Bioinformatics, 37(13), 2021, 1909–1911
doi: 10.1093/bioinformatics/btaa538
Advance Access Publication Date: 25 May 2021
Applications Note

Genome analysis

Allele-specific multi-sample copy number segmentation in ASCAT

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

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK and [2]The Francis Crick Institute, London NW1 1AT, UK

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Lenore Cowen

Received on September 9, 2019; revised on April 8, 2020; editorial decision on May 16, 2020; accepted on May 19, 2020

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

Contact: florian.markowetz@cruk.cam.ac.uk

1 Introduction

Allele-specific copy number alterations (CNAs) are commonly used to trace the evolution of tumours. One of the most frequently used algorithms to infer these copy number changes is ASCAT (Van Loo *et al.*, 2010), which segments each sample separately. Due to measurement noise, the inferred locations of breakpoints shared between samples often differ. These differences can impair analyses of phylogenetic relationships between the samples, because evolutionary methods depend on the assumption that shared breakpoints appear at exactly the same location. Previous approaches to address this problem include extensive experimental breakpoint validation (Schwarz *et al.*, 2015), an expensive approach that is not always feasible, or size-based heuristic filters (Mangiola *et al.*, 2016). Another approach infers allele and clone-specific CNA from multi-sample data by binning without segmentation (Zaccaria and Raphael, 2018).

To rigorously address the problem of multi-sample breakpoint detection, we have developed `asmultipcf` (allele-specific multi-sample piecewise constant fitting), a robust allele-specific multi-sample segmentation algorithm that is tightly integrated into the ASCAT framework (Van Loo *et al.*, 2010). The ability of `asmultipcf` to improve phylogenetic inference was shown in a large case study on 181 samples from 10 patients with lethal metastatic breast cancer (De Mattos-Arruda *et al.*, 2019).

2 Approach

`asmultipcf` incorporates and extends two copy number segmentation algorithms previously developed by Nilsen *et al.* (2012), which leverage vector operations for efficient implementation: first, `aspcf` (an allele-specific segmentation method for single samples), and second, `multipcf` (a multi-sample segmentation method, which is not allele-specific). Additionally, `asmultipcf` handles missing values, making extensive data filtering unnecessary.

2.1 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 single nucleotide polymorphism (SNP) array data, and (ii) B allele frequencies (BAF), describing the allelic imbalance of SNPs. The algorithm presented here can handle missing values and thus loci with incomplete data across samples do not need to be excluded.

2.2 Pre-processing

`asmultipcf` uses the same pre-processing steps as the allele-specific single sample algorithm of Nilsen *et al.* (2012), including (i) mirroring BAFs to obtain a single track in regions of allelic imbalance and (ii) removing extreme outliers from logR and BAF data [see Nilsen *et al.* (2012) for details]. Given 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.

2.3 An exact algorithm for weighted segmentation

We evaluate the fit of a segmentation solution to the data with a weighted least squares function that models 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-not assigned (NA)] value. Our aim is to find a segmentation $S = \{I_1, \ldots, I_M\}$ that minimizes the cost function

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

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

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

$$\bar{y}_{i,I} = \sum_{j \in I} w_{ij} y_{ij} / \sum_{j \in I} w_{ij}$$

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

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

To find an optimal solution to the cost function, we adapt the dynamic programming algorithm of Nilsen *et al.* (2012) to our weighted problem. The algorithm iteratively minimizes the total errors $\mathbf{e}_k$ at locus k across all samples using the errors $\mathbf{e}_{k-1}$ up to k, the costs of the current segments, $\mathbf{d}_k$, and the penalty γ, together with intermediate variables $\mathbf{A}_k$ and $\mathbf{C}_k$:

2.4 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 (length 5000 with an overlap of 1000), combine all of the inferred breakpoints and use them as input for the subsequent global segmentation. Algorithm 2 describes the fast heuristic version of `asmultipcf`.

2.5 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 breakpoints.

2.6 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. The manual contains example use cases, including a comparison to HATCHet (Zaccaria and Raphael, 2018).

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. Initialize $\mathbf{A}_0 = [\,]$, $\mathbf{C}_0 = [\,]$, $\mathbf{e}_0 = 0$ and iterate for $k = 1, \ldots, p$
   - $\mathbf{A}_k = [\mathbf{A}_{k-1} 0] + \mathbf{w}_{.k}\mathbf{y}_{.k}$
   - $\mathbf{C}_k = [\mathbf{C}_{k-1} 0] + \mathbf{w}_{.k}$
   - $\mathbf{d}_k = -1^T(\mathbf{A}_k \circ \mathbf{A}_k \circ \mathbf{C}_k^{\circ -1})$ 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 $s_1 = t_p$, $s_2 = t_{s_1 - 1}$, …, s_M=1, where $M \leq 1$.
3. Find segment averages

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

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, …, 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.

3 Discussion

The independent segmentation of related samples can artificially inflate tumour heterogeneity. The algorithm presented here addresses this problem by joint segmentation. While this approach can potentially underestimate tumour 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

This research was supported by the Cancer Research UK Cambridge Institute with core grant C14303/A17197 and the Francis Crick Institute with core funding from Cancer Research UK [FC001202], the UK Medical Research Council [FC001202] and the Wellcome Trust [FC001202]. P.V.L. is a Winton Group Leader, F.M. is a Royal Society Wolfson Research Merit award holder.

Conflict of Interest: none declared.

References

De Mattos-Arruda,L. *et al*. (2019) The genomic and immune landscapes of lethal metastatic breast cancer. *Cell Rep*., **27**, 2690–2708.e10.

Mangiola,S. *et al*. (2016) Comparing nodal versus bony metastatic spread using tumour phylogenies. *Sci. Rep*., **6**, 33918.

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

Schwarz,R.F. *et al*. (2015) Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med*., **12**, e1001789.

Van Loo,P. *et al*. (2010) Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA*, **107**, 16910–16915.

Zaccaria,S. and Raphael,B.J. (2018) Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *bioRxiv*, 496174v1.