ABSTRACT

Motivation: Identification of somatic DNA copy number alterations (CNAs) and significant consensus events (SCEs) in cancer genomes is a main task in discovering potential cancer-driving genes such as oncogenes and tumor suppressors. The recent development of SNP array technology has facilitated studies on copy number changes at a genome-wide scale with high resolution. However, existing copy number analysis methods are oblivious to normal cell contamination and cannot distinguish between contributions of cancerous and normal cells to the measured copy number signals. This contamination could significantly confound downstream analysis of CNAs and affect the power to detect SCEs in clinical samples.

Results: We report here a statistically principled in silico approach, Bayesian Analysis of COpy number Mixtures (BACOM), to accurately estimate genomic deletion type and normal tissue contamination, and accordingly recover the true copy number profile in cancer cells. We tested the proposed method on two simulated datasets, two prostate cancer datasets and The Cancer Genome Atlas high-grade ovarian dataset, and obtained very promising results supported by the ground truth and biological plausibility. Moreover, based on a large number of comparative simulation studies, the proposed method gives significantly improved power to detect SCEs after in silico correction of normal tissue contamination. We develop a cross-platform open-source Java application that implements the whole pipeline of copy number analysis of heterogeneous cancer tissues including relevant processing steps. We also provide an R interface, bacomR, for running BACOM within the R environment, making it straightforward to include in existing data pipelines.

Availability: The cross-platform, stand-alone Java application, BACOM, the R interface, bacomR, all source code and the simulation data used in this article are freely available at authors’ web site: http://www.cbil.ece.vt.edu/software.htm.

Supplementary Information: Supplementary data are available at Bioinformatics online.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
The BACOM software instantiates the algorithms described in this
We first discuss a deletion-focused latent variable model for the copy
and subsequently recover the cancer-specific DNA copy number profile.
approach to statistically characterize distinctive copy number signals due
(i) inference of deletion types, (ii) estimation of the normal tissue fraction
and (iii) recovery of the copy number profile in cancer cells.
2.1 Copy number signal model
Supplementary Figure S17 shows SNP array intensity signals that serve as
of DNA copy number signals from both normal and cancer cells, given
mathematically by
\[ X_i = \alpha \times X_{\text{normal}}, + (1 - \alpha) \times X_{\text{cancer}}. \]
where \( X_i \) is the observed DNA copy number signal at locus \( i \), \( \alpha \) is the unknown fraction of normal cell subpopulation in the sample and \( X_{\text{normal}}, \) and \( X_{\text{cancer}} \) are the unknown latent DNA copy number signals in normal and cancer cells at locus \( i \), respectively. It should be noted that, in model (1), we have chosen not to consider CNVs in normal cells, because these are much rarer than CNAs in cancer cells.

\[ X_i = \alpha \times X_{\text{normal}}, + (1 - \alpha) \times X_{\text{cancer}}. \]
Since human somatic cells are diploid, the expected DNA copy number at locus \(i\) in normal cells is two, i.e. \(E[X_{normal,i}]=2\). In contrast, if there is a homo-deletion or semi-deletion at locus \(i\) in cancer cells, then the expected DNA copy number becomes zero or one, i.e. \(E[X_{cancer,i}]=0\) or 1.

By focusing on deletion-only CNA loci and taking the expectations on both sides of Equation (1), we have

\[
E[X_i] = \begin{cases} 
2 + (1-\alpha) \times 0 & \text{if homo-deletion,} \\
2 + (1-\alpha) \times 1 & \text{if semi-deletion.} 
\end{cases}
\] (2)

Equation (2) indicates that, as a function of normal cell fraction \(\alpha\), the expected copy number at a deletion locus depends on the deletion type and is distinctive except when \(\alpha=1\). Inspired by this observation, we propose to explore a statistically principled solution (detectability): if a Bayesian hypothesis test can be constructed to differentiate between homo-deletion and semi-deletion segments based on allele-specific signals, we could, in principle, estimate \(\alpha\) by the sample average over the deletion segments.

