Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA

Roger Pique-Regi1,2,*, Antonio Ortega1, Shahab Asgharzadeh2,*

1Signal and Image Processing Institute, Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, EEB 400, 3740 McClintock Ave, Los Angeles, CA 90089-2564 and 2Departments of Pediatrics and Pathology, Saban Research Institute, Childrens Hospital Los Angeles, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Received on October 14, 2008; revised on February 17, 2009; accepted on February 26, 2009
Advance Access publication March 10, 2009
Associate Editor: John Quackenbush

ABSTRACT

Motivation: The complexity of a large number of recently discovered copy number polymorphisms is much higher than initially thought, thus making it more difficult to detect them in the presence of significant measurement noise. In this scenario, separate normalization and segmentation is prone to lead to many false detections of changes in copy number. New approaches capable of jointly modeling the copy number and the non-copy number (noise) hybridization effects across multiple samples will potentially lead to more accurate results.

Results: The new proposed approach is compared with the currently used strategy of separate normalization followed by independent segmentation of each array. Real microarray data obtained from the HapMap samples are randomly partitioned to create different multiple sample model. The copy number component is independent for each sample and uses a sparse Bayesian prior, while the reference hybridization effects are less variable if the reference set of measurements are unbiased (i.e. the noise distribution is centered at zero) and thus making it more difficult to detect them in the presence of significant measurement noise. In this scenario, separate normalization and segmentation is prone to lead to many false detections of changes in copy number. New approaches capable of jointly modeling the copy number and the non-copy number (noise) hybridization effects across multiple samples will potentially lead to more accurate results.

Results: The new proposed approach is compared with the currently used strategy of separate normalization followed by independent segmentation of each array. Real microarray data obtained from the HapMap samples are randomly partitioned to create different multiple sample model. The copy number component is independent for each sample and uses a sparse Bayesian prior, while the reference hybridization effects are less variable if the reference set of measurements are unbiased (i.e. the noise distribution is centered at zero) and thus making it more difficult to detect them in the presence of significant measurement noise. In this scenario, separate normalization and segmentation is prone to lead to many false detections of changes in copy number. New approaches capable of jointly modeling the copy number and the non-copy number (noise) hybridization effects across multiple samples will potentially lead to more accurate results.

Availability: http://biron.usc.edu/~piquereg/GADA

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Copy number variations (CNVs) represent deviations from the normal number of DNA copies generally found in the genome of organisms (e.g. two for diploid cells). In recent years, advances in microarray technology have revealed the presence of short copy number variants that are frequently repeated across normal genomes (i.e. copy number polymorphism) (Feuk et al., 2006; Friedman et al., 2004; Freeman et al., 2006; Iafrate et al., 2004; Redon et al., 2006) constituting a new form of natural genetic variation. Currently, some known CNV regions (CNVRs) cataloged in the database of genome variants (DGV) (Iafrate et al., 2004) tend to be large and not very clearly delimited regions and also miss smaller CNV (∼10kb) due to the lower resolution of the technologies used in the past (McCarroll et al., 2004; Perry et al., 2008). Moreover, the copy number structure of some of the most highly polymorphic regions have a much higher complexity than what was initially thought (Perry et al., 2008). In order to understand the role of CNVs as a genetic determinant we still require a better characterization and delimitation of these polymorphisms along the entire genome using higher resolution arrays and more accurate detection algorithms. There are several array platforms that can be used to measure CNV. In this article, we focus on the latest high density array platforms that use millions of short oligonucleotide probes distributed along the genome. The high resolution of these arrays is particularly well suited to detect new short CNVs and more accurately delimitate the position and structure of known CNVs. Additionally, some of these probes are also used to target the possible allelic variants of a single nucleotide polymorphism (SNP), making it possible to obtain with the same experiment the SNP and CNV genetic profile of a sample. Commercially available platforms with these characteristics include Affymetrix SNP 6.0 and Illumina Human 1M microarrays. The basic premise for copy number detection is that the hybridization intensities of probes falling under a CNV region will have higher (or lower) values than those expected on a non-copy number variant (non-CNV) region. However, probe hybridization intensities also depend on other non-copy number-related events, which can be regarded as experimental noise that makes CNV detection a challenging problem. Correct estimation of a reference hybridization intensity expected on a non-CNV probe is essential to consider that the noise has zero mean.

