Estimation and assessment of raw copy numbers at the single locus level

H. Bengtsson1,*, R. Irizarry2, B. Carvalho2 and T. P. Speed1,3

1Department of Statistics, University of California, Berkeley, USA, 2Department of Biostatistics, Johns Hopkins University, Baltimore, USA and 3Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia

Received on September 18, 2007; revised on December 25, 2007; accepted on January 9, 2008
Advance Access publication January 19, 2008
Associate Editor: Alfonso Valencia

ABSTRACT
Motivation: Although copy-number aberrations are known to contribute to the diversity of the human DNA and cause various diseases, many aberrations and their phenotypes are still to be explored. The recent development of single-nucleotide polymorphism (SNP) arrays provides researchers with tools for calling genotypes and identifying chromosomal aberrations at an order-of-magnitude greater resolution than possible a few years ago. The fundamental problem in array-based copy-number (CN) analysis is to obtain CN estimates at a single-locus resolution with high accuracy and precision such that downstream segmentation methods are more likely to succeed.

Results: We propose a preprocessing method for estimating raw CNs from Affymetrix SNP arrays. Its core utilizes a multichip probe-level model analogous to that for high-density oligonucleotide expression arrays. We extend this model by adding an adjustment for sequence-specific allelic imbalances such as cross-hybridization between allele A and allele B probes. We focus on total CN estimates, which allows us to further constrain the probe-level model to increase the signal-to-noise ratio of CN estimates. Further improvement is obtained by controlling for PCR effects. Each part of the model is fitted robustly. The performance is assessed by quantifying how well raw CNs alone differentiate between one and two copies on Chromosome X (ChrX) at a single-locus resolution (27kb) up to a 200kb resolution. The evaluation is done with publicly available HapMap data.

Availability: The proposed method is available as part of an open-source R package named aroma.affymetrix. Because it is a bounded-memory algorithm, any number of arrays can be analyzed.

Contact: hb@stat.berkeley.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Due to high demand and strong competition, the number of loci interrogated by DNA microarrays has in recent years been pushed to levels that are of orders of magnitude greater than were available only a few years ago. In 2003, Affymetrix released a high-density oligonucleotide array named Mapping 10 K targeting genotyping and estimation of copy numbers (CNs) of 10 K (10,000) single-nucleotide polymorphisms (SNPs). Since then Affymetrix has released platforms that interrogate 100 K SNPs and 500 K SNPs (‘SNP chips’). More recently, Affymetrix released the Genome-wide SNP 5.0 and Genome-wide SNP 6.0 chip types (‘SNP & CN chips’). In addition to SNPs, these interrogate a large number of non-polymorphic (NP) loci targeting CN analysis by (often single) CN probes. The 5.0 chip has the same set of SNPs as the 500 K chip as well as 400 K NP loci. The 6.0 chip interrogates 900 K SNPs and 900 K NP loci. Illumina has released BeadArray platforms interrogating 240, 300, 370, 550, 650 and recently 1000 K loci. It is reasonable to expect that this competitive growth in density of markers will continue for at least another couple of years.

The Affymetrix SNP chips consist of a large number of units (‘probe sets’), each interrogating a particular SNP. Each unit consists of a set of 25-mer probe pairs quantifying the amount of allele A and allele B DNA target. Although aligned between the two alleles, the sequences across probe pairs are slightly shifted relative to each other such that they together cover a wider region around the SNP loci than if all pairs would be centered at the SNP loci. The number of probe pairs per unit and the amount of probe-pair shift varies with chip type and also between SNPs. For the SNP & CN chip types, the probe pairs are no longer shifted relative to each other, but instead they are technical replicates. Moreover, for the 5.0 chip type the two alleles within a probe pair are not necessarily aligned. Because of these updated chip designs, the model introduced in this article should only be applied to the SNP & CN chips with great care. We are currently developing alternative method that we believe are better suited for these new chip types.

We have developed a preprocessing method for estimating raw CNs from Affymetrix SNP arrays. By raw CNs we mean single-locus CN estimates on a continuous scale that have been obtained independently of each other in the sense that the information of their chromosomal positions have not been utilized. Raw CNs constitute the input to downstream methods for identifying chromosomal regions with CN aberrations, e.g. CN segmentation methods. The purpose of our method, referred to as ‘Copy-number estimation using
Robust Multichip Analysis’ (CRMA), is to provide downstream methods with raw total CNs of as high accuracy and precision as possible, something which is likely to make those methods perform better.

The outline of this article is as follows. In Section 2, we give further details on the layout of Affymetrix SNP chips together with details on the most important systematic effects observed in these chip types. In Section 3, in addition to describing the data set used for the analysis, the CRMA method is given in detail. Other existing methods used for comparison are also described. At the end, the methods used for evaluation are explained. In Section 4, results are given. Because parts of the preprocessing can also be applied to genotyping, their outcomes are presented first. Thereafter, the method is evaluated and compared with the other methods. The proposed method and the results discussed in Section 5. Conclusions and future research directions are given in Section 6.

