid" is that it is slightly more generic than writing "SNPs that are on autosomal chromosomes but not on sex chromosomes", which is the main purpose. This is why we write "i.e. ChrX and ChrY data are excluded" at the end of the sentence "[...] this model are estimated based on the subset $\mathcal{J^*}$ of loci that are likely to be copy neutral regardless of sample, i.e. ChrX and ChrY data are excluded." follow Eqn (2).

A: If one includes data from sex chromosomes it is likely that one "shrink" females and males towards each other. This is true for any normalization method, not just the methods we are suggesting. However, with our model, method and implementation it is possible to exclude loci beyond the sex chromosomes that may add to the risk of doing the same, e.g. loci in known CN polymorphic regions. This should be compared to for instance existing *rank-based* quantile normalization methods for which this is impossible (even for sexchromosome loci). Note that all our calibration and normalization methods use robust estimators. In other words, having a small fraction of CN aberrations will not affect the model fit. In the aroma.affymetrix implementation of CRMA v2 (and CRMA v1), the default is to exclude ChrX and ChrY data points.

R1: Is quantile normalization of the intensities typically not needed after the suggested preprocessing?

A: We do not use quantile normalization for any of our CN analysis. One reason is that the empirical distributions of probe intensities are very similar after doing crosstalk calibration. Please see Figure 1 or Bengtsson et al. (2008) illustrating exactly this. There we show that much of the difference in empirical distributions observed across samples are controlled for by the affine correction done by the allele-crosstalk calibration. Equivalently, we argue that discrepancies in signal distributions can often be explained by a simple affine (offset and scale) transform. This can also be seen if one plot the probe intensities before and after quantile normalization against each other; the transform is close to linear offset and scale (expect near the extreme tails where rank-basedQN methods show problems).

R1: Axis labels in the figures need to be larger

A: We are sorry, but just before the resubmission deadline we realized that we forgot to do this. For this reason we have committed the revision without these updates, but if the editor allows we will commit updated figures with larger labels as soon as possible.

R1: Figure 3 caption: slightly

A: Corrected.

R1: p.2 l.34 These constraints

A: Corrected.

- -
Reviewer: 2
- -
Comments to the Author

R2: The main contribution of this article is to propose a single-array preprocessing method for Affymetrix SNP arrays that appears to work as well as other multi-array preprocessing methods. The authors compared their method, CRMA v2, with affymetrix CN5 and dChip on estimating copy numbers fromchrX and chrY.

R2: 1. The competing methods CN5 and dChip should be spelled out.

A: All details on these methods are in the Supplementary Notes #2 (as named in the revised versions). It is our aim that the manuscript together with the Suppl Materials is self-contained, and that any one should be able to reproduce the results exactly given these documents. However, the page limitation does not allow us to do this. FYI, in the review process of our previous CRMA (v1) method (Bengtsson et al. 2008), the reviewer asked us for theopposite - to move all such details in the Suppl. Materials- which we then did.

R2: 1b. In addition, allelic crosstalk need to be explained briefly rather than referring the readers to Bengtsson et al (2008b)

A: We have added an additional paragraph to Section 2.2.1 that in words clarifies the algorithm. We now also reference the technical report that describes the algorithm (same reference as in Bengtsson et al. 2008).

R2: 2. How sensitive are cross talk parameter estimates to the scale of copy number abberations (cancers, for instance)? My concern is that relying on the robustness of the estimator might not be enough to counteract the adverse influence of large scale CN aberration and will lead to attenuation of signals when comparing to normals. Perhaps some pre hoc filtering is in order here before subjecting to the cross talk model.

A: The offset (apex) and the crosstalk parameters (direction of homozygote arms) will not be affected much by CN aberrations, because such aberrations occur within polyhedral cone defined by the homozygote arms and the estimator is robust against changed within. Moreover, the direction of the homozygote arms (A, AA, AAA, ..., B, BB, BBB, ...) will be the same regardless of CN level. The crosstalk estimator is robust against outliers relative to these arms.