2.2 Inference of deletion type

Affymetrix SNP chips provide both allele-specific signals (A allele and B allele) and their summed intensity (observed DNA copy number signal). If we denote the signals of alleles A and B at locus \(i\) by \(X_{A,i}\) and \(X_{B,i}\), respectively, then the observed DNA copy number signal \(X_i\) in model (1) can be rewritten as

\[X_i = X_{A,i} + X_{B,i} .\] (3)

To fully exploit allele-specific information readily provided by the SNP arrays and associated genotype calling algorithms, our method will focus solely on AB genotype (not considering AA or BB genotypes). For a length-\(L\) homo/semi-deletion segment \([X_i = 1, 2, \ldots, L]\), we make the following realistic assumption on the allele-specific signals.

**Assumption 1.** For a length-\(L\) homo/semi-deletion segment \([X_i = 1, 2, \ldots, L]\), each of the allele-specific signals \(X_{A,i}\) and \(X_{B,i}\) are independently distributed Gaussian random variables with distinct means but common variance, \(\sigma^2\), for \(i = 1, 2, \ldots, L\).

It should be noted that \(X_{A,i}\) and \(X_{B,i}\) are not statistically independent but, rather often correlated, referred to as the cross-talk between alleles A and B (Bengtsson et al., 2008). Thus, under Assumption 1, the observed copy number signals \(X_i\) are independent and identically distributed random variables following a normal distribution \(N(\mu_{X_{A,i}}, \sigma^2 X_{A,i})\) whose mean \(\mu_{X_{A,i}}\) and variance \(\sigma^2_{X_{A,i}}\) can be readily estimated by using the observed signals \(X_i\) for \(i = 1, 2, \ldots, L\).

To statistically differentiate between hemi-deletion and homo-deletion, we define a novel summary statistic, given mathematically by the following newly defined random variable

\[Y = \frac{1}{\sigma^2_{X_{A,i}}} \sum_{i=1}^{L} (X_{A,i} - X_{B,i})^2,\] (4)

where \(\sigma^2_{X_{A,i}}\) is the variance of \(X_{A,i} - X_{B,i}\). Under Assumption 1, it can be shown that \(Y\) follows either a non-central or a standard \(\chi^2\) distribution, depending upon the deletion type. We, therefore, present the following two lemmas with proofs to show that the key parameter associated with these \(\chi^2\) distributions can be estimated using signals \(X_{A,i}\) and \(X_{B,i}\).

**Lemma 1.** Suppose that, within a length-\(L\) semi-deletion segment, each of the allele-specific signals \(X_{A,i}\) and \(X_{B,i}\) are independently distributed Gaussian random variables with distinct means and common variance. Then, the summary statistic random variable \(Y\) defined in (4) follows an \(L\) degrees of freedom non-central \(\chi^2\) distribution with non-centrality parameter \(\Lambda = L(2 - \mu_{X_{A,i}})^2 \sigma^2_{X_{A,i}} / \sigma^2_{X_{B,i}}(1+\rho)/(1-\rho)\), where \(\rho\) is the correlation coefficient between \(X_{A,i}\) and \(X_{B,i}\).

**Proof.** Applying Equation (1) to the loci within a semi-deletion segment, where one of the alleles (but not both) is deleted, we have, for \(i = 1, 2, \ldots, L\)

\[
\mu_{X_{A,i}} = E[X_{A,i} - X_{B,i}] = E[\alpha (X_{normal,i} - X_{normal,i})]
\]

\[+ (1-\alpha) (X_{cancer,i} - X_{cancer,i}) = \alpha \times E[X_{normal,i} - X_{normal,i}]\]

\[+ (1-\alpha) \times E[X_{cancer,i} - X_{cancer,i}] = \alpha \times (1-\alpha) \times (X_i - 0) = \alpha (1-\alpha), \quad i = 1, 2, \ldots, L\]

While from Equation (2), we have \(\mu_{X_{A,i}} = E[X_i] = 1+\alpha\) which implies \(\alpha = \mu_{X_{A,i}} - 1\).