Existing CNV detection algorithms assume that the measurements are unbiased (i.e. the noise distribution is centered at zero) and require a separate preprocessing step for normalization and reference extraction (Fig. 1A). Different normalization approaches have been proposed, including CNAT (Huang et al., 2004), dChip (Zhao et al., 2004), CNAG (Nannya et al., 2005), GEMCA (Komura et al., 2006),
workflow using a joint estimation model for CNVs and the probe et al. 

While the spatially correlated portion of the bias can be removed by these preprocessing methods (Global effects Normalization in Figure 1A) there is still a probe-specific bias due to its own binding affinity that needs to be corrected. This is usually done by taking a robust average (trimmed mean, median, clustering) across a set of reference samples. However, if the aim is to identify CNVs that are frequent in population, the application of median normalization on reference robust average (trimmed mean, median, clustering) across a set of regions of the genome with complex CNV patterns. Clearly defined methodology currently available to establish a good priori, would lead to biased results. This has been already pointed out as potentially problematic by several authors (Bengtsson et al., 2008; Komura et al., 2006; Rigaill et al., 2008), and there is no probe hybridization intensity due to altered copy number. $r_s$ is the reference probe hybridization intensity expected for a non-CNV state and $\epsilon_{\text{sm}}$ is a zero-mean array noise.

Using this model, we created an artificial microarray dataset with $N=20$ samples and one single chromosome with $M=10,000$ probes. We generated a copy number profile $x_{\text{sm}}$ containing two CNVRs of different type (Fig. 1A). The first one is a region with aligned CNVs (breakpoints occurring at same position across samples). The second type is a region with non-aligned CNVs. The non-CNV portion of the genome, has an expected log2-ratio $\gamma_{\text{sm}} = 0$. The first variation is chosen to represent a loss $x_{\text{sm}} = -1$ (copy number 1), and the second variation is chosen to have copy number 3 $x_{\text{sm}} = \log_2(3) - 1 = 0.58$.

The noise introduced in the dataset is white i.i.d. Gaussian $\epsilon_{\text{sm}} \sim \mathcal{N}(0,1)$. The major difference of this simulation model as compared with others (Lai et al., 2005; Willenbrock and Fridlyand, 2005) is the inclusion of the probe hybridization bias effect $r_s = 0.5 \sin(2\pi x + 0.1) + A(0.25)$ with two main components: (i) a sinusoidal wave with spatial correlation similar to what has been observed in some actual array experiments (Marioni et al., 2007), and (ii) a noise wave without spatial correlation simulating each probe specific affinity. Supplementary Materials includes additional probe hybridization bias effects including scenarios where $r_s$ changes in amplitude or across different batches.

The preprocessing software packages that are available for the Affymetrix SNP Array 6.0 platform include: the Affymetrix Genotyping Console 3.0, the Affymetrix Genotyping Console GTC3.0.1. with GC correction. The algorithm has been implemented in C, and tested using Matlab. There is also an R package in preparation that will include the new model.

### 2.2 Affymetrix SNP 6.0 data description and normalization

The Affymetrix dataset contains 270 samples from the International HapMap Project consisting of 30 CEPH trios (CEU), 30 Yoruban trios (YRI), 45 unrelated Han Chinese samples (CHB) and 45 unrelated Japanese samples (JPT). The Affymetrix Genome-Wide Human SNP Array 6.0 integrates about 1.9 million probesets (931 946 SNPs, 946 000 CN).

The preprocessing software packages that are available for the Affymetrix SNP Array 6.0 platform include: the Affymetrix Genotyping Console 3.0 (GTC3) with a GC content normalization (see Supplementary Materials and normalization (GADA-IRN), with the currently used approach of using the median to compute the reference hybridization (GADA with separate median normalization, GADA-SMN), and with the Affymetrix Genotyping Console GTC3.0.1. with GC correction. The presented results demonstrate that the detected CNV are significantly more consistent within the HapMap family trios. CNV detected by the new approach are also less variable if we change the set of samples used to create the reference. Finally, the computation time to fit the new model GADA-IRN is very competitive, since the resulting algorithm (as was the case for GADA-SMN) has complexity that grows linearly with the number of samples and probes.