We believe the reviewer's concern is instead related to the overall scale. We constrain the crosstalk calibration model such that the median of all probe signals will equal an arbitrary constant (=2200; same for all arrays). If it is true that the median *true* signals is not the same between arrays, then the relative CNs will indeed be biased. This scaling issue is not specific to the crosstalk model perse , but related to any calibration/normalization method. Again, we would like to point out that this problem is even more sever for the quantile normalization method that is more or less thede

facto standard method.

Filtering will be similar to using a different percentiles (than 50%) as the scaling factor. Which level to use is not clear. It can of course be provided as an options in the implementation. Because CRMA v2 is sequential and single array, it likely that this filtering/scale correction does not have to be done until the very last step when calculation the relative CN ratios. Since this issue has been raised by several people in the microarray community over many years without any definite answer, we believe it is beyond this manuscript to discuss it here.

R2: 3. Affymetrix probe intensities are well known to correlate with GC contents, the probe-sequence effect model adopted from Carvalho et al, corrects for position specific sequence effect, but does not adjust for number of specific nucleotide per sequence, especially number of Gs and Cs, if I am reading it correctly. I think adding these few parameters can potentially improve the variability even further.

A: The nucleotide-position model does to some extent also correct for the number of nucleotides of a certain type, because for GC-rich probes more nucleotide-position terms for G & C will be included. However, it is correct that there are no explicit count terms which have been suggested by other (e.g. Johnson et al, 2006), e.g. n_C, n_G & n_T.

A: While preparing CRMA (v1) we did investigate how much the GC content (GC fraction) affected the CNs on the 500K arrays. We did implement a GC-content normalization method very similar to the PCR fragment-length normalization method, which was also applied on the summarized signals. Based on similar ROC analysis we concluded that effect was minor or even slightly negative. This is mentioned in our previous paper (Bengtsson et al. 2008). For CRMA v2 study presented here, we observe the same - the GC-content effect remaining after applying CRMA v2 is minor orneglectable . We have added a new Section 2.2.5 'Normalization for GC-content effects' that clarifies this. It should be noted that the GC-content annotation data for this was obtained from Affymetrix which calculated the GC-content in a much larger window around the probe, i.e. that is not just the GC-content of the probe.

A: FYI, although independent of this manuscript and method, in aroma.affymetrix there are other options for sequence models, such as a nucleotide-count model and nucleotide-pair-position model (as well as GC-content normalization on the probe summaries). These work the same for 10K-GWS6.

R2: 4. I think it will be helpful for the readers if the authors can produce boxplots of variance along the preprocessing procedures to showcase the effect on variation reduction and and the amount of reduction in each step. A scatter plot of (log-intensities, or log-ratios against probe position) will also be fine.

A: Boxplots of *variances* alone will not reveal the improvement of a method and may even be misleading, because they do not take into account the amplitude of the CN changes. Different steps as well as different methods produce raw CN ratios on different scales, that is, they are more or less compressed toward the copy neutral state (CN=2). For instance, the offset correction in the crosstalk calibration step will *increase* the variances, but at the same time it will increase the separation of CN mean levels. A possible proxy would be to present how the SNR between CN change and variance (of any two states) differ between steps and methods. However, we argue that the ROC analysis is better and assumes less about the noise distribution.

A: For the new tumor-normal data set, we now depict raw CN estimates along the genome.

- -
Reviewer: 3
- -
Comments to the Author

R3: This paper proposed a single-array analys copy number method for Affymetrix SNP 5.0 and 6.0 arrays. This has certain advantages such as analyzing data in sequential or parallel manner, and paired normal and tumor samples can be analyzed

without using further other samples. It's good that the software is available in R to users.

R3: 1. The authors modified previous methods in two papers to derive copy number estimates from single arrays. Some aspects are novel but many are technical extension of the existing methods, such as sequence effect model (2.3.2) or normalization of fragment-length effects (2.3.4).

A: Please not that there are some novelties in how we estimate the sequence effect model and correct for the effects. For instance, we do *not* use ChrX and ChrY loci for estimation, because then the estimates could be biased depending on gender. Moreover, we apply the estimated normalization function to the intensity scale (as a scale factor) and not to the log scale (as an additive term). This way we avoid problems with zero (and negative) signals, which may occur with copy-number data (depending on preprocessing method). Normalization on the log-scale would introduce missing values. The reason for not also estimating the model on the intensity scale is simply that there are many more existing methods for estimating additive models (we are aware that there is a risk for biased estimates when we fit the model on the log scale, but we assume this effect is small).