2 APPROACH

Let \( C_i \) be the true number of DNA copies in a sample \( i = 1, \ldots, I \) at locus \( j = 1, \ldots, J \) with position \( x_j \) on some chromosome. In normal individuals, with the exception for CN polymorphic regions, we expect that \( C_i = 2 \) for loci on autosomal chromosomes, as well as on ChrX in females. Furthermore, if the sample is homogeneous, we expect \( C_j \) to be discrete, otherwise continuous. One fundamental assumption in CN analysis is that loci close together are likely to have the same underlying CN.

2.1 Design of SNP arrays

Ignoring so-called mismatch (MM) probes, a SNP on an Affymetrix SNP chip is interrogated by a set of perfect-match (PM) probe pairs \( \{PM_{jkA}, PM_{jkB}\} \) where each probe pair is indexed by \( (j,k) \) for \( k = 1, \ldots, K_j \), where \( K_j \) is the number of probe pairs for SNP \( j \). The number of probe pairs per unit varies between different chip types and in some cases also between units.

2.2 Crosstalk between allele A and allele B

The PM\(_{jkA}\) and PM\(_{jkB}\) probes differ only by one nucleotide in the position that corresponds to the SNP locus in the genome. Their close similarity give rise to crosstalk between the two allele signals. Due to the way Affymetrix label a probe allele A or allele B, there are six possible nucleotide pairs at the interrogating position on the forward strand, namely \( (A,C), (A,G), (A,T), (C,G), (C,T), (G,T) \). The corresponding ones on the reverse strand are \( (T,G), (T,C), (T,A), (G,C), (G,A), (C,A) \). Note that the two sets become identical if the alleles on one of the strands are swapped. Allelic crosstalk can be observed when plotting the probe-pair signals stratified by nucleotide pair as illustrated in Figure 1 and further explained in Section 3.2.1.

2.3 Sequence-dependent effects

Ishikawa et al. (2005) and Nannya et al. (2005) observed a bias in the observed CNs (log ratios) with a strong dependency on PCR fragment length and GC content. During a PCR cycle, shorter DNA sequences are more likely to be fully replicated compared with longer ones. The outcome is that the longer the extracted fragments are, the less target DNA gets hybridized, and the weaker the observed signals become. The number of G and C nucleotides in a probe has a similar effect on the efficiency of the PCR and the hybridization to the probes. As further explained in Section 3.2.5, if these effects vary between samples, they give rise to non-linear effects in the observed log ratios.

3 METHODS

3.1 Data set

Mapping250K_Nsp CEL files (dated July 28 and 29, 2006) for the 30 male and 30 female CEU founders were downloaded from The International HapMap Project (Altshuler et al., 2005; The International HapMap Consortium, 2003). Offsprings were excluded in order to avoid biological relationships. Because female NA12145 has low true ChrX CN level (Ting et al., 2006), it was excluded from the evaluation step.

3.2 Proposed model

The CRMA method pre-processes the probe signals, summarizes them, post-processes the summarized signals, and calculates raw CNs relative to a reference. As described in detail below, the main pre-processing method is allelic crosstalk calibration, with the option to perform a follow-up quantile normalization. A multichip log-additive model fitted robustly is used to summarize the probe signals. Summarized signals are normalized for sequence-dependent effects that are related to main covariate fragment length with the option to also include covariate GC content. For paired analysis, a copy-neutral sample should be used as a reference. For non-paired analysis, a robust average across a pool of samples, where the majority are copy neutral, may be used as a reference. Here we use the pool of all samples.
3.2.1 Calibration for offset and crosstalk between alleles

By swapping the alleles for the reverse-strand probe pairs, we can model these together with the corresponding probe pairs on the forward strand, leaving us with the six unique classes identified in Section 2.2. Let \( x_{ik} = (x_{ikA}, x_{ikB}) \) and \( y_{ik} = (y_{ikA}, y_{ikB}) \) be the true and observed signals for probe pair \((j, k)\) in sample \(i\). Let \( \{x_{ikA}\} \) and \( \{y_{ikB}\} \) be the corresponding signals with the probe pairs on the antisense strand swapped. For a particular nucleotide pair, we model the allelic crosstalk and shift observed in \( \{y_{ikB}\} \) by a sample-specific affine transformation as

\[
y'_{ikB} = a_i + S_i x_{ikB} + e_{ikB},
\]

where \( a_i = (a_{iA}, a_{iB})^T \) denotes the offset,

\[
S_i = \begin{bmatrix} S_{iAA} & S_{iAB} \\ S_{iBA} & S_{iBB} \end{bmatrix}
\]

is the crosstalk matrix, and \( e_{ikB} \) is random noise. Because either \( x_{ikA} \) or \( x_{ikB} \) is zero for homozygote \((i, j)\) data points, it is possible to estimate \((a_i, S_i)\) from the observed data \( \{y_{ikB}\} \) stratified by nucleotide pair. We constrain \( S_i \) such that for a subset of probe pairs that are likely to be diploid the average of \( \{x_{ikA}\} \) and the average of \( \{y_{ikB}\} \) are an equal constant \((=2000).\) This constant should be the same for all arrays in order to bring the arrays to the same scale.