2 MATERIALS AND METHODS

All the samples used to evaluate the methods described in this article are publicly available as the Affymetrix SNP 6.0 Dataset (Affymetrix, 2007). The implementation for the new GADA method is publicly available at http://biron.usc.edu/piquereg/GADA. The algorithm has been implemented in C, and tested using Matlab. There is also an R package in preparation that will include the new model.
et al. (2001) and Aroma.Affymetrix which implements the CRMA (Bengtsson et al., 2008). The CRMA method was chosen, since it has been shown to be more robust and accurate than other methods (Bengtsson et al., 2008). The Aroma.Affymetrix package performs the following four correction steps: (1) allelic crosstalk calibration, ACC, for SNP probes; (2) probe-level modeling PLM, which gives a single signal for the SNP probes; (3) fragment length normalization, FLN, which corrects the differences in the PCR reaction due to the length and GC content of the fragment; and finally, (4) log ratio extraction, LRE, which calculates the log hybridization intensity relative to the expected diploid signal reference intensity using the median. In Figure 1A, the steps ACC, PLM and FLN are grouped together as global effects normalization, and LRE is the probe bias correction.

In this article, we argue that while Steps 1–3 may be performed safely during preprocessing, finding the non-CNV reference intensity is problematic in regions rich in copy number polymorphisms. In normal samples, most of the probes (>90%) will fall on regions without CNV which implies that the normalization model parameters which have a global effect on all probes can be safely estimated using robust strategies as those employed by CRMA (Bengtsson et al., 2008). On the other hand, in any given region of the genome containing a highly polymorphic CNV it is not known a priori which samples do not have a CNV, and this CNV effect should be removed before estimating the probe hybridization intensity associated with the non-CNV state. In this article, we will extract the corrected probe intensities after Steps 1–3, and use our new proposed model (GADA-JRN) to jointly estimate the reference and the copy number component (Fig. 1B). Results from Aroma.Affymetrix GTC software with GC correction are also obtained for which samples do not have a CNV; and this CNV effect should be removed.

There are two basic premises that allow the joint estimation of the reference \( r \) and the copy number \( x \) component. First, the copy number component \( x \) on each sample is PWC with a small number of breakpoints, and can be efficiently represented with \( x = \eta \varepsilon \). It should be noted that the number and position of those breakpoints is in general different for each sample. Second, the probe hybridization bias (or non-CNV reference intensity) \( \eta \) is not necessarily PWC, but is exactly the same across multiple arrays. Our model could also be extended for cases in which the amplitude of \( \eta \) may change, i.e. \( \eta_{mn} = \eta_{0} + \delta_{mn} + \epsilon_{mn} \) (See Supplementary Materials).

The copy number component is modeled using and independent SBL observation model for the log-hybridization of each probe on sample \( n \) can be rewritten as:

\[
y = x + \varepsilon + \eta
\]

(4)

The properties of the SBL prior are detailed in Tipping (2001), Wipf and Rao (2004) and Pique-Regi et al. (2008b). Setting \( b=0 \) and \( a \) to be small encourages a sparse number of breakpoints but is uninformative with respect to the number of breakpoints delimiting CNV \( K \). The number of breakpoints \( K \) is very small compared with the number of probes, making it possible to identify the most likely regions of CNVs. The second step uses a backward elimination (BE) procedure to statistically rank the identified CNVs, allowing a flexible control of the false discovery rate (FDR). Compared with other state-of-the-art methods, using standard evaluation datasets and benchmarks (Willenbrock and Fridlyand, 2005), GADA obtained the highest accuracy and was 100 times faster.

2.4 GADA model with joint reference intensification

The new proposed model does not assume that the probe hybridization bias for each probe \( \eta_{0} \) in (1) has been removed, it is instead estimated jointly with the copy number from a large number samples. In vector form, the observation model for the log-hybridization of each probe on sample \( n \) can be rewritten as:

\[
x = \eta^{c}\varepsilon + \eta^{a}
\]

(2)

In other words, comparing the models in (1) and (2), if the reference \( \eta_{0} \) were known, we could move it to the left side and \( \beta_{0} = \beta_{0} \) would become the log-ratio intensities with only two remaining variations: the CNV effect \( \varepsilon \), and the zero-mean hybridization noise \( \eta_{0} \). The \( n \) in the notation can now be dropped because each sample \( n \) can be analyzed separately. Figure 2 shows an example to illustrate that \( \eta^{a} \) is PWC.

