Multi-platform segmentation for joint detection of copy number variants

Shu Mei Teo\textsuperscript{1,2,3}, Yudi Pawitan\textsuperscript{2}, Vikrant Kumar\textsuperscript{4}, Anbupalam Thalamuthu\textsuperscript{4}, Mark Seielstad\textsuperscript{4}, Kee Seng Chia\textsuperscript{1} and Agus Salim\textsuperscript{1,\ast}

\textsuperscript{1}Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, National University of Singapore, Singapore, \textsuperscript{2}Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden \textsuperscript{3}NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore and \textsuperscript{4}Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore

ABSTRACT

Motivation: With the expansion of whole-genome studies, there is rapid evolution of genotyping platforms. This leads to practical issues such as upgrading of genotyping equipment which often results in research groups having data from different platforms for the same samples. While having more data can potentially yield more accurate copy-number estimates, combining such data is not straightforward as different platforms show different degrees of attenuation of the true copy-number or different noise characteristics and marker panels. Currently, there is still a relative lack of procedures for combining information from different platforms.

Results: We develop a method, called MPSS, based on a correlated random-effect model for the unobserved patterns and extend the robust smooth segmentation approach to the multiple-platform scenario. We also propose an objective criterion for discrete segmentation required for downstream analyses. For each identified segment, the software reports a $P$-value to indicate the likelihood of the segment being a true CNV. From the analyses of real and simulated data, we show that MPSS has better operating characteristics when compared to single-platform methods, and have substantially higher sensitivity compared to an existing multiplatform method.

Availability: The methods are implemented in an R package MPSS, and the source is available from http://www.meb.ki.se/$\sim$yudpaw.

Contact: agus_salim@nus.edu.sg

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on December 17, 2010; revised on March 8, 2011; accepted on March 26, 2011

1 INTRODUCTION

Copy-number variants (CNVs) are defined as duplications or deletions in the number of copies of a DNA segment (larger than 1 kb in length) when compared to a reference genome. Currently, common technologies used to detect CNVs include high-density single nucleotide polymorphism (SNP) arrays and comparative-genomic hybridization (CGH) arrays. In recent years, whole-genome studies using commercial genotyping arrays to detect CNVs have been rapidly expanding. With decreasing cost of commercially available platforms and the fast evolution of these platforms, it is not unusual for research groups to have data from multiple platforms for each sample. For example, The Cancer Genome Atlas Research Network, a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to explore genomic changes involved in human cancers, used Agilent 244K, Affymetrix SNP 6.0 and Illumina 550K platforms to measure copy number alterations in its pilot study. Our own collaborators, and perhaps many other researchers, collected genotype data using both Illumina HumanHap300 and HumanHap240S arrays for each sample in order to get higher genome coverage.

Marker density is an important factor for comprehensive and accurate detection of CNVs and their breakpoints, and different platforms have different probe coverage and density; see Curtis et al. (2009) for a summary of probe coverage of the different platforms. Combining data from different platforms can potentially yield more precise and accurate detection of CNVs and its breakpoints. However, combining such data is not straightforward because it is known that estimates from different platforms show different degrees of attenuation of the true copy-number changes (Bengtsson et al., 2009) as well as different noise characteristics. Furthermore, different platforms have different marker panels and molecular assay methods (Zhang et al., 2010).

Currently, there is still a relative lack of formal procedures for combining information from different platforms for copy-number calling. Most studies with multiple platforms interrogating the same samples process the data from the different platforms independently, then combine the segments in an ad hoc manner. This approach does not fully utilize information from the different platforms, and when the segmented results from the different platforms differ, it is difficult for researchers to come to a consensus in a statistically rigorous manner.

