Allele-specific multi-sample copy number segmentation

Edith M. Ross [1], Kerstin Haase [2], Peter Van Loo [2,3] and Florian Markowetz [1]

[1] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK
[2] The Francis Crick Institute, London NW1 1AT, UK
[2] Department of Human Genetics, University of Leuven, Leuven B-3000, Belgium

Abstract

Motivation: Allele-specific copy number alterations are commonly used to trace the evolution of tumours. A key step of the analysis is to segment genomic data into regions of constant copy number. For precise phylogenetic inference, breakpoints shared between samples need to be aligned to each other.

Results: Here we present `asmultipcf`, an algorithm for allele-specific segmentation of multiple samples that infers private and shared segment boundaries of phylogenetically related samples. The output of this algorithm can directly be used for allele-specific copy number calling using ASCAT.

Availability: `asmultipcf` is available as part of the ASCAT R package (version $\geq$ 2.5) from `github.com/Crick-CancerGenomics/ascat`

1 Introduction

Allele-specific copy number alterations (CNAs) are commonly used to trace the evolution of a tumour. One of the most frequently used algorithms to infer these copy number changes is ASCAT (Van Loo *et al.*, 2010). It segments each sample separately and due to the noise in the data the inferred locations of shared breakpoints are likely to differ between samples. These differences can impair the analysis of phylogenetic relationships between the samples as it depends on the assumption that shared breakpoints appear at exactly the same location. To address this problem, Schwarz *et al.* (2015) performed extensive experimental break point validation, an expensive approach which has often been omitted by similar papers. Mangiola *et al.* (2016), for example, used a size-based heuristic filter for CNAs instead.

To rigorously address the problem of multi-sample breakpoint detection, we have developed `asmultipcf` (allele-specific multi-sample piecewise constant fitting), an algorithm performing allele-specific segmentation of multiple samples for the inference of private and shared segment boundaries of phylogenetically related samples. `asmultipcf` enforces joint segment boundaries across samples unless the data contains significant evidence that the breakpoints differ.

2 Approach

`asmultipcf` is based on the copy number segmentation algorithms developed by Nilsen *et al.* (2012), which use penalized least square principles to fit piecewise constant segments to the data and which are used by ASCAT (Van Loo *et al.*, 2010) to infer allele specific copy numbers. `asmpultipcf` combines two algorithms by Nilsen *et al.* (2012), aspcf (an allele-specific single-sample segmentation method) and multipcf (a non-allele-specific multi-sample segmentation method), to enable the joint segmentation of multiple related samples in an allele specific manner. Additionally, `asmpultipcf` handles missing values, making extensive data filtering unnecessary.

Input data: For each sample the following input data are required across germline heterozygous sites: (i) log ratios (logR), representing log-transformed copy numbers derived from sequencing depth or SNP array data, and (ii) B allele frequencies (BAF), describing the allelic imbalance of SNPs. Both measurements can be derived from whole genome sequencing data. The algorithm presented here can handle missing values and thus loci with incomplete data across samples do not need to be excluded.

Preprocessing: `asmultipcf` uses the same pre-processing steps as the allele-specific single sample algorithm proposed by Nilsen *et al.* (2012). Most importantly BAFs are mirrored in order to obtain a single track in regions of allelic imbalance and extreme outliers, which are likely to be noise, are removed from logR and BAF data (see Nilsen *et al.* (2012) for details of the pre-processing). Given the input data of n samples across p SNP loci, the pre-processing yields a single matrix $\mathbf{Y} = (\mathbf{y}_{ij}) \in \mathbb{R}^{2n \times p}$ that contains both logR and BAF values.

An exact algorithm for weighted segmentation: We extend the penalized least squares approach of Nilsen *et al.* (2012) to evaluate the fit of a segmentation solution to the data, and use a weighted least squares function to model missing values in the data matrix. A weight matrix $\mathbf{W} = (\mathbf{w}_{ij}) \in \mathbb{R}^{2n \times p}$ is derived by assigning w_{ij} a weight of 0 if y_{ij} is missing and 1 otherwise. Then all missing values in $\mathbf{Y}$ are assigned an arbitrary (non-NA) value. Our aim is to find a segmentation $S = \{I_1, \ldots, I_M\}$ that minimizes the cost function

$$L(S|\mathbf{Y}, \mathbf{W}, \gamma) = \sum_{i=1}^{2n} L(S|\mathbf{y}_{i.}, \mathbf{w}_{i.}, \gamma) \tag{1}$$

$$= \sum_{i=1}^{2n} \sum_{I \in S} \sum_{j \in I} w_{ij}(y_{ij} - \overline{y_{i,I}})^2 + \gamma|S|, \tag{2}$$

where the best fit on a given segment I is the weighted average of the observations on that segment

$$\overline{y_{i,I}} = \frac{\sum_{j \in I} w_{ij} y_{ij}}{\sum_{j \in I} w_{ij}}$$

and where γ is a penalty parameter that controls the number of segments. Expanding the square in (2) and omitting the term that is independent of the segmentation we find

$$L'(S|\mathbf{Y}, \mathbf{W}, \gamma) = -\sum_{i=1}^{2n} \sum_{I \in S} \frac{\left(\sum_{j \in I} w_{ij} y_{ij}\right)^2}{\sum_{j \in I} w_{ij}} + \gamma|S|.$$

To find an optimal solution to the cost function we adapt the dynamic programming algorithm presented by Nilsen *et al.* (2012) to our weighted problem.