In GADA, a compact linear algebra representation for these PWC vectors \( x=(x_{1}, \ldots, x_{M})^{T} \) was introduced:

\[
x = F w
\]

(3)

which represents a linear combination of step vectors \( F_{j} \) (columns of \( F \)) as in Figure 3. Using this representation, any PWC vector \( x \) with \( K \) breakpoints
In GADA-JRN, \( \tau \) is considered unknown and will be jointly estimated with the copy number estimate.

### 2.5 Fitting the model with the EM algorithm

The model parameters are estimated by finding the maximum \( a \) posteriori (MAP) using a similar evidence maximization procedure as was used in our previous work. The EM algorithm starts by setting \( \nu_0 \) and \( \sigma_0 \) to zero; then, it iterates the following E and M steps:

**E-step:**

\[
E_p: \quad \gamma_n = (\mathbf{w}_n^T \mathbf{F} + \text{diag}(\nu_n))^{-1}
\]

where \( F(w'y, \mathbf{r}, \alpha_m) = \mathcal{N}(w'y, \Sigma) \) exploiting the conjugacy properties between the gamma and normal distributions. The same notation as in our previous work is used and the super/sub-scripts \( n \) and \( m \) are added to identify the parameters that correspond to each sample and, respectively. For example, \( \Sigma_{nm} \) refers to the diagonal terms of the covariance matrix for the breakpoint weights \( \mathbf{w}_n \) posterior distribution of sample \( n \). Convergence of the model is reached with very few EM iterations; and all required operations in each iteration can be performed in a linear number of steps \( O(MN) \) exploiting the properties of our PWC representation (i.e. the matrix structure for \( \mathbf{F} \)).

In relation to the previous GADA model (Pique-Regi et al., 2008b), if the new \( r \) component was modeled by a fixed point estimate (e.g. the median across samples), then the entire model will be completely equivalent to processing each sample independently using GADA (i.e. GADA-SMN). This can also be seen on the EM steps, if \( r \) is a fixed point estimate then (7–11) can be updated separately for each \( n \). Therefore, in GADA-JRN the CNV are placed independently in each sample and the coupling is only through the estimation of \( r \).

During the EM algorithm most of the weight parameters \( \nu_n = \mu_n \) are driven to 0 to fulfill the sparseness constraints imposed by the hierarchical prior. Upon convergence to zero, the corresponding weight parameter and hyperparameter can be eliminated from the model and the EM algorithm can continue with a model of reduced dimensions. This elimination makes the algorithm run faster since it has to update fewer parameters (i.e. only those that are non-zero).

### 2.6 Backward elimination

The sensitivity versus FDR tradeoff of our model is controlled by the hyperparameter \( a \). For higher values of \( a \), a more sparse solution with fewer CNVs is obtained, reducing both the FDR and the sensitivity. In order to efficiently explore solutions with different levels of sparseness without having to run the algorithm all-over again, the same BE strategy described in our previous work is employed. Separately, for each of the samples the breakpoints with the lowest scores are recursively removed:

\[
I_{\text{Be}} = \left\{ \frac{\mu_n^2}{\sum_{\nu_n}^} \right\}
\]

The sensitivity versus FDR tradeoff is controlled by stopping the procedure when all the remaining breakpoints have a score higher than a given critical value \( \tau \). Therefore, as the algorithm continues to remove all the breakpoints and keeps track of which score \( \tau \) a particular breakpoint is eliminated, the breakpoints can be rapidly adjusted to any desired level based on their rank. The final result is reported as a set of segment breakpoints and amplitudes that represent the CNVs. The parameter \( \tau \) is physically more informative than the parameter \( a \) because it can be interpreted as the standard error the user is willing to tolerate to call a CNV significant. More information on how to set \( \tau \) and \( a \) can be found in our previous work (Pique-Regi et al., 2008b) and its supplementary material.

### 2.7 Performance metrics and evaluation methods

In order to compare the performance of the different approaches for combining normalization of the non-CNV probe reference intensities and copy number detection, the following methods and performance metrics are introduced. From the Affymetrix dataset 270 samples are randomly partitioned into reference sets of different sizes (10, 20, 30, 45, 70, 90 and 135 samples). In each partition, one given sample would have been grouped with a different set of samples. For example, if we create 10 random partitions into 9 groups of size 30, each sample will have been grouped randomly with 29 different samples of the remaining 269 samples.