Thus, \(\mu_{X_{A,i}}\) can be expressed in terms of \(\mu_{X_{A,i}}\) as

\[
\mu_{X_{A,i}} = \alpha (1-\alpha) = \alpha (1-\mu_{X_{A,i}}) = \alpha (2-\mu_{X_{A,i}}).
\]

Furthermore, Assumption 1 implies that

\[
\sigma^2_{X_{A,i}} = 2\sigma^2(1+\rho) \quad \text{and} \quad \sigma^2_{X_{B,i}} = 2\sigma^2(1-\rho).
\]

Although direct estimation of \(\sigma^2_{X_{A,i}}\) is a non-trivial task, simple mathematical manipulation shows that \(\sigma^2_{X_{A,i}}\) can be expressed in terms of \(\sigma^2_{X_{B,i}}\) as

\[
\sigma^2_{X_{A,i}} = \sigma^2_{X_{B,i}} (1-\rho)/(1+\rho).
\]

By the definition of the non-centrality parameter \(\lambda\) and Equation (4), we conclude

\[\lambda = \frac{\sigma^2_{X_{A,i}}}{\sigma^2_{X_{B,i}}} \sum_{i=1}^{L} \frac{(X_{A,i} - X_{B,i})^2}{\sigma^2_{X_{A,i}}(1+\rho)}\]

\[= L(2 - \mu_{X_{A,i}})^2 \sigma^2_{X_{A,i}} / \sigma^2_{X_{B,i}} (1+\rho)/(1-\rho)\].

Accordingly, the conditional \(L\) degrees of freedom non-central \(\chi^2\) distribution of \(Y\) under semi-deletion is given by

\[\chi^2 (L, \lambda) \sim \frac{\chi^2(\lambda+2L-2)}{\chi^2(\lambda+2L-2)} \sum_{i=1}^{L} \frac{X_{A,i}^2 - X_{B,i}^2}{\chi^2(\lambda+2L-2)} \]

for \(y > 0\),

\[0 \quad \text{for} \ y \leq 0\] (5)

where \(\Gamma()\) denotes the Gamma function.

Q.E.D.

**Lemma 2.** Suppose that, within a length-\(L\) homo-deletion segment, each of the allele-specific signals \(X_{A,i}\) and \(X_{B,i}\) are independently distributed Gaussian random variables with distinct means and common variance. Then, the summary statistic random variable \(Y\) defined in (4) follows an \(L\) degrees of freedom standard \(\chi^2\) distribution.

**Proof.** Applying Equation (1) to the loci within a homo-deletion segment, where both alleles are deleted, we have, for \(i = 1, 2, \ldots, L\)

\[
\mu_{X_{A,i}} = E[X_{A,i} - X_{B,i}] = E[\alpha (X_{normal,i} - X_{normal,i})]
\]

\[+ (1-\alpha) (X_{cancer,i} - X_{cancer,i}) = \alpha \times E[X_{normal,i} - X_{normal,i}]\]

\[+ (1-\alpha) \times E[X_{normal,i} - X_{normal,i}] = \alpha \times (1-\alpha) \times (0-0) = 0.
\]
Theorem 1: (deletion-type identifiability). Suppose that, within a length-L deletion segment, each of the allele-specific signals $X_{A,i}$ and $X_{B,i}$ are independently distributed Gaussian random variables with distinct means and common variance. Then, the summary statistic random variable $Y = \sum_{i=1}^{N} (X_{A,i} - X_{B,i})^2$ follows an $\chi^2$ distribution under homo-deletion, and a non-central $\chi^2$ distribution under hemi-deletion, with a parameter that can be estimated based on available signals $X_{A}, X_{A,i}, X_{B}$, and $X_{B,i}$. Furthermore, having determined the deletion-type conditional probability density functions, we can then identify the deletion type of the segment using Bayesian hypothesis testing. The conclusion is summarized in the following theorem.

Proof: From Lemma 1, the summary statistic random variable $Y$ under hemi-deletion follows an $\chi^2$ distribution with $L$ degrees of freedom non-central distribution.

Having determined the distribution of the deletion-type conditional models, we can infer the type of each deletion segment by applying Bayesian hypothesis testing based on (7). Subsequently, we can estimate the fraction of normal cells in the sample specified by (2), i.e., $u_j = \frac{\mu_j}{\mu_j + 2}$ for homo-deletion and $u_j = \frac{\mu_j}{2}$ for homo-deletion, where $\mu_j$ is the sample average of the copy number signals of the $j$-th deletion segment. Moreover, assume that there are $K$ deletion segments, we can calculate the ensemble estimate of the normal cell proportion via segment-length weighted average.

\[
p = \frac{\sum_{j=1}^{K} u_j \cdot L_j}{\sum_{j=1}^{K} L_j},
\]

where $L_j$ is the length of the $j$-th deletion segment.