To obtain estimates \((\hat{a}, \hat{S}),\) we utilize a robust simplex fitting algorithm (Wirapati and Speed, 2001) implemented in the \(R\) package \textit{sfit} (http://www.braju.com/R/). One advantage of this algorithm is that it is also robust against outliers in between the two homozygote arms (Fig. 1), which is the case for samples with CN gains. Estimates \( \{\hat{x}_{ikB}\} \) of the true signals (up to a multiplicative constant) are obtained by backtransforming as

\[
\hat{x}_{ikB} = \hat{S}_i^{-1}(y'_{ikB} - \hat{a}_i),
\]

and by reversing the swap of the probes on the antisense strand.

It is obvious why different nucleotide pairs should contribute with different offsets. An alternative is to constrain Model (1) to have a common offset across nucleotide pairs and alleles. The small differences in offset estimates observed between groups (see Supplementary Materials) support the latter. One rationale for such a model is that there exists a scanner offset (Bengtsson et al., 2004) that dominates other offsets. However, due to limitations in the implemented algorithm available we use Model (1) and assume the model error is small.

3.2.2 Normalization for differences in probe signal distributions

Depending on the experimental setup, a reasonable assumption is that the distribution of the true amount of DNA targets over all probes should be the same across samples. This should also be true for the observed signals when all systematic effects have been removed. If this is not the case, probe signals may be transformed further by a process called quantile normalization, such that this assumption is met.

As illustrated in Section 4.1, the affine transform of the above allelic crosstalk model in Equations (1)–(3) explains most of the differences observed in the empirical distributions of the observed probe signals. However, if for reasons other than offsets and allelic crosstalk, there exist differences in signal distributions, an option is to apply quantile normalization following the crosstalk calibration.

To simplify the notation, let \( k' = 1, \ldots, K' \) be the indices for all \( K' \) PM probes on the array. Then, for quantile normalization, the (calibrated) probe signals \( \{\hat{x}_{ik'k'}\} \) are transformed for each array \(i = 1, \ldots, I\) such that their distribution become the same as a given target distribution. Various methods have been proposed to normalize empirical distributions, cf. Bolstad et al. (2003). For reasons explained below, instead of modeling the empirical distributions explicitly, we choose to normalize signals \( \{\hat{x}_{ik'}\} \) as follows:

**Pass A:**

A1. Choose a subset of the PM probes for which one can assume that the distribution of the true signals are equal across arrays. Let \( K^* \) be the number of such probes.

A2. For each array \(i = 1, \ldots, I\), construct the set \( \{\hat{x}_{ik'k'}\} \) consisting of the ordered above subset such that \( \hat{x}_{i(1)} \leq \hat{x}_{i(2)} \leq \cdots \leq \hat{x}_{i(K^*)}. \)

A3. For each rank \((k^*)\) \(=1, \ldots, K^*\), calculate the target distribution as the average signal across arrays by

\[
\bar{x}_{i(k')} = \frac{1}{I} \sum_{i=1}^{I} \hat{x}_{i(k')}.
\]

**Pass B:**

For each array \(i = 1, \ldots, I\):

B1. Fit a smooth spline \( f_i(\cdot) \) robustly to \( \{(\hat{x}_{ik'k'}, \bar{x}_{i(k')})\}_{k'} \).

B2. For all probes, normalize (backtransform) signals as

\[
\tilde{y}_{ik'} = f_i^{-1}(\hat{x}_{ik'_k}).
\]

This algorithm will make the distribution of probe signals \( \{\tilde{y}_{ik'}\} \) approximately the same across arrays for the subset of probes chosen in Step A1. The distributions are only approximately equal because smooth splines are used in Step B1.

There are at least two advantages with this algorithm compared with existing quantile normalization algorithms. First, because Step A3 uses a summation to calculate the average, it can be calculated incrementally such that only 2\(K^*\) probe signals are kept in memory at any time. As Steps B1-B2 is done for each array independently, the memory complexity of the above two-pass algorithm is \( O(K) \). Second, probes that interrogate loci for which we know that the true CNs differ across samples can be excluded while estimating the normalization function \( f_i(\cdot) \) in Step B1. Note that all probe signals will still be normalized in Step B2. This allows us to exclude for instance ChrX probes in Step A1 to avoid bringing female and male distributions toward each other. Loci known to carry CN polymorphisms may also be excluded for the same reason.

The method just described is not always appropriate, for example, when samples are expected to contain a large number of CN aberrations. If not used, replace \( \hat{x}_{ikB} \) by \( \hat{x}_{ikA} \) in what follows. We will see in Section 4 that this is not necessarily a bad choice.