The Affymetrix dataset used in this analysis were processed under ideal conditions (i.e. processed on the same day using three plates; personal communications). Thus, we expect copy number estimates \( \hat{F}_2 \) would be the same regardless of the subgroup chosen as the reference. Although, significant changes in \( F_2 \) are not observed in this dataset, it is noteworthy that the \( \hat{F}_2 \) will likely change with various laboratory conditions or across different batches. Section 4 will address possible methods to analyze samples from different batches. Variance in non-copy number and copy number estimates \( \hat{V}_2 \) across different subgroups \( g \) can be used to assess the performance, with smaller variance indicating better performance.

\[
V = \frac{\text{median}_n \left( \sum_{g=1}^G (\hat{F}_2 - \bar{F}_2)^2 \right) }{\text{median}_n \left( \sum_{g=1}^G (\hat{V}_2 - \bar{V}_2)^2 \right) }
\]

Since this dataset contains 180 samples that are related in family trios, we also propose an additional measure of trio consistency. Identified CNVs in each HapMap trio are classified for each probe as in Table 1. Then, the failed trio consistency rate (FTCR) metric is defined as the ratio of inconsistent CNV probes in a trio among all identified CNV probes (except those considered uninformative).

\[
\text{FTCR} = \frac{I}{C + I}
\]

A smaller FTCR value indicates that copy numbers within a family trio are more consistent. This measure ignores less frequent scenarios, e.g. if both mother and father have one chromosome with a CNV gain and the other chromosome without variation, it will still be possible (25% of the time) for

### Table 1. Consistency on HapMap trios

<table>
<thead>
<tr>
<th>Offspring</th>
<th>Father and Mother pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{C} )</td>
<td>( \hat{L} )</td>
</tr>
<tr>
<td>Gain</td>
<td>C</td>
</tr>
<tr>
<td>Loss</td>
<td>I</td>
</tr>
<tr>
<td>Neutral</td>
<td>I</td>
</tr>
<tr>
<td>FTDR</td>
<td>I</td>
</tr>
</tbody>
</table>
The currently used approach of separate preprocessing estimates the artificially generated data (see Section 2.1) illustrates a scenario from randomly grouping unrelated samples into trios. Additionally, GADA-JRN can better detect the small CNV on the son to not inherit the CNV. In order to assess the validity of the FTCR measure, the FTCR scores of true trios are compared with those obtained from randomly grouping unrelated samples into trios.

3 RESULTS

3.1 Results with simulated data

The artificially generated data (see Section 2.1) illustrates a scenario in which there are two relatively high-frequency CNVs (Fig. 4A) with a bias on the hybridization measurement from the array experiment (Fig. 4B). If the probe hybridization bias $r$ is not removed from the data, the results will be contaminated with a large number of spurious segments not associated with true CNVs (Fig. 4C). While other approaches (Bengtsson et al., 2008; Diskin et al., 2007) can correct the 'smooth-wave' (GC correlated) part of the bias (see next section), GADA-JRN can also correct the non-smooth (uncorrelated) probe-specific bias. The currently used approach of separate preprocessing estimates $r_{np}$ as the median across a set of reference samples (here the simulated samples themselves) before extracting the CNVs. This can eliminate $r_{np}$ on the areas of the genome without CNVs ($r_{np} \approx 0$ regions); but it is problematic on CNVRs containing a large amount of CNV across samples (Fig. 4D and E). The new joint reference normalization approach (GADA-JRN in Fig. 4F) can correctly delimitate the CNV on the samples with CNV on the region CNVR-2, while the separate median normalization (GADA-SMN in Fig. 4E) incorrectly reports CNV on samples $n = 1, \ldots, 10$. Additionally, GADA-JRN can better detect the small CNV on CNVR-1, in which the separate median normalization tends to make the amplitude of the variation smaller and thus more difficult to detect.

3.2 Results with Affymetrix microarray data

The hybridization intensities are obtained after using ACC and applying FLN corrections from Aroma.Affymetrix, on the 270 HapMap samples analyzed with Affymetrix SNP 6.0 arrays. This preprocessing step appropriately scales and centers the data removing the spatially correlated part of the bias. We next compare GADA-JRN and GADA-SMN employing the evaluation methods introduced in Section 2.7.