One published method, multiple platform circular binary segmentation (MPCBS) (Zhang et al., 2010), is able to jointly use information from different platforms for CNV calling. The MPCBS method extends the circular binary segmentation (CBS) algorithm (Olshen et al., 2004) by detecting coupled changes in multiple sequences. Briefly, it uses a weighted sum of $t$-statistics
We assume the errors and the random-effects to be independent. The error et al. the likelihood of the Cauchy random-effects model (see Huang et al., 2007, 2009). The Cauchy distribution on location 0 and scale factor \( \sigma_j \):

\[
log(f(\mu|\sigma_j)) = -(\mu - 2 \log(\pi \sigma_j) - \sum_{i=1}^{n_j} \log(1 + \sigma_j^{-2}))
\]

Differentiating (3) with respect to \( f \), we get:

\[
\frac{d \log(f(\mu|\sigma_j))}{d\mu} = \frac{2 \sigma_j^{-2}}{1 + \sigma_j^{-2}}
\]

Combining (5) and (6), we obtain the score function, the first derivative of \( \log(\hat{L}(\sigma_j^2); \sigma_j^2) \) as:

\[
S(\mu) = (Z'WZ)^{-1} - (\Delta^2 D'^{-1} \Delta^2)'y'Z'WZ y
\]

We estimate \( f \) from (7) by exploiting the band-limited property of \( Z'WZ + (\Delta^2 D'^{-1} \Delta^2) \) and use well-tested fortran subroutines available in Linpack (see Huang et al., 2007 and Dongarra et al., 1979).

2.2 Estimation of \( \sigma_j \)

Given \( \hat{f} \), at each probe position \( i \), the deviance is defined as

\[
d_i = (k + 1) \log \left( 1 + \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}_i^2} \right)
\]

This can be approximated by the gamma distribution with mean \( \mu_i \) and dispersion \( \phi \). To estimate \( \mu_i \), we use a generalized linear model with a log-link function, so \( b_i = \log(\mu_i) \) and \( \hat{b}_i = \sum c_j \), where the dimension of \( s_i \) and \( \alpha \) is equal to the number of platforms. We solve using IWLS with robust weights:

(1) Start with an initial \( a_{0i} \). We estimate \( \phi \) once using \( \hat{\phi} = \frac{(w_j - w_{0j})}{\alpha} \).

(2) We write

\[
Y^* = X\hat{\mu} + \epsilon^*
\]

where \( Y^* \) is the working vector with elements

\[
y^*_i = \frac{\partial}{\partial \hat{\mu}_i} (d_i - \hat{\mu}_i + \hat{\epsilon}_i)
\]

(3) We use robust weights

\[
w_j = \frac{1}{\text{var}(\epsilon_j^2)} \times \text{whaheur}
\]

where \( w_{\text{heur}} \) is the commonly used Huber weight function defined as

\[
\text{whaheur}(\epsilon) = \begin{cases} 
1 & \text{if } |\epsilon| < c_j \\
\frac{\epsilon}{|\epsilon|} & \text{if } |\epsilon| > c_j
\end{cases}
\]

where \( c_j = 1.345 \). As an initial estimate, we use a robust measure of spread, \( d_i = \text{median}(|\epsilon_i|) / 0.745 \). Then \( \alpha \) can be solved using the usual weighted least squares solution:

\[
\hat{\alpha} = (X'WX)^{-1}X'Wy^*
\]
The Fisher information for degrees of freedom associated with the elements of $\sigma$. To compute (12) without explicitly obtaining the inverse of a matrix, we use an approximation (Pawitan, 1996) for $\lambda$ that minimizes the Akaike information criterion (AIC).

2.3 Choosing optimal $\lambda$

The degrees of freedom associated with $\lambda$ is given by (Pawitan, 2001, p. 448)

$$\Delta f = \text{trace}(Z'WZ + (\Delta^2 D^{-1}(\Delta^2)^{-1})Z'WZ)$$

where $W$ and $D$ are computed using $\hat{f}$. This expression is hard to obtain computationally, so we use an approximation (Pawitan, 1996)

$$\Delta f = \sum_{k=1}^{n-2} \frac{\hat{w}_k}{\hat{d}_k}$$

where $\hat{w}$ and $\hat{d}$ are the average diagonals of $Z'WZ$ and $D$, and

$$\hat{W} = [1 - \cos(\pi(k-1)/n)]$$

is the $j$th eigenvalue of the second derivative matrix $\Delta^2$. We choose $\lambda$ that minimizes the AIC.