3.2.3 Construction of non-polymorphic probe signals

In this article, we focus on optimal estimation of total CNs. In other words, we are not interested in the amount of target DNA in the individual strands but only the total amount. For this reason we sum the allele signals at the probe level as

\[
\tilde{y}_{ik} = \hat{y}_{ikA} + \hat{y}_{ikB}
\]

for all samples \(i = 1, \ldots, I\) and all probes \((j, k)\). This is also an approach taken by the dChip tool (private communication). We refer to \( \tilde{y}_{ik} \) as non-polymorphic probe signals to indicate that the signals no longer contain allele-specific information but only information on the amount of DNA at a specific locus. One advantage of summing the signals at the probe level is that we partly avoid the problem of modeling signals close to the signal-to-noise level (‘close to zero’), which becomes a problem especially for homozygote SNPs, as well as for the log-additive model introduced next.
3.2.4 Probe-level modeling Let \( \{ y_{ijk} \} \) be as in Equation (6). For each locus \( j \), we model the signals \( \{ y_{ijk} \}_a \) across samples and across probes by

\[
\log_2 (\hat{y}_{ij} + e) = \log_2 \hat{\theta}_j + \log_2 \phi_k + \xi_{ij},
\]

where \( e \) is a global constant (explained below), parameters \( \{ \theta_j \} \) are the chip effects for samples \( i = 1, \ldots, I \) at locus \( j \), and parameters \( \{ \phi_k \}_l \) are the affinities for probes \( k = 1, \ldots, K \) at locus \( j \), and \( \xi_{ij} \) are zero-mean error terms with variance \( \sigma^2 \). For Equation (7) to be identifiable, probe affinities are constrained such that \( \sum_k \phi_k = 1 \). This model is similar to the one Irizarry et al. (2003) suggested for expression arrays. The model is fitted robustly using M estimators via an iterative reweighted least squares (IWLS) method (Bolstad, 2004) implemented in the affyPLM package (Ben Bolstad, 2007). In cases where the true CN is zero, the corresponding chip effect (and raw CN) cannot be defined on a logarithmic scale.

By allowing for a small bias in the chip effects, which in turn will give a small bias in the raw CNs, we correct for differences in sequence effects which, depending on array, correspond roughly to PM quantiles in Wong et al. (2001; Zhou and Rocke, 2005).

Furthermore, non-positive probe signals in \( \{ y_{ijk} \} \) introduced by the backtransformation in Equation (3), and that remain after the fit and still be able to normalize all loci. Also, when normalizing \( \{ M_{ij} \}_j \), directly there is no obvious way to go back to chip effects if additional normalization of such is wanted. For further clarification, see Bengtsson (2004) and Bengtsson and Hössjer (2006).

3.2.5 Normalization for sequence effects By modeling the means of the chip effects as a smooth function of PCR fragment length and GC content, we correct for differences in sequence effects across arrays observable in the summarized signals (not shown). Let \( \lambda_j \) be the length in basepairs of the restriction fragment containing locus \( j \). These lengths are predetermined by the restriction sites of the enzymes. Let \( \kappa_j \in [0, 1] \) be the fraction of G and C nucleotides for the same sequence. For reasons outlined below, we model the sequence effects on \( \log_2 \hat{\theta}_j \) instead of \( \{ M_{ij} = \log_2 (\hat{\theta}_j / \hat{\theta}_R) \} \), where \( \{ \hat{\theta}_R \} \) are reference signals at each locus (Section 3.2.7). Analogously to quantile normalization, normalization for PCR effects can be performed using a two-pass procedure as follows:

Pass A:

1. Choose a subset of the loci for which one can assume to be diploid in most or in all samples. Let \( J^* \) be the number of such loci. In case non-diploid samples or loci are included by mistake, we rely on the robustness of the estimators.
2. For each locus \( j = 1, \ldots, J^* \), calculate the average log-signal by

\[
\hat{\overline{y}}_j = \text{median} [\overline{y}_j],
\]

where \( \overline{y}_j = (\lambda_j / \kappa_j) \). This constitutes the average PCR effect across samples.

Pass B:

1. For each sample \( i = 1, \ldots, I \):
   - Fit a smooth spline \( g_i(x) \) robustly to \( \{ y_{ij}, \lambda_j, \kappa_j \} \). This constitutes the PCR effect for sample \( i \).

B.2. For all loci, estimate the PCR discrepancies for sample \( i \) by

\[
\hat{d}_i = g_i(x_i) - g_i(\hat{z}_i).
\]

B.3. Finally, normalize all loci as

\[
\log_2 \hat{\theta}_j = \log_2 \hat{\theta}_j - \hat{d}_j.
\]

Steps B1-B3 allow us to use a pre-estimated \( g_i(x) \), which turns the problem into a single-array estimation problem. This is computationally more efficient for large data sets. For further computational efficiency, constrain the fragment-length and GC-content effects to contribute additively (on the logarithmic scale), that is,

\[
g_i(x) = g_{FL}(x_i) + g_{GC}(x_i).
\]

We note that the above normalization could have been done on the log-ratios \( \{ M_{ij} \}_j \) (Equation (12)), cf. Nannya et al. (2005). However, in that case it is not clear how to exclude ChrX and other non-diploid loci from the fit and still be able to normalize all loci. Also, when normalizing \( \{ M_{ij} \}_j \), directly there is no obvious way to go back to chip effects if additional normalization of such is wanted. For further clarification, see Bengtsson (2004) and Bengtsson and Hössjer (2006).