A: Yes, some problems are technical, but in order to solve those, a deep understanding of the technology as well as a sound underlying statical model is required. We would like to assure the reviewers that the proposed method is not the first possible one that runs on the data, but has evolved from detailed studies of the technology, gained knowledge from analyzing many data sets, and having collaborators and beta users testing the method on data sets that are unknown to use.

R3: 2. The usefulness of the copy number estimates lies beyond estimating chromosome X's 0, 1, or 2 copies. Take examples of CNVs in normal samples or copy alterations in cancer sample. The author only used Hapmap samples to compare methods via their estimates of copy number 0, 1, or 2. On explanation is the current lack of more public data sets for the new array types. But given the small number of required arrays for the proposed methods, a few cancer samples may be generated and their copy number are then estimated and compared with RT-PCR or across multiple analysis methods.

A: It is true that at the time when we submitted the manuscript there was not many public GWS6 data sets. Since then, at least one has become available on GEO (GSE13372), which we now have downloaded and processed with CRMA v2, dChip and Affymetrix GTC. It is a tumor-normal data set consisting of 68 hybridization/arrays in total. Our additional evaluation is done based on a few losses and gains in a single pair. For CRMA v2 it was enough to process only this single pair of arrays in order to obtain CN ratios. For the multi-array methods dChip and CN5, raw CN were estimated using all 68 arrays. Even though the latter methods "were allowed to" borrow strength for the other 66 arrays in the data set, our conclusion is still that CRMA v2 performs equally well or better than these two both for gains and losses. The details are in the Supplementary Notes #3 and part of the results are added to the Results section. Also, we would like to assure the reviewer that we, as well as others, have satisfactory applied CRMA v2 to several tumor data sets, cf. aroma.affymetrix group/website/vignettes.

We do not have the option to conduct the assays ourselves or to run RT-PCR - we are part of a statistical department. Moreover, it is not clear to what extend such a comparison would be useful beyond verifying that some of the regions identified are true. It would not allow us to compare to existing methods.

- -
Reviewer: 4
- -
Comments to the Author

R4: This is a review of the CRMA v2 manuscript by Bengtsson, Wripati, and Speed. In my
experience the original CRMA is the best method for estimating copy number from Affymetrix SNP arrays. A single-array extension is a real service to the community. The authors show that v2 is superior to two other multi-array methods for distinguishing between 1 and 2 copies, and that it is comparable to the better method

for separating 0 and 1 copies. I think, however, that the manuscript can be improved.

R4: 1. The most glaring issue is that the original CRMA is not included in any of the comparisons. To me and many others, that is the standard. If the results are so similar that the authors do not wish to add them to figures, I would like this mentioned.

A: Please see our detailed reply to Reviewer #2 (under 'Specific comments (Major)') who raises the same issue.

R4: 2. I would also be interested in a discussion, and maybe some evaluation, of the impact of reference samples in tumor studies. Does one need paired normals, or would historical normals from the HapMap project be sufficient? Data from the Cancer Genome Atlas might be relevant.

A: From our experience, which also others have confirmed, we strongly recommend people to use an in-house reference. We do not recommend the use of external data sets such as HapMap as a reference. This is rather suboptimal and you can do much better with even a small (anonymous) in-house data set. This is what HB (author) officially recommend in the aroma.affymetrix forum. Moreover, thesefindings imply that there are additional batch/lab effects that we still do not control for. Finally, we believe a discussion on this is beyond this paper and important enough to be a standalone publication itself and we prefer not to discuss it in the manuscript.

A:4Unfortunately we cannot use Affymetrix data from the The Cancer Genome Atlas (TCGA) project for our publication. Data from SNP platforms is not public and requires special rights to access.

R4: 3. I think there is too much dependence on the previous CRMA manuscript. As an example, there is nothing in the results section as to how many SNPs are evaluated. I realize this information is in the other manuscript, but I would prefer not to have to flip back and forth. I think something should be added about the algorithm for estimating the offset and crosstalk.

A: We have added a paragraph to Section 2.5.1 'Differentiating CN=1 and CN=2 (ChrX)' restating the evaluation method in Bengtsson et al. (2008). Reviewer #2 pointed out the same thing.