Finally, the estimated normal cell fraction can be used to recover the true copy numbers in cancer cells in the sample. Since $X_{normal} = 2$ and based on (1), it is straightforward to estimate the DNA copy number of pure cancer cells by

\[
\hat{X}_{\text{cancer}} = X_{\text{cancer}} - 2p
\]

3 BACOM SOFTWARE

3.1 Stand-alone Java application

To better serve the research community, we developed a cross-platform and open-source BACOM Java application, which implements the entire pipeline of copy number change analysis for heterogeneous cancer tissues (Supplementary Material). The BACOM software instantiates not only the novel algorithms described here but also other relevant processing steps, including extraction of raw copy number signals from CEL files, iterative data normalization, identification of AB loci, copy number detection and segmentation, probe sets annotation, differentiation of deletion types, estimation of the normal tissue fraction and correction of normal tissue contamination. Interested readers can freely download the software and source code at http://www.cbii.ece.vt.edu/software.htm.

3.2 Running BACOM in R environment

To take advantage of many widely used packages in R and its associated powerful and versatile visualization capabilities, we also implemented an R interface, hacomR, that enables users to smoothly incorporate BACOM into their routine copy number analysis pipeline or integrate BACOM with other R or Bioconductor packages. Users can use their preferred methods to perform routine tasks such as array normalization and DNA copy number segmentation and estimation, while using the newly added BACOM to estimate the normal cell fraction and subsequently recover the true copy number profiles in pure cancer cells.

4 RESULTS

4.1 Simulation studies

We first consider a realistic synthetic dataset from a mixture of normal and simulated cancer copy number profiles, as shown in

\[\mu_A = \sum_{i=1}^{N_{normal}} X_{A,i}, \quad \mu_B = \sum_{i=1}^{N_{cancer}} X_{B,i}\]

\[\rho = \frac{\sum_{i=1}^{N_{normal}} (X_{A,i} - \mu_A)(X_{B,i} - \mu_B)}{\sqrt{\sum_{i=1}^{N_{normal}} (X_{A,i} - \mu_A)^2 \sum_{i=1}^{N_{normal}} (X_{B,i} - \mu_B)^2}}\]
Figure 2. The DNA copy number profile and Bayesian analysis of the deletion segment of the simulation dataset 1 when $\alpha = 0.7$. (A) Copy number profile and correction results. (B) Bayesian analysis to determine the deletion type of the deletion segment.

Figure 2a. The cancer copy number profile is simulated based on the real DNA copy number profile of a normal tissue sample assayed on the Affymetrix Genome-Wide 6.0 SNP array, consisting of two simulated four-copy amplification segments and one simulated hemi-deletion segment. The normal and cancer copy number profiles are numerically mixed based on known proportions to produce the observed copy number signal. Since there is only one deletion segment (loci 25k–30k), it is theoretically impossible to tell the deletion type by examining the observed copy number signal, given the fact that the cancer copy number signal has been severely contaminated by a normal copy number signal. The single deletion inclusion in this dataset has been chosen in order to illustrate the unsupervised learning ability of BACOM in determining deletion types.

To determine the deletion type, we first estimate the posterior probability models of the summary statistic using allele-specific signals provided by SNP chips, and plot the observed value of the summary statistic associated with the deletion segment, shown in Figure 2b. The plot clearly suggests the hemi-deletion type of the deletion segment. We then estimate the normal tissue fraction in the sample based on the sample average of the deletion segment $\alpha = H \exp \{1 \}$. This leads to an estimate of $\alpha = 0.692$ and the accordingly corrected cancer copy number profile shown in Figure 2a. The results show the effectiveness of the BACOM approach in that the deletion type is correctly determined, the estimated normal tissue fraction is very close to the true value $\alpha = 0.7$ and the recovered amplification signals indicate the two expected four-copy segments.