3.2.6 Calculation of raw copy numbers The logarithm of the relative CN estimate, or shorter, the raw CN, for sample \( i \) and locus \( j \), is defined as

\[
M_{ij} = \frac{\log_2 \hat{\theta}_j}{\hat{\theta}_R}.
\]

where \( \hat{\theta}_R \) is a reference (hence the subscript \( R \)) signal at locus \( j \) typically representing the mean diploid signal.

3.2.7 Reference signals In this article, we consider the case of estimating the reference signals \( \{ \hat{\theta}_R \} \) from a pool of samples. Consider locus \( j \). Our objective is to estimate the mean \( \mu_{j2} \) of the diploid signal for this locus such that the estimate has low variance and low bias. Furthermore, consider the set of samples that are truly diploid at this locus. An estimate is

\[
\hat{\mu}_{j2} = \text{median} [\hat{\theta}_j],
\]

where the average is taken over the set of truly diploid samples. Note, we do not know exactly what samples are diploid at locus \( j \), but instead we assume that the majority of the samples are diploid, and rely on the robustness of the estimator. Among the methods used for comparison (Section 3.4), CNAT uses Equation (13), dChip uses a trimmed sample mean, and CNAG uses the sample mean and robustifies by optimizing the selection of diploid samples (Nannya et al., 2005).

For loci on ChrX, the above estimate will be biased if male samples are included. For this reason, both CNAT and CNAG exclude males when estimating the means on ChrX. In dChip this is optional. When excluding non-diploid samples, the estimates \( \hat{\mu}_{j2} \) have larger variances for loci on ChrX than on other chromosomes. An alternative is to utilize non-diploid samples as follows. Let

\[
\hat{\mu}_{j2} - \hat{\mu}_{j2} \]

be the robust average across all non-diploid samples. Then the alternative estimate of the diploid mean is

\[
\hat{\mu}_{j2} = (\hat{\mu}_{j2} + \hat{\mu}_{j2} - \hat{\mu}_{j2}) / 2,
\]

where

\[
\hat{\mu}_{j2} = \text{median} [\hat{\mu}_{j2}] / \text{median} [\hat{\mu}_{j2} - \hat{\mu}_{j2}].
\]
Equations (14)-(16) are applicable to all chromosomes, as well as to cases where some samples are known to be non-diploid, or where the ploidy of some samples are unknown. The approach may also be extended to handle shorter regions with known aberrations but also regions of CN polymorphism (Redon et al., 2006; Sebat et al., 2004; Iafrate et al., 2004).

3.3 Implementation

Not only does the number of loci investigated grow fast, but also the number of samples. Whilst smaller projects still involve 1 to 10s of samples, the larger projects now involve 1000s of samples. In order to handle such large data sets in everyday high-throughput analysis, as well as for developing new methods and evaluating them against existing ones, a new computational framework is needed. We have developed an open-source cross-platform environment in R (R Development Core Team, 2007) named aroma.affymetrix (http://www.braju.com/R/aroma.affymetrix/) that enables us to analyze every-thing from 1 to 1000s of Affymetrix arrays within virtually any given memory constraints. The CRMA method was implemented as part of this framework.

3.4 Raw copy numbers by other models

In order to assess the performance of CRMA, we compared it with dChip (Li and Wong, 2001), CNAT (Affymetrix, 2007), and CNAG (Nannya et al., 2005). For each method evaluated, we used the default settings.

dChip: For the dChip model, we used the dChip 2006 (Build: March 24, 2007; http://www.dchip.org). Probe-level data was normalized using the invariant-set method (Li and Wong, 2001), allele signals for the PMs were summed, and the same for the MMs. Then MBEI scores (Li and Wong, 2001) corresponding to \( \theta_{ij} \) were estimated from \( \max(PM-MM, 1) \). The dChip software can calculate raw CNs from a pool of samples by using a trimmed-mean estimate. However, because there are no documented guidelines on how to choose the trimming parameter in \( [0, 1] \), we choose to export the MBEI scores to R and use the more robust median for the estimates. For differences in using mean and median, see Section 4.4.

CNAT: For the CNAT v4 model, we used gtype-probeset-genotype v1.5.6 with options -analysis quant-norm-sketch = 500000, usepen = true, pm-only, sumz: to obtain estimates of allele-specific chip effects \( \left[ \theta_{ij} \right] \) on an allele-specific version of the log-additive model in Equation (7) fitted using median polish and applied to quantile normalized PM signals. The CNAT tool did not support using the pool of samples as a reference. Instead, a pooled reference was created by taking an identical copy of the CEL where the 60 samples was used as both test samples and reference samples, which was done by first creating an identical copy of the CEL set to be used as the reference set. This approach was taken because CNAG did not permit use of the pool of test samples as a reference set. Signals were extracted and the 60 test and 60 reference samples were labeled male or female. A ‘non-self analysis’ was used to calculate raw CNs. CNAG had to be forced to use all reference samples, because if allowing CNAG to choose an optimal subset of the reference samples, it would choose the single reference sample that is an identical copy of the test sample. This approach is slightly different from the one proposed by Nannya et al. (2005) and the reason for the asterisk.