2.4 Summary of MPSS algorithm

For a given $\lambda = \frac{\text{average}}{\text{std deviation}}$, we employ the following algorithm:

(a) Start with an initial value for $f_0$ and $\sigma_f^2$. To adjust for multiple hypothesis testing involving a large number of hypotheses, users can pick the threshold that maximizes the total log-likelihood.

(b) Compute $\sigma_f^2 = \text{average}(\text{std deviation})$.

(c) Compute $Z'WZ$ and $D^{-1}$ and update $f$ using (7).

(d) Update $\sigma_f^2$ as described in Section 2.2.

(e) Repeat (b)–(d) until convergence.

2.5 $P$-values for segments

The Fisher information for $f$ is the negative of the second derivative of the log-likelihood.

$$f_0 = Z'WZ + (\Delta^2 D^{-1}(\Delta^2)^{-1})$$

At convergence, the estimated variance matrix is

$$V = \hat{f}^{-1}(f)$$

If we have a segment $S$, defined for instance by setting a threshold, then $f_S$ is a vector that contains the estimated values in the segment and zero everywhere else:

$$f_S = \begin{cases} \hat{f} & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases}$$

The significance of the segment can be assessed using the statistic

$$\chi^2 = f_S' V^{-1} f_S$$

(12)

To compute (12) without explicitly obtaining the inverse of a matrix with extremely large dimension, we write (12) as

$$f_S' Z'WZ + (\Delta^2 D^{-1}(\Delta^2)) f_S = \sum(f_S' w_i) + \sum(d_i d)$$

where $f_S$ contains the second-order differences of $f_S$ and $d$ contains the diagonal elements of $D^{-1}$. We compare this statistic to the chi-squared distribution with $q$ degrees of freedom, where $q$ is the number of probes in $S$. To adjust for multiple hypothesis testing involving a large number of segments, we compute the false discovery rate (FDR) for each segment.

2.6 Objective threshold segmentation

A segment whose random-effects parameter $f$ consistently and significantly deviates from zero is evidence of a deletion/duplication. We obtain potential copy-number segments by setting thresholds for $f$, where duplicates are sets of consecutive probes for which $f$ is consistently greater than or equal to a specified threshold, and deletions are sets of consecutive probes for which $f$ is consistently smaller than or equal to a specified threshold. For automatic threshold selection, users can pick the threshold that maximizes the total $\chi^2$ values (scaled by its associated degrees of freedom).

To avoid oversegmentation, we merge the segments if the distance between adjacent segments is less than 5 kb. For each segment, we compute its associated $P$-value/FDR as described in the previous section. We further filter the segments by their length (those that are less than 1 kb are omitted). We remove the segments where the FDR for the test is $<0.01$. We call these segments ‘discrepant segments’. Discrepant segments are removed because they are likely to be unreliable.

2.8 Comparisons using simulated data

We conduct a simulation study to evaluate the performance of MPSS as well as to compare against the MPCBS method. To get a realistic noise pattern, we use the empirical CNV profile of chromosome 1 of the Hapmap sample NA10851. We use data from both Affymetrix 6.0 and Illumina 1M platforms (see Section 3.1) and apply the MPSS algorithm with segmentation threshold of 0.05 and FDR threshold of $10^{-5}$. We remove segments with $<4$ probes as extremely short segments are more likely to be false positives due to noise. We label the different segments of the chromosome as CNV or ‘NULL’. In total, there are 12 CNV segments and 15 ‘NULL’ segments. We perform the simulation study at three different noise levels; the input values are the smoothed intensities plus 1 times the residuals.

2.9 Comparisons using real data

We compare MPSS against the single-platform smoothing as well as with the simulations in a real data setting. We use the integer copy-numbers for a total of 5037 CNV loci from Conrad et al. (2010)’s study as well as McCarron et al. (2008)’s study as a reference list. A set of 20 NimbleGen arrays, each comprising 2.1-million long oligonucleotide probes were used to first generate a new map of CNV locations. Subsequently, a customized Agilent CGH-platform
comprising of 105,000 long oligonucleotide probes was used to detect the loci and the genotypes were estimated for 450 HapMap samples using a Bayesian algorithm with stringent selection for optimal normalization and cluster locations for every locus [see Supplementary Material in Conrad et al. (2010) for more details]. We remove segments in the reference if the number of probes from the combined probe list from the two platforms we are using is less than 10 or if the segment size is less than 1kb. There is a median of 163 CNV segments per individual.

It should be noted, however, that this reference list cannot really be considered the gold standard, as even sequencing data do not have 100% sensitivity and specificity in CNV detection (Xue et al., 2009). For each method, we perform individual-specific comparisons with Conrad's CNVs and compute the number of bases that are called as CNV both by the method and by Conrad et al. We report the number of overlapping bases as a proportion to the total length of CNVs identified by the method and as a proportion of total length of Conrad's CNVs. While these may not be considered a 'true discovery rate' and 'sensitivity', since Conrad's CNVs are well-validated, a higher proportion of overlap is an indication of better performance.

2.10 Implementation and computing time
The methods are implemented in an R package MPSS. The main inputs are vectors of genomic positions, chromosome numbers and log2-intensity ratios from each platform. It is recommended that users check if data from the different platforms are well-normalized. If not, background correction should be performed first; the package rsmooth from http://www.meb.ki.se/~yudpaw can be used for background correction. All computations for this article were done on a 3GHz Intel Core 2 Duo processor. For 1 individual, with more than 2.5 million combined probes from Affymetrix 6.0 and Illumina 1M, and for a user-specified λ, the algorithm takes <1 min. It takes <6 min if the AIC criteria is used to find the optimal λ.

3 RESULTS
3.1 Datasets and background correction
We use nine HapMap samples (International HapMap Consortium, 2005 and see Supplementary Materials for sample ID and population). These samples were previously genotyped by two SNP arrays (Illumina 1M and Affymetrix 6.0) in our research lab. We first perform background correction on the log2-intensity ratios from each platform using a robust smoother in the rsmooth package from http://www.meb.ki.se/~yudpaw. This normalization assumes that the majority of the genome does not contain CNVs, which is the case for germline samples.

To investigate if the input intensities are well-normalized, we randomly sample a non-CNV segment of 100 consecutive probes and test if the mean intensity for the Affymetrix platform is equal to the mean intensity of the Illumina platform (using t-test corrected for autocorrelation). We repeat the process 1000 times and record the P-values that are less than 0.01. The normalization results look reasonable for all individuals with the percentage of P-values less than 0.01 ranging from 0.0043 to 0.02.

3.2 Estimated parameters
For each chromosome, we use the λ that minimizes the AIC criterion. A large variation in the optimal λ is observed across the genome, indicating the need for the selection of different λs for different chromosomes. For example, for individual NA19139, the optimal λ ranges from about 47 for Chromosome 15 to about 4900 for Chromosome 19; see Figure 1.

3.3 Choice of thresholds
For each individual, we choose the deletion and duplication thresholds that give the largest total scaled chi-squared value. For individual NA19139, the deletion and duplication thresholds was chosen to be 0.13 (Fig. 2). At the chosen threshold values, after removing the discrepant segments (see Section 2.7), the algorithm identifies a median of 137, 129, 117 and 110 segments that passed the FDR threshold of 10−6, 10−7, 10−8 and 10−9, respectively. The median length of the segments are 15.6, 16.1, 16.7 and 17.3 respectively.

At the same segmentation and FDR thresholds, the single platform algorithm identifies a median of 68, 63, 58 and 56 segments (median length of 39.6, 41.4, 43.3 and 42.8 kb) for the Illumina platform and 81, 77, 69 and 66 segments (median length of 40.8, 41.6, 44.5 and 45.1kb) for the Affymetrix platform.

We apply the MPCBS method on the signals, post-background correction, and use the modified Bayesian information criterion (BIC) approach as suggested by the authors to estimate the number of segments. The maximum number of change points per chromosome is set to 30. MPCBS outputs the breakpoints of the segments as well as the estimated response from each platform. We calculate the estimated response for each segment as the average of the responses from the two platforms. For each individual, we vary the segmentation thresholds such that the total length of CNVs identified by MPCBS is similar to MPSS. Similar results are obtained if we control the number of CNVs, so the results are not shown here. We require that the segments have a minimum of 10 SNPs, a minimum...
With the multiplatform approach, we see strong evidence of a deletion in Chromosome 16 of individual NA19139. Single platform smoothseg on Illumina platform was unable to identify the deletion due to lack of probes in the region. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to insufficient signal. A deletion in Chromosome 16 of individual NA19139. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to complete lack of probes in the region. (A deletion in Chromosome 22 of individual NA19139.

We also ran the single platform CBS algorithm. With \( \alpha = 0.01 \) and segmentation threshold of \( \pm 0.01, 0.02, 0.03 \) and 0.04, we obtain a median of 30, 28, 26 and 26 segments (median length of 21.9, 19, 16.1 and 14.8 kb) for the Illumina platform and a median of 75, 73, 71 and 69 segments (median length of 37.8, 34.8, 34.3 and 31.5 kb) for the Affymetrix platform.

### 3.4 Comparison: simulated data

The average sensitivity and specificity across 100 bootstrap samples are summarized in Table 1. For most scenarios, both MPSS and MPCBS have high specificity (greater than 99%), though MPSS has slightly lower specificity when noise level is decreased. For both algorithms, sensitivity increases with decreased noise level and vice versa. In all cases, MPSS has substantially higher sensitivity than MPCBS. Mean sensitivity for MPSS can be as high as 80% when the noise level is decreased, whereas MPCBS only attains a mean sensitivity of about 41%. When noise level is high, both algorithms perform poorly—MPSS with a mean sensitivity of about 18% and MPCBS with a mean sensitivity of about 10%. However, note that with twice the magnitude of the residuals, the platform variability is increased to four times the original variability. With such high level of noise, unless the CNV signal is very strong, no algorithm is likely to identify the CNV.

### 3.5 Comparison: real data

When signals from the different platforms are consistent, we get increased power to detect the CNVs when we combine the information from the different platforms, especially in areas where a single platform has low density of probes. Figure 3a shows that the Illumina platform has a single probe in the deletion region, and while this probe exhibits strong evidence of a decreased intensity (log2-intensity ratio less than \(-3\)), the single platform approach was unable to identify the deletion. On the other hand, the Affymetrix platform has several probes in the region with moderately decreased log2-intensity ratio values, and the single platform approach detects a slight dip but the evidence is not strong. With the multiplatform approach, we see strong evidence of a deletion in that area. The gray shaded area indicates the CNV region identified by the HapMap 3 project release 3 (downloaded from ftp://ftp.ncbi.nlm.nih.gov/hapmap/cnv_data/hm3_cnv_submission.txt on 20 July 2010); this particular individual NA19139 was found to have a homozygous deletion in this region. In some cases, a single platform is unable to detect the CNV due to complete lack of probes in that region. (Fig. 3b).

When signals from different platforms are inconsistent, it is difficult for the multiplatform method to detect the CNV. Even if the CNV segments are identified, they are likely to be false positives. For example, at the FDR threshold of \( 1\times 10^{-6} \), the true discovery rate for the non-discrepant segments is 15.5% but it is 5.1% for the discrepant segments. On average, we remove 22 discrepant segments per individual.

Figure 4 plots the proportion of bases that overlap with Conrad’s CNVs as a function of total length of CNVs for MPSS and MPCBS. MPSS has a higher proportion of base overlap with Conrad’s CNVs as compared to MPCBS. Figure 5, which plots the amount of overlapping bases as a proportion of Conrad’s CNVs versus the amount of overlapping bases as a proportion of each method’s CNVs, also shows the better performance of MPSS as compared to all the other methods.