Algorithm 1: asmultipcf

Input: Matrix $\mathbf{Y}$ of log-transformed copy numbers and B allele frequencies; Weight matrix $\mathbf{W}$; penalty $\gamma > 0$;
Output: Segment start indices and segment averages

1. Calculate scores by setting $\mathbf{A}_0 = [\,]$, $\mathbf{C}_0 = [\,]$, $\mathbf{e}_0 = 0$ and iterate for $k = 1, \ldots, p$
   - $\mathbf{A}_k = [\mathbf{A}_{k-1}\ 0] + \mathbf{w}_{.k}\mathbf{y}_{.k}$
   - $\mathbf{C}_k = [\mathbf{C}_{k-1}\ 0] + \mathbf{w}_{.k}$
   - $\mathbf{d}_k = -\mathbf{1}^T \left(\mathbf{A}_k \circ \mathbf{A}_k \circ \mathbf{C}_k^{\circ -1}\right)$ where $\circ$ denotes an element-wise matrix product and $\mathbf{C}_k^{\circ -1}$ the element-wise inverse
   - $\mathbf{e}_k = [\mathbf{e}_{k-1}\ \min(\mathbf{d}_k + \mathbf{e}_{k-1} + \gamma)]$

   storing also the index $t_k \in 1, \ldots, k$ at which the minimum in the last step is achieved.
2. Find segment start indices from right to left as $s_1 = t_p$, $s_2 = t_{s_1 - 1}$, $\ldots$, $s_M = 1$, where $M \leq 1$.
3. Find segment averages

$$\overline{\mathbf{y}_m} = \frac{(\mathbf{w}_{.s_m}\mathbf{y}_{.s_m} + \cdots + \mathbf{w}_{.s_{m-1}-1}\mathbf{y}_{.s_{m-1}-1})}{(\mathbf{w}_{.s_m} + \cdots + \mathbf{w}_{.s_{m-1}-1})}$$

A heuristic algorithm for large data sets: Algorithm 1 is of order $O(p^2)$, which means that the segmentation becomes computationally expensive for long sequences. However, instead of allowing breakpoints at any of the p positions, we can pre-select potential breakpoints and thereby reduce the runtime to $O(q^2)$ where q is the number of potential breakpoints. To identify potential breakpoints, different heuristics can be used. Here, we apply Algorithm 1 to overlapping subsequences, combine all of the inferred breakpoints and use them as input for the subsequent global segmentation. As in the implementation by Nilsen *et al.* (2012) we use subsequences of length 5000 with an overlap of 1000. Algorithm 2 describes the fast heuristic version of `asmultipcf`.

Algorithm 2: Fast asmultipcf

Input: Matrix $\mathbf{Y}$ of log-transformed copy numbers and B allele frequencies; Weight matrix $\mathbf{W}$; penalty $\gamma > 0$;
Output: Segment start indices and segment averages

1. Split data set into overlapping subsequences and apply steps 1 and 2 of Algorithm 1 to each of them in order to find potential breakpoints $r_0, r_1, \ldots, r_q$ where $r_0 = 1$ and $r_1 = p + 1$.
2. Aggregate sequences between breakpoints by setting $x_{ik} = \sum_{j=r_{k-1}}^{r_k-1} w_{ij} y_{ij}$ and $v_{ik} = \sum_{j=r_{k-1}}^{r_k-1} w_{ij}$.
3. Calculate segmentation solution by using the aggregated matrices $\mathbf{X}$ and $\mathbf{V} \in \mathbb{R}^{2n \times q}$ as input to Algorithm 1 instead of $\mathbf{Y}$ and $\mathbf{W}$, respectively.

Post-processing: Both algorithms yield a single segmentation solution S for all samples. However, we expect that only some of the segments will be shared between all samples while others will be private. While ASCAT can be run directly on the global segmentation solution, removing unnecessary breakpoints on a per sample base can reduce noise in the segment average estimates by generating larger segments. To refine breakpoints individually for each sample, we simply use the breakpoints inferred from the multi-sample segmentation and rerun steps 2 and 3 of Algorithm 2 on each sample individually based on these potential break points.

Implementation: `asmultipcf` is part of the ASCAT R package from version 2.5 onwards. The asmultipcf function contains a parameter to select whether the exact or the fast algorithm should be run, as well as an option to include the per-sample breakpoint refinement. Furthermore, samples can be weight adjusted to account for quality differences in the data. An example of how to use this function can be found in the package manual.

3 Discussion

The independent segmentation of related samples can artificially inflate tumor heterogeneity. The algorithm presented here addresses this problem by joint segmentation. While this approach can potentially underestimate tumor heterogeneity, because CNAs that are shared by many samples are more likely to be detected than CNAs that are private or shared by only few samples, in practice the penalty parameter γ can be adjusted to ensure sensitivity. Overall, `asmultipcf` substantially improves the analysis of copy number changes of multiple samples.

Funding

EMR and FM would like to acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. Parts of this work is funded by CRUK core grant C14303/A17197 and A19274. This research is supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute.

References

Mangiola, S. *et al.* (2016). Comparing nodal versus bony metastatic spread using tumour phylogenies. *Scientific Reports*, **6**, 33918 EP –.

Nilsen, G. *et al.* (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*, **13**(1), 591.

Schwarz, R. F. *et al.* (2015). Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic analysis. *PLOS Medicine*, **12**(2), 1–20.

Van Loo, P. *et al.* (2010). Allele-specific copy number analysis of tumors *Proceedings of the National Academy of Sciences*, **107**(39), 16910–16915.