As an example of a more complex simulation, we consider a dataset from a mixture of normal and simulated cancer copy number profiles, as shown in Figure 3a. The cancer copy number profile includes one homo-deletion, two hemi-deletions and three different amplification (copy numbers 3, 4 and 5) segments. The simulated cancer copy number signal, with a total of six altered copy number segments, not only retains the statistical characteristics of real SNP array intensity data, but also provides a more complete picture of copy number alterations and genomic instability in cancer cells. Once again, the normal and cancer copy number profiles are numerically mixed based on known proportions to produce the observed copy number signal. The multiple type-deletion inclusions in this dataset have been chosen in order to illustrate the consistency and applicability of BACOM in estimating normal tissue fraction and cancer-associated copy number alterations.

We first estimate separately the individual normal tissue fractions $\alpha_j$ from one homo-deletion and two hemi-deletion segments, where the posterior probability models and observed values of the summary statistic associated with the deletion segments are shown in Figure 3b. We then use the average value $\bar{\alpha}$ to recover the cancer-associated copy number profile, shown in Figure 3a, where the solid line segments are the recovered cancer-associated copy number changes. We tested BACOM on six simulation datasets with different $\alpha$ values, as given in Table 1. The BACOM approach again achieved very promising results in which the deletion types are correctly determined, the estimated normal tissue fractions from different deletion segments are highly consistent, with the average value very close to the true value and the recovered signals of all six deletion and amplification segments indicate the expected integer-valued copy number changes. Table 1 summarizes the experimental results from all 12 simulated copy number profiles.

4.2 Analysis of real DNA copy number data

To test the applicability of our proposed method, we consider a real copy number profile for a prostate cancer sample assayed on the Affymetrix SNP 500K array. We first applied the BACOM algorithm to estimate the fraction of normal cell population in the sample, resulting in $\bar{\alpha} = 0.784$, which indicates significant normal tissue contamination. We then used the estimated $\bar{\alpha}$ value to recover cancer-specific copy number signal by Equation (13). The resulting corrected copy number profile for Chromosome 10 is shown in Figure 4, where dotted signals are the mixed copy number signals arising from the tumor sample with blue-colored regions being the detected deletion segments, green solid lines are the normal copy number segments and blue solid lines are the corrected cancer-specific deletion segments. In this experiment, our analysis readily reveals and distinguishes both deletion types and their occurred genomic locations. It is worth noting that BACOM algorithm identified a homo-deletion segment around locus 18 500 in Chromosome 10, that contains the well-known tumor suppressor gene PTEN.
Fig. 3. The DNA copy number profile and Bayesian analysis of deletion segments of the simulation dataset 2 when $\alpha = 0.7$. (A) Copy number profile and correction results. (B) Bayesian analysis to determine the deletion types of the three deletion segments.

Table 1. Estimation results on two simulation datasets

<table>
<thead>
<tr>
<th></th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\alpha_1$</td>
<td>$\alpha_1$</td>
</tr>
<tr>
<td>0.3</td>
<td>0.291</td>
<td>0.293</td>
</tr>
<tr>
<td>0.4</td>
<td>0.391</td>
<td>0.393</td>
</tr>
<tr>
<td>0.5</td>
<td>0.491</td>
<td>0.493</td>
</tr>
<tr>
<td>0.6</td>
<td>0.591</td>
<td>0.593</td>
</tr>
<tr>
<td>0.7</td>
<td>0.693</td>
<td>0.693</td>
</tr>
<tr>
<td>0.8</td>
<td>0.793</td>
<td>0.793</td>
</tr>
</tbody>
</table>

As an example of a somewhat independent verification, we applied the BACOM algorithm to the copy number profile of another prostate cancer sample assayed on the Affymetrix Genome-Wide 6.0 platform (Liu et al., 2009). The estimated fraction of normal cells in the sample is $\alpha = 0.691$ and the results of similar analyses are given in Figure 5. Different from the previous example, this copy number profile contains two amplification segments that are purple colored. Denoted by red solid lines, the corrected copy numbers of amplification segments are integer valued, consistent with our theoretical expectation. This observation serves as a convincing validation of the proposed method, since the normal cell fraction $\alpha$ was independently estimated from only deletion segments.