A: We have added information on the number of loci evaluated for the ChrX and ChrY evaluations as well as for the new two tumor-normal regions.

A: Regarding the crosstalk algorithm, please see our reply to Reviewer #2 who had the same concern; we have added a paragraph to the manuscript.

R4: 4. I think the information in section 2.2 about the data should be moved to after the model development and possibly merged with section 2.5.

A: We appreciate the suggestion. The section on the data used has been moved to Section 2.4 after the section on 'Implementation' and before the 'Results'.

R4: 5. The writing while clear is a bit terse, particularly in the model development. In addition, I think the manuscript needs one more edit. For example, in the first paragraph of page 3, the sentence starting with "This also suggests" and the sentence starting with "In Bengtsson et al. (2008b)" are not quite correct.

A: Thank you for the comment. We have gone through the manuscript and corrected the language at a few places.

- -
REFERENCES
- -
Bengtsson, H. & Hössjer, O. Methodological study of affine transformations of gene

expression data with proposed robust non-parametric multi-dimensional normalization method.BMC Bioinformatics, 2006, 7, 100.

Bengtsson, H.; Irizarry, R. A.; Carvalho, B. & Speed, T. P. Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics, 2008, 24, 759-767.

Bengtsson, H. Poster: Anonymous in-house and public reference sets for copy-number analysis of a small number of Affymetrix SNP arrays, RECOMB 2007, April 2007. URL: http://www.stat.berkeley.edu/share/hb/talks/BengtssonH_20070422-RECOMB2007,poster.A4.pdf

Johnson, W. E.; Li, W.; Meyer, C. A.; Gottardo, R.; Carroll, J. S.; Brown, M. & Liu, X. S. Model-based analysis of tiling-arrays for ChIP-chip. PNAS, 2006, 103, 12457-12462.

Step 2: Type, Title, & Abstract Edit

Category: Original Paper

Title: A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6

Abstract: Motivation:
High-resolution copy-number (CN) analysis have in recent years gained much attention, not only for the purpose of identifying CN aberrations associated with a certain phenotype but also for identifying CN polymorphisms. In order for such studies to be successful and cost effective, the statistical methods have to be optimized. We propose a single-array preprocessing method for estimating full-resolution total CNs. It is applicable to all Affymetrix genotyping arrays, including the recent ones that also contain non-polymorphic probes. A reference signal is only needed at the last step when calculating relative CNs.

Results:
As with our method for earlier generations of arrays, this one controls for allelic crosstalk, probe affinities and PCR fragment-length effects. Additionally, it also corrects for probe-sequence effects and co-hybridization of fragments digested by multiple enzymes that takes place on the latest chips. We compare our method with Affymetrix' CN5 method and the dChip method by assessing how well they differentiate between various CN states at the full resolution and various amounts of smoothing. Although the others use data from all arrays for their preprocessing, we observe that CRMA v2 performs as well as or better even if it is a single-array method. This shows that it is possible to do online analysis in large-scale projects where additional arrays are introduced over time.

Availability:
A bounded-memory implementation that can process any number of arrays is available in the open-source R package aroma.affymetrix.

req **Subject Category**

req Please select your manuscript subject category: Genome analysis

Step 3: Attributes Edit

Keywords: copy number* , microarray* , Software* , DNA* , SNPs*

Step 4: Authors & Institutions Edit

1. Bengtsson, Henrik; University of California, Berkeley, Department of Statistics
2. Wirapati, Pratyaksha; Swiss Institute of Bioinformatics, Bioinformatics Core Facility
3. Speed, Terence; University of California, Berkeley, Department of Statistics; Walter & Eliza Hall Institute of Medical Research, Bioinformatics Division

✓ **Step 5: Details & Comments** Edit

Cover Letter:

Dear Editor,

please find our revised manuscript and updated supplementary notes.

One of the reviewers pointed out that some of the labels in the figures are too small. Unfortunately, we just realized that we forgot to update those. Since we might only have 30 minutes left before submission deadline (if GMT counts), we have decided to commit the revised version without the "large-label" figures. However, if possible, we would like to submit these afterward (probably offline because the online submission will be locked). Is this possible? We are sorry about this.

