Genome analysis

Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution

Björn Nilsson1,2,3,*, Mikael Johansson4, Fatima Al-Shahrour1,3, Anne E. Carpenter1 and Benjamin L. Ebert1,3

1Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA, 2Department of Hematology and Transfusion Medicine, Lund University Hospital, SE-221 85 Lund, Sweden, 3Hematology Division, Brigham and Women’s Hospital, Harvard Medical School, One Blackfan Circle, Boston, MA 02115, USA and 4Department of Automatic Control, Royal Institute of Technology, SE-100 44 Stockholm, Sweden

Received on January 12, 2009; revised on February 6, 2009; accepted on February 13, 2009

ABSTRACT

Motivation: Multimillion-probe microarrays allow detection of gains and losses of chromosomal material at unprecedented resolution. However, the data generated by these arrays are several-fold larger than data from earlier platforms, creating a need for efficient analysis tools that scale robustly with data size.

Results: We developed a new aberration caller, Ultrasome, that delineates genomic changes-of-interest with dramatically improved efficiency. Ultrasome shows near-linear computational complexity and processes latest generation copy number arrays about 10 000 times faster than standard methods with preserved analytic accuracy.

Availability: www.broad.mit.edu/ultrasome.

Contact: bnilsson@broad.mit.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Microarray-based DNA copy number profiling has transformed the identification and characterization of gains and losses of chromosomal material. The technology is evolving rapidly in terms of genomic resolution. The most recent generation of microarrays, including Affymetrix SNP6.0 (McCarroll 

et al., 2005), measure copy number at millions of chromosomal locations, an increase of up to 10-fold compared with earlier platforms. Even denser arrays are underway, and copy number profiling based on next-generation sequencing is rapidly gaining traction.