### 3.6 Application: breast cancer data

To demonstrate the applicability of the method for large studies, we apply the method to samples from the Cancer Hormone Replacement Epidemiology in Sweden (CAHRES) study, a population-based study which includes women aged 50–74 years, born in Sweden and resident there between October 1, 1995 and March 31, 1995 (Li et al., 2008). A subset of 804 subjects were selected for genotyping on the HumanHap300 and HumanHap240S arrays. Clinical information

---

**Table 1. Sensitivity and specificity of MPSS and MPCBS using simulation data**

<table>
<thead>
<tr>
<th>Data input</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPSS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1*residuals</td>
<td>0.779</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>0.5*residuals</td>
<td>0.804</td>
<td>0.918</td>
</tr>
<tr>
<td>2*residuals</td>
<td>0.127</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>MPCBS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1*residuals</td>
<td>0.527</td>
<td>0.999</td>
</tr>
<tr>
<td>0.5*residuals</td>
<td>0.627</td>
<td>&gt;0.999</td>
</tr>
<tr>
<td>2*residuals</td>
<td>0.177</td>
<td>&gt;0.999</td>
</tr>
</tbody>
</table>
individual. The median length of deletions is 113 kb and that for

out segments with FDR more than 0.01, number of probes less than

objectively chosen values in the previous examples. We further filter

for deletions and duplications, respectively. These are similar to

threshold is fixed at the 5th and 95th percentile of the intensities

\( \lambda \)

to MPSS algorithm. We choose the optimal smoothing parameter,

background correction with the smoothing parameter set to \( \lambda = 10^5 \).

The background-corrected intensity data is then used as inputs to

MPSS algorithm. We choose the optimal smoothing parameter, \( \lambda \) based on the AIC criterion. For convenience, the segmentation threshold is fixed at the 5th and 95th percentile of the intensities for deletions and duplications, respectively. These are similar to objectively chosen values in the previous examples. We further filter out segments with FDR more than 0.01, number of probes less than 10, length of segments less than 1 kb and segments with discrepant signals from the two platforms.

An average of 14 deletions and 4.5 duplications are identified per

histologic grade. Prior to combining data from the two platforms, we

indicates better performance.

of the total length of CNVs identified by the method. A higher proportion of overlap indicates better performance.

Fig. 4. Proportion of bases that overlap with Conrad’s CNVs as a function of of the total length of CNVs identified by the method. A higher proportion of overlap indicates better performance.

made available to us includes lymph node status, tumor size and histologic grade. Prior to combining data from the two platforms, we use rsmooth package(http://www.meb.ki.se/~yad/paw) to perform background correction with the smoothing parameter set to \( \lambda = 10^5 \).

Ten segments are associated with tumor grade (see Supplementary Table T3). Segment 548 in Chromosome 9 overlaps with TPM2 gene, where its protein products were found to be differentially expressed between tumor and non-tumor forming breast cancer cell lines (Harris et al., 2002). Segment 691 in Chromosome 11 overlaps with the PKNOX2 gene, previously shown to be deleted in breast cancer (Iseki et al., 2001).

The 240K array was designed to supplement the 300K array, hence the probes on the two arrays are non-overlapping. The validation of the method in the earlier sections was performed on Affymetrix 6.0 and Illumina 1M arrays which have overlapping probes. Here, we are interested to know if the algorithm works for non-overlapping platforms. However, we do not have a ‘gold standard’ for CNVs of these individuals to make comparisons with. Instead, we take a random sample of non-overlapping 240 000 and 300 000 probes from the Illumina 1M platform for the 9 HapMap samples and make comparisons with the reference CNVs in the same way as before. The true discovery rate and sensitivity for the multiplatform approach is higher than that of the single platform approach: true discovery rate of 0.29 for the multiplatform approach, 0.24 for the 300K array and 0.22 for the 240K array. Sensitivity of 0.027 for the multiplatform approach, 0.027 for the 300K array and 0.017 for the 240K array.

4 DISCUSSION

We have described a new method for identifying CNVs by using data from multiple platforms simultaneously. This method allows researchers to come to a formal consensus result when data from different platforms but for the same individuals are available. The model assumes a random-effects parameter that is common to all platforms, meaning that each platform is assumed to have the same underlying copy-number pattern. We also develop an objective method to segment the estimated random effects parameter (which describes the underlying copy-number pattern) into discrete segments. In addition, we provide a method for calculating a P-value associated with a segment of interest. The P-value would indicate how likely that the segment is a deletion/duplication, and is useful for filtering out likely false positives.

duplications is 140 kb. We use the method in Teo et al. (2010) to form common CNV segments, defined as segments with consecutive probes where there is at least 0.5% of the subjects (~1 subjects) whose individual segments overlap with the probes. We identified 942 common segments (median length of 114.5 kb).

We test each segment for association with tumor size (\( n = 540 \) with size < 2 cm versus \( n = 60 \) with size > 3 cm), lymph node status (\( n = 242 \) lymph-node positive versus \( n = 561 \) negative) and tumor grade (\( n = 118 \) grade-1, \( n = 377 \) grade-2 and \( n = 308 \) grade-3). Fisher’s exact test is used to compute the \( P \)-values. There are no significant associations with tumor size. For lymph-node status, 6 segments have \( P < 0.01 \) (see Supplementary Table T2). Of notable interest is segment 159 in Chromosome 3 which overlaps with the protein tyrosine phosphatase, receptor type, G (PTPRG) gene; overexpression of PTPRG was found to inhibit anchorage-independent growth and proliferation of breast cancer cells (Shu et al., 2010). Another interesting segment is segment 845 in Chromosome 17, which overlaps with the ITGB4 gene, where studies have shown its expression to be correlated with tumor size and nuclear grade (Diaz et al., 2005) and significantly association with basal-like breast cancer (Lu et al., 2008).

An average of 14 deletions and 4.5 duplications are identified per

individual. The median length of deletions is 113 kb and that for
Background correction is needed to make the data from the different platforms comparable; we use a robust smoother that assumes the majority of each chromosome has normal copy-number. While this assumption is likely to be true for germ-line samples, it may not hold for cancer/tumor samples. Recently, Bengtsson et al. (2009) developed a normalization method to bring data from different platforms to the same scale. The method uses a technique based on principal curves to estimate the normalization functions. This method was tested on data from The Cancer Genome Atlas Research Network and seems to work well on tumor samples where there is sufficient deletions and duplications in the genome, but we found that it did not work well with the germ-line samples we use. When we performed Bengtsson et al. (2009)’s normalization on our samples, the correlation in the copy-number estimates between the platforms increased only very slightly after normalization (see Supplementary Fig. S1 and Table T3). This could be due to insufficient CNVs in the data for the principal curves to be identified.

We illustrate the performance of MPSS using real and simulated data sets. In the comparisons using real datasets, we show that MPSS CNVs has greater amount of overlap with the reference as compared to the other methods. In the comparisons using simulated datasets, we show that the proposed method achieve high specificity and accuracy at reasonable noise levels.

In general, for all methods, the proportion of overlapping bases with the highly comprehensive CNV map published by Conrad et al. (2010) is low. However, we believe it is due to the limitation of the SNP arrays rather than the inadequacy of the algorithms. This was also noted by Zhang et al. (2010) where the authors investigated and found that in the regions where the reference CNVs lie, both Affymetrix and Illumina platforms do not have a shift in the intensities and hence the algorithm would not pick out the region as a CNV. Moreover, we do not know if the reference list we have used can be considered the gold standard, since it is not likely to have 100% sensitivity and specificity. Even sequencing methods only show between 72.2% and 96.5% specificity (Xie et al., 2009). The arrival of higher density arrays, for example, the Illumina HumanOmniv2.5 and HumanOmniv5 arrays will likely improve the sensitivity of CNV identification.

Another kind of multiplatform problem arises when there is some stratification of cohorts by chips; for example, if the cases and controls were typed on different chips. Differential sensitivity or false positive rates between the platforms will lead to confounding bias in the case-control comparisons. The method presented here assumes that the data from the different platforms are available for each individual, hence the algorithm could not address this problem. This is an important and valid concern and warrants further investigations.

ACKNOWLEDGEMENT

The authors thank Drs Per Hall, Kamila Czene, Liu Jianjun and Li Jingmei for providing the Swedish breast cancer data.

Funding: Swedish Science Council and National University of Singapore Start-up (Grant No. R-186-000-103-133). National University of Singapore Graduate School for Integrative Sciences and Engineering (NGS) Scholarship (to S.M.T.).

Conflict of Interest: none declared.

REFERENCES