All the best,

Henrik Bengtsson et al.

req **Is this is a resubmission of a manuscript which was previously submitted under a different manuscript number?**

Yes

✓ No

If yes, what is the manuscript ID of the previous submission?

req **Do you or your co-authors have any conflict of interest?**

As corresponding author it is your responsibility to confirm with your co-authors whether they have any conflicts to declare. If you are unable to do this you will need to co-ordinate the completion of written forms from all co-authors, and submit these to the editorial office before the manuscript is accepted.

If you are in any doubt what constitutes a conflict, please read the FAQs or contact the editorial office.

If the answer is yes, please provide details of potential conflicts of interest in the space provided stating which authors they apply to. Also, you will need to include a prominent paragraph in your submitted manuscript.

Yes

✓ No

If yes, please state:

Please enter the Word Count for your manuscript:

8 pages

Please enter Number of Tables:

0

Please enter Number of Figures:

7

Does the manuscript include Colour Figures?

No

If yes, please state which figures:

Figures 1-7 are "online-only colour" figures. The intensity of the colors were chosen such that they separate well went printed as grayscale.

req If this manuscript is accepted for publication, I agree to pay the reproduction cost of £400 per colour figure reproduced in the print version of the journal. (Please read the Instructions to Authors before selecting the agreement checkbox.) ✓

req I agree to pay the relevant page charges for this manuscript if it is accepted for publication. (Page charges are levied on manuscripts over the permitted published length; please read the Instructions to Authors before selecting the agreement checkbox.) ✓

req This submission is on behalf of all authors and signifies that they are in complete agreement with the contents of the paper, that they are prepared to abide by the policies of the journal and that this paper has not been submitted elsewhere. ✓

Software that is the main focus of your submission should be available to the reviewers. If software forms part of your submission please enter below details of the web site or FTP server through which reviewers can access the software.

aroma.affymetrix
URL: http://www.braju.com/R/aroma.affymetrix/

req **Confirm whether your paper contains supplementary data to go online only.**

Yes ✓

No

If yes, please state:

There are 3 (three) Supplementary Notes (in PDF) with this manuscript.

✓ **Step 6: File Upload** Edit

1. BengtssonH_2009b-CRMAv2.pdf
2. BengtssonH_2009b-CRMAv2,SupplNotes1,AnnotationSummary.pdf
3. BengtssonH_2009b-CRMAv2,SupplNotes2,Methods.pdf
4. BengtssonH_2009b-CRMAv2,SupplNotes3,GSE13372,HCC1143.pdf
5. CRMAv2,chrX,all,ROC,d=0_15.eps
6. CRMAv2,chrX,all,smooth1-4,ROC,d=0_10.eps
7. CRMAv2,chrX,all,tpResolution,fpRate=2_00,y=0_50-1_00.eps
8. CRMAv2,chrY,all,tpResolution,fpRate=2_00,y=0_85-1_00.eps
9. CRMAv2,chrY,all,ROC,d=0_10.eps

10. CRMAv2,chrY,all,smooth1-4,ROC,d=0_06.eps
11. CRMAv2,chrX,snp,smooth1-4,ROC,d=0_10.eps
12. CRMAv2,chrX,cn,smooth1-4,ROC,d=0_10.eps
13. CRMAv2,chrX,snp,tpHist,fpRate=3_45.eps
14. CRMAv2,chrX,cn,tpHist,fpRate=3_45.eps
15. CRMAv2,chrY,snp,smooth1-4,ROC,d=0_06.eps
16. CRMAv2,chrY,cn,smooth1-4,ROC,d=0_06.eps
17. CRMAv2,chrY,snp,tpHist,fpRate=3_45.eps
18. CRMAv2,chrY,cn,tpHist,fpRate=3_45.eps
19. GSM337641,chr01,100_1-107_5,ratios,track.eps
20. GSM337641,chr01,100_1-107_5,ratios,ROC.eps
21. GSM337641,chr10,61-69,ratios,track.eps
22. GSM337641,chr10,61-69,ratios,ROC.eps

Offline Files:

1. Figures with greater labels. Please see Cover Letter.

Step 7: Review & Submit

HTML PDF View MedLine Format

Save and Go Back Submit