Using the randomly chosen reference sets of different size, we evaluated the variability $V_r$ and $V_x$ in the reference intensities and the copy number estimates. Figure 5 shows that as the number of samples in the reference set increases the variability on CNV estimates $V_r$ decreases. More importantly, we can see that using GADA-JRN achieves a considerably better performance when compared with GADA-SMN. In some cases GADA-JRN requires about half of the samples in order to obtain estimates of similar accuracy to those achieved with GADA-SMN. In terms of variance or the reference intensity estimates $V_x$, GADA-JRN also achieves a
significantly smaller values ($P < 10^{-7}$ Kolmogorov–Smirnov test, data not shown).

This improvement in performance can also be observed using the trio consistency measure (FTCR) described in Section 2.7. In Figure 6, the trio consistency improves (i.e. FTCR decreases) with the size of the reference set, and GADA-JRN also achieves significantly better consistency. The results are also similar with change on the sparseness parameters $\alpha$ and $T$ that set the tradeoff between sensitivity and the FDR. Figure 7 illustrates for one of the reference sets (90 CEU samples) the consistency that is obtained for different settings of the parameter $T$, which controls the BE step. The FTCR measure improves (decreases) with increasing $T$, since the number of detected false CNVs (i.e. FDR) decreases; but for higher values of $T$ true CNVs may also fail to be detected on the offspring (i.e. lower sensitivity), and thus FTCR stops decreasing. The FTCR obtained on randomly formed trios of unrelated samples assures that this metric is actually capturing the increase on shared CNVs within a family ($P < 0.01$) and can be used to compare different normalization and copy number detection approaches. In Table 2, GADA-JRN obtains a better FTCR (16.7%) than GADA-SMN (19.5%) and GTC (19.45%) when using 90 reference samples. On a larger reference set of 180 samples, GADA-SMN (FTCR=18.3%) and GTC (FTCR=17.31%) improve but GADA-JRN still retains a better FTCR (16.5%). Overall, the new approach is especially indicated for small reference sets and for regions with highly polymorphic CNVs. Figure 8 shows the copy number estimates obtained on an already known highly polymorphic region of chromosome 17. The predicted gains and losses of GADA-JRN are retained when the reference set of 90 reference (CEU) samples is enlarged to include 180 (CEU+YRI) samples. Finally the computational time required for fitting the model is longer on the new approach, but still retains a very competitive linear complexity in the number of probes and samples (Fig. 9).

### 4 DISCUSSION

The application of the proposed GADA with joint reference estimation is not only limited to Affymetrix SNP arrays but it can also be applied to other platforms such as Illumina BeadStudio or NimbleGen array comparative genomic hybridization (aCGH). In Illumina BeadStudio the probe hybridization intensities $I$ (obtained after ACC) can be extracted instead of the LRR values used.

<table>
<thead>
<tr>
<th>Number of reference samples</th>
<th>90 CEU</th>
<th>180 CEU+YRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTCR (%) No. of CNVs</td>
<td></td>
<td>FTCR (%) No. of CNVs</td>
</tr>
<tr>
<td>GADA-JRN</td>
<td>16.7</td>
<td>16.5</td>
</tr>
<tr>
<td>GADA-SMN</td>
<td>19.2</td>
<td>18.3</td>
</tr>
<tr>
<td>GTC</td>
<td>18.45</td>
<td>17.31</td>
</tr>
</tbody>
</table>

FTCR and the median number of CNVs per sample when using three different algorithms GADA-JRN, GADA-SMN and GTC are shown in the table. The threshold values for GADA-JRN and GADA-SMN were chosen to obtain approximately equal number of CNVs among the algorithms. The lowest FTCR values are highlighted in bold.

### Table 2. Comparison on HapMap trio consistency FTCR

**Fig. 5.** Variability on the copy number estimates if the set of reference samples changes.

**Fig. 6.** Consistency of the copy number estimates on HapMap trios if the set of reference samples changes.

**Fig. 7.** Consistency within HapMap trios using a different sparseness setting $T$. The dashed and solid lines correspond to a 90 (CEU) and 180 (CEU+YRI) sample reference set, respectively. The cloud of points are the FTCR values obtained from 100 randomly formed trios. The FTCR values of GADA-JRN (blue) are smaller than those of GADA-SMN (green).
Estimation of CNV and reference intensities using GADA

Fig. 8. Section of the chromosome 17 that contains an already known CNV. Each row corresponds to one of the 90 CEU HapMap samples and are grouped in trios (father, son/daughter, mother) delimited by horizontal dotted lines. On the left of the thick vertical line are shown the CNVs estimated using GADA-SMN using a reference set of 90 and 180 reference samples. On the right, copy number estimated using GADA-JRN shows a higher consistency when the reference set is changed as well as within HapMap trios.

Fig. 9. Computational time required to fit the model is linear on the number of samples for both approaches. Execution times required to process the models are measured on the same machine.

expected \( R \) values of non-CNVRs (Peiffer et al., 2006) which have similar drawbacks as separate median normalization. In aCGH, the reference DNA from a single sample or a pool of samples is used as a reference, but these log-ratio intensities may still contain a remaining ‘wavy’ artifact (Marioni et al., 2007) that our proposed approach could eliminate.

ITALICS (Rigaill et al., 2008) is an iterative approach that alternates two separate steps of copy number detection and normalization. The iterative concept is similar to the EM algorithm employed to fit the GADA-JRN model in this article, but there are two fundamental differences. First, ITALICS operates on a single sample and only includes a small set of parameters correcting for global array effects such as the fragment length and GC content; ITALICS considers that the reference non-CNV probe intensity is fixed and extracted from a separate reference set of samples. Second, the copy number extraction and the normalization model are iterated in practice only twice and not integrated under an unifying probabilistic model as in GADA-JRN. In contrast, GADA-JRN first proposes a multiple sample probability model which include parameters for the CNV component and the reference non-CNV probe hybridization intensity of every position of the genome; and then an iterative approach to fit the model is derived using the EM algorithm, which guarantees that the parameters converge.

GEMCA (Komura et al., 2006), only available for Affymetrix 500K platform, approaches the problem by finding the reference after copy number detection. First, the copy number is estimated on the difference between all the possible pairs. The CNVRs are defined as those identified in certain number of pairs. Then the largest subgroup of samples with no relative variations (found using a maximum clique algorithm) is used to establish
the reference set for that particular region of the genome. The Canary algorithm in Birdsuite (Korn et al., 2008) uses a Gaussian mixture model (GMM) mixture model to identify a posteriori the copy number variation state of already delimited regions of CNV. In both GEMCA and Canary, the underlying assumption that CNVs are tightly aligned across samples and do not overlap with other possible CNVs represents a challenge in dealing with complex polymorphic CNVRs of the genome (e.g. Fig. 8).

GADA-JRN may also prove to be a flexible algorithm in variety of other applications. For example, samples processed by different labs or on different days may not necessarily have the same non-CNV reference hybridization intensity even if the platform is exactly the same. GADA-JRN could be applied to assess if this effect has changed between different experimental batches. Moreover, one or more control samples replicated across experimental batches could be used to better characterize and correct the probe hybridization noise.

In addition to the probe hybridization reference, other parameters could also be introduced to the GADA framework to model other known sources of variation. For example, allowing changes in scale of the bias (see Supplementary Materials), using an independent bias correction for each SNP allele, and including probe-specific noise variances (heteroscedastic model). The residuals of the model can be used to assess the impact of these other sources of variation on the results. In particular, correlation on the model residuals can be used to discover hidden batch effects indicating the need for subgroup analysis (see Supplementary Materials).

Finally, the GADA-JRN model could be used in combination with the N-GADA model (Pique-Regi et al., 2008a) that the authors proposed for modeling breakpoints of CNVs across multiple samples.

5 CONCLUSION

In this article, we introduce a new method in which the reference probe hybridization intensity is jointly estimated with the copy number component. This type of methods are essential for the characterization of CNV on normal population using the latest array technologies, in which the underlying genome CNVs are not known a priori. The currently used approach of separate array technologies, in which the underlying genome CNVs are not known a priori. The currently used approach of separate

ACKNOWLEDGEMENTS

The authors would like to thank the two anonymous reviewers for valuable comments and suggestions.

Funding: NIH’s Child Health Research Career Development Award Program (SA) (TK12-CA60104); Alex’s Lemonade Stand Foundation (SA).

Conflict of Interest: none declared.

REFERENCES