To further assess the applicability of BACOM, we report results on additional 55 prostate samples and The Cancer Genome Atlas high-grade ovarian cancer dataset, showing a considerable range of variation in normal cell contamination (36–70%) and co-deletion of two tumor suppressor genes—TP53 and BRCA1 (Supplementary Table S2 and Figs S3–S10).

4.3 Impact on detecting significant consensus events

Somatic copy number alterations in genomes underlie almost all human cancers. One of the systematic efforts to characterize cancer genomes is to identify significant consensus events (SCGEs) from random background aberrations. To test the utility of our method to address an important biological question, we applied BACOM together with genomic identification of significant targets in cancer (GISTIC) (Beroukhim et al., 2007) to specifically designed copy...
To analyze the impact of correcting normal tissue contamination on false discovery rates (FDRs) and (Figure 6, consistently show significantly improved power using and ground truth. Comparative experimental results, given in detecting SCEs, we calculated power based on GISTIC outcomes position, using a semi-exact approach to determine the significance.

cancer-specific copy numbers by BACOM. To detect SCEs from normal copy number and various deletion/amplification segments (150–250 loci). The consensus events at certain loci are inserted into the base profile according to a specified frequency, while random background aberrations are simulated with randomly assigned length and loci. Simulation parameters include sample size, consensus frequency and normal cell fraction. We generated 1000 simulation datasets for each combinatorial parameter setting, resulting in a total of 20,000 simulation datasets, each containing 30–90 samples.

Next, for each of the mixed copy number profiles, we recovered cancer-specific copy numbers by BACOM. To detect SCEs from both mixed and deconvolved copy number profiles, we applied GISTIC, a statistical method that calculates a score that is based on both the amplitude and frequency of copy number changes at each position. Using a semi-exact approach to determine the significance. To analyze the impact of correcting normal tissue contamination on detecting SCEs, we calculated power based on GISTIC outcomes and ground truth. Comparative experimental results, given in Figure 6, consistently show significantly improved power using deconvolved cancer-specific profiles (see more discussion and results in the Supplementary Material).

5 CONCLUSIONS AND DISCUSSIONS

In this article, we report a statistically principled in silico approach to estimate copy number deletion types and normal tissue contamination, and to extract the true copy number profile in cancer cells. The BACOM algorithm utilizes the allele-specific information provided by SNP chips to differentiate between hemi-deletion and homo-deletion and subsequently estimates the fraction of normal cells in tissues. We tested the proposed method on two simulated datasets, two prostate cancer datasets and The Cancer Genome Atlas high-grade ovarian dataset, and obtained highly promising results. We expect the newly developed BACOM software to be a useful tool in copy number analysis of heterogeneous tissues. There are some questions worth further exploration. Specifically, so far we have focused on normal tissue contamination by assuming a homogeneous cancer cell population, while in reality, cancer cells are often clonally heterogeneous leading to cancer subtypes. The ability to further dissect genomic heterogeneity of cancer cells is of great interest and will facilitate pathogenesis studies with far-reaching clinical implications.

In addition to heterogeneity of copy number, more mutations in cancer cells are expected and may have some unknown implications. However, since the summary statistic was defined on the whole deletion segment and the final normal tissue fraction was estimated using segment-length weighted average over multiple deletion segments, such mutations will only have negligible effects on the estimation accuracy as long as the mutations are sporadic compared with copy number alterations. In our experiments on real datasets, we have not observed any major effects caused by such mutations.

In relation to earlier work, the use of allele-specific signals is shared by Goransson et al. for developing a histogram matching-based CNNLOH method (Goransson et al., 2009). Besides high computational complexity in histogram simulation and matching, the mechanism for evaluating observed histogram is prone to high estimation bias and variance as CNNLOH uses truncated, evenly weighted and coarse histograms in winner-take-all selection among chromosomes. We have conducted a comprehensive simulation study to compare the performance of CNNLOH and BACOM. Experimental results consistently show that BACOM outperforms CNNLOH in terms of much lower estimation bias and variance, as well as dramatically reduced computation time (see more discussions and results in Supplementary Material).

Funding: US National Institutes of Health under Grants (EB000830, CA109872, CA129080 and CA103937) in part.

Conflict of Interest: none declared.

REFERENCES