3.5 Methods for evaluation

We assess the performance of the different methods by testing how well they can differentiate between one and two copies using raw or smoothed CN estimates. For this we use estimates from a set of females and males across the 5608 non-pseudoautosomal SNPs that are available on ChrX (located after position 2.8 Mb). For such a locus \( j \), we assume that \( C_j = C_i = 2 \) if sample \( i \) is a female, and \( C_j = C_i = 1 \) if it is a male.

3.5.1 Single-locus classification

For a given locus \( i \), we classify each sample \( i \) as \( C_j = 2 \) or \( C_j = 1 \) from raw CN estimates \( M_{ij} \) by the rule

\[
\tilde{C}_j = \tilde{C}(t_j) = \begin{cases} 1: & M_{ij} < t_j \\ 2: & \text{otherwise} \end{cases}
\]

where \( t_j \) is an arbitrary threshold. The performance of this classification can be assessed by the Receiver Operator Characteristic (ROC) curve \( \tau \mapsto (\beta(\tau), \gamma(\tau)) \), which describes the relationship between the true-positive (TP) and false-positive (FP) at any given threshold. For a given \( \tau \), we estimate the TP rate \( \hat{\beta}(\tau) \) and the FP rate \( \hat{\gamma}(\tau) \) from \( (C_i, C_j) \) as

\[
\hat{\beta}(\tau) = 1 / P \sum_j \mathbb{I}(M_{ij} < \tau \cap C_j = 1),
\]

\[
\hat{\gamma}(\tau) = 1 / P \sum_j \mathbb{I}(M_{ij} < \tau \cap C_j = 2),
\]

where \( \mathbb{I}(\cdot) \) is the indicator function. See Figure S1 (Supplementary Materials) for an example of ROC curves for two different loci.

3.5.2 Multilocus classification with common threshold

We now assess how well individual estimators from a set of \( J^* \) loci differentiate between one and two copies, on average, when all loci truly have the same copy number. We can then assess how well individual estimators from a set of females can differentiate between one and two copies using raw or smoothed CN estimates. For this we use estimates from a set of females and males across the 5608 non-pseudoautosomal SNPs that are available on ChrX (located after position 2.8 Mb). For such a locus \( j \), we estimate the TP rate \( \hat{\beta}(\tau) \) and the FP rate \( \hat{\gamma}(\tau) \) from \( (C_i, C_j) \) as

\[
\hat{\beta}(\tau) = 1 / P \sum_j \mathbb{I}(M_{ij} < \tau \cap C_j = 1),
\]

\[
\hat{\gamma}(\tau) = 1 / P \sum_j \mathbb{I}(M_{ij} < \tau \cap C_j = 2),
\]

where \( \mathbb{I}(\cdot) \) is the indicator function. See Figure S1 (Supplementary Materials) for an example of ROC curves for two different loci.

3.5.3 Resolution at given FP and TP rates

Consider an ensemble of \( S \) disjoint (non-overlapping) sets each with \( H \) consecutive loci. For each such set \( s = 1, \ldots, S \), we construct

\[
\hat{M}_s = \frac{1}{H} \sum_{i=1}^{H} M_{ij},
\]

where index \( j \) runs over the \( H \) loci in set \( s \). Let the average distance between loci on the investigate chromosome be \( \rho_s \). We can then assess the performance at resolution \( H \cdot \rho_s \) by using Equation (19) to estimate ROC curves for \( \{M_s\} \). Let \( \hat{\beta}^{(s)}(\tau), \hat{\gamma}^{(s)}(\tau) \) be the FP and TP rates for threshold \( \tau \). For ROC estimates of CRMA and dChip with \( H = 1, 2, \ldots \).
3, 4, see Figure 6. By constructing $\hat{H}_m$ with a mix of $H$ and $|H|$ (the greatest/least integer less/greater than $H$) loci per window together with resampling techniques, we can obtain estimates of the average performance also for non-integer values on $H$. For an example using $1 \leq H \leq 7$, see Figure 7.

### 4 RESULTS

In order for the results below to be less specific to an evaluation on ChrX alone, raw CNs are calculated using reference signals as in Equation (13) based on all samples. This will introduce a global bias in the ChrX estimates, but it has no effect on the evaluation procedure proposed in Section 3.5. The effect of using the more efficient estimate in Equation (16) is evaluated in Section 4.4.

#### 4.1 Improvements by various transformation steps

The effect of calibrating probe-pair signals for allelic crosstalk is illustrated in Figure 1. Before calibration, an allele A probe picks up signal from allele B, and vice versa. After calibration, much of this crosstalk is removed.

Note, for total CN estimates crosstalk that is perfectly symmetric between the two alleles will cancel out. This can be seen if using a symmetric crosstalk matrix in Equations (1)–(2) and then taking the sum in Equation (6). The crosstalk will not cancel out if it is asymmetric.

What is probably of greater importance is the correction for the offset in Equation (1). Offset in signals introduces intensity-dependent biases in log ratios (Bengtsson and Hössjer, 2006; Bengtsson et al., 2004). Such biases are of great concern in CN analysis, because they will add to the between-locus variability and because downstream segmentation methods assume no or equal bias for neighboring loci.

The crosstalk calibration method is applied to each array independently. Interestingly, this single-array method corrects for many of the systematic effects seen across arrays, such as differences in empirical distributions of probe signals. As shown in Figure 2, there are large differences before calibration, but afterward the distributions are approximately equal. As explained in detail in Bengtsson and Hössjer (2006) and Bengtsson (2004), observed differences in empirical distributions can often be explained by different offsets and different scale factors as modeled in Equations (1) and (2).

Note, although quantile normalization [Equations (4) and (5)] corrects for differences in distributions, it does not remove offset. It is reasonable to believe that this is the reason for seeing a better separation between one and two copies in the raw estimates when using allelic crosstalk alone compared with using quantile normalization alone, cf. left panel of Figure 3. The fact that the performance goes down when we apply quantile normalization following crosstalk calibration may be explained by the fact we are overfitting the data, at least for this particular data set. This is another reason why we suggest only using quantile normalization when large and unexpected differences in distributions remain after crosstalk calibration. Here we do not apply quantile normalization.

The existence of differences in PCR-effects across arrays affects the overall ROC performance. When correcting for PCR-fragment length effects, the TP rate increases 1% unit at a 7% FP rate (2% unit at a 4% FP rate), cf. left panel of Figure 3. We see no or even a negative effect when doing a GC-content normalization (not shown), why we suggest excluding the GC-contents function in Equation (11) unless a strong effect is observed.

#### 4.2 Performance of different methods

The distributions of TP rates $[\gamma(\cdot)]$ for the different methods, while fixing the FP rate at $\beta(\cdot) = \beta = 1/2 \cdot 1/29 = 0.0172$ as described at the end of Section 3.5.1, are depicted in Figure 4.
CRMA has a greater set of loci with high TP rates than other methods have. There is a notable increase of SNPs with very low TP rates. This may be due to some regions on ChrX being homologous to regions on chrY, but also due to misannotation or the particular model being used. For instance, the overall performance of CNAT is penalized from approximately 10% poorly performing SNPs, which we suspect is because it fits the two alleles separately adding extra variance, especially for homozygote SNPs.

From the average ROC across all ChrX loci, we get an estimate of how well on average the different methods can separate one from two copies (based on a random locus). In Figure 5, the average ROC according to Equation (19) is shown for the four methods. We see that both CRMA and dChip outperform CNAG* and CNAT by obtaining higher TP rates at any given FP rate, and that CRMA gets slightly higher TP rates than dChip for FP rates less than 7%. See also the TP rates for the single-locus resolution in Table S1.

4.3 Gain in TP rate & loss in resolution by smoothing

By averaging raw CN estimates across a set of loci as in Equation (20), we obtain more sparse but also more precise CN estimates. As a result of the greater precision, the TP rate of Rule (17) increases at any given FP rate. This is depicted in Figure 6. Note that we neither imply that Equation (20) is optimal for smoothing raw CNs nor that raw CNs should be smoothed before passing them to the segmentation method. We use it solely for comparing methods at various amounts of smoothing.

As described in Section 3.5.3, by mixing differently-sized (non-overlapping) windows, we can obtain ROC estimates for any resolution. Thus, by fixing the FP rate, we can assess the performance as a function of resolution (in kb) for each of the methods evaluated. At a FP rate of 1.0%, the TP rate of CRMA is greater than dChip, CNAG* and CNAT at any given resolution. For TP rates for the various methods at certain resolutions, see Table S1 as well as Figure 7.

Furthermore, using the above mixing approach, it is possible to find the amount of averaging required to obtain a certain TP rate at any given FP rate. For instance, based on hybridizations on the Mapping250K_Nsp chip type alone, using CRMA it is possible to obtain a 95% TP rate at a 1.0% FP rate at a 52 kb resolution. For dChip, CNAG*, and CNAT the same is obtained at 62, 78, and 121 kb resolution, respectively. See also Table S2.

4.4 Pooled diploid and non-diploid reference set

In the above evaluation, no distinction was made between female and male samples when estimating the reference signals from the pool of samples. This was done in order to not restrict the comparison to ChrX alone, but also to make the comparison between methods more fair. For ChrX loci we can obtain estimates of the diploid reference levels with greater precision by including male samples in the reference model, see Section 3.2.7. The impact of this varies with the ratio of males to females in the data set studied (here 29 females and 30 males). Thus, if using only female samples to estimate the reference level [Equation (13)], more than half of the data is discarded resulting in roughly twice the variance. This great loss in precision result in significant lower TP rates as depicted in the right panel of Figure 3. For instance, at a 5.0% FP rate, there is a decrease from 94.0% to 91.5% in TP rate when
excluding male samples. By using the proposed improvement for calculating reference levels [Equations (13)–(16)] we obtain almost the same performance as if all samples were females.

Finally, in Equations (13)–(16) the median is used for estimating the reference signals instead of the more efficient sample mean. For the data set tested, the result differ only marginally when using the sample mean (not shown). For this reason, we recommend to use the median because of its robustness against outliers and model errors.

## 5 Discussion

In the above evaluation, CNAT performs significantly worse than other methods. This may be explained by the fact that CNAT first estimates the chip effects for the two alleles separately and then sums them to obtain an estimate of the total amount of target DNA. Because the CNAT model operates on the logarithmic scale there will be a large variance in $\theta_{ijA}$ for a SNP with genotype BB, because on the intensity scale the probe signals for allele A are close to zero, and vice versa for allele B. This variance is likely to dominate the total CN estimate. See also Section 3.2.3. Furthermore, because of inflexibilities in CNAT, the reference signals had to be calculated for each allele separately, and then added. This could in addition contribute to the weaker performance.

Carvalho et al. (2007) show that by estimating chip effects for the two strands separately, the overall concordance rate for genotype calls are better. We investigated the same idea for total CN analysis by fitting one model for each strand and then averaging the two estimated chip effects (both on the intensity scale and the logarithmic scale). The result was that the ROC performance went down significantly (not shown). A reason for this may be that the allelic-crosstalk calibration removes most of the between strand effects and that when stratifying by strand more parameters are estimated based on fewer data points resulting in an increase in overall variances. If not controlling for allelic crosstalk, the systematic effect between strands are larger, which may explain why Carvalho et al. (2007) reports a gain for genotyping using an alternative approach.

Analysis of loss-of-heterozygosity (LOH) is known to provide information on copy-number losses, and can confirm or strengthen the confidence in CN regions identified by segmentation method. More generally, the genotype dimension, that is, the ratio of allele A and allele B signals, or flavors thereof, carries information on CN aberrations. In Peiffer et al. (2006), the authors use an estimate called ‘Allele Frequency’, which in our terms corresponds roughly to $\theta_{ijA}/(\theta_{ijA} + \theta_{ijB})$, in order to confirm regions of loss and gain. A similar approach was adopted in the aforementioned GTC software, where Affymetrix present the ‘Allelic Difference’ score $\log_2(\theta_{ijA}/\theta_{ijB})$. In copy-neutral regions, these ‘raw genotyping scores’ will present themselves as a mixture of three (AA, AB and BB) distributions along the genome. If there is a deletion, there will be a mixture of two (A and B) distributions, and if there is a gain, there will be a mixture of four (AAA, AAB, ABB and BBB), and so on. These scores allow us to identify different classes of gains, but also copy-neutral aberrations such as uniparental disomy. In cases of heterogeneous DNA extracts, the structure will be more complicated as there will be a mixture of mixtures.

The above is beyond the problem of estimating raw total and/or allele-specific CNs as it requires making use of the genomic locations of the SNPs. Here we have deliberately left such ideas to the various downstream methods which each may target a different question. Given optimal raw CN and raw genotyping estimates one can for instance elaborate on methods where segmentation and genotyping calls are combined in a single- or a multi-array model. Having said this, we still want to emphasize that, if the preprocessing fails to control for all systematic effects, such as crosstalk between alleles, the raw genotypes and the raw CNs will not be orthogonal. In such cases, the accuracy of the raw CN estimates can be improved further by incorporating the raw genotypes but still without utilizing genomic locations. Related to the above, we have found that more precise total CNs can be obtained if estimated directly and not via allele-specific estimates, hence Equation (6). We are working on a method for estimating the allele-specific CNs given an estimate of the total CN.

## 6 Conclusion

In this article, we have proposed a method that, compared with popular existing methods, produces single-locus estimates that better differentiate between one and two copies on ChrX. One advantage of using ChrX data for the evaluation is that it can be applied to any data set consisting of male and female samples. As we find it to be a simple and readily accessible evaluation method for raw CNs, we propose its use in any CN analysis.

The CRMA method was designed with both paired and non-paired analysis in mind as well as analysis of other than germline data, e.g., tumor data. However, the performance of CRMA in combination with downstream segmentation methods still has to be investigated, especially in cases where the underlying assumption in the preprocessing method might no longer hold, e.g., in late-stage tumors with a very large number of genomic aberrations. We look forward to comparison on other kinds of data in order to determine the extent to which the presented conclusions apply to other chromosomes and other types of CN changes.

Finally, we would like to emphasize that parts of the CRMA method are likely to apply to the new SNP & CN chip types, e.g., the calibration of allelic crosstalk in SNPs. However, we also believe that other parts need modifications to deal with data from these new chip types. Moreover, since most of the NP units on these chips are single-probe units, the current approach where a PLM is fitted robustly for each unit can no longer be used, and other means for dealing with outliers must be found. We are currently developing such methods.

## Acknowledgements

We wish to thank P. Wirapati (ISREC) for help on the simplex-fitting algorithm and suggesting the two-pass quantile normalization approach, K. Simpson (WEHI) for significant contributions to aroma.affymetrix, J. Bullard & K. Hansen (UC Berkeley) for affxparser, and B. Bolstad (Affymetrix) for
modifying **affyPLM** according to our needs. We also thank the reviewers for valuable feedback helping improve this text. H.B. was supported by grants from the Wenner-Gren Foundation, the American-Scandinavian Foundation, and the Solander Foundation.

**Conflict of Interest:** none declared.

**REFERENCES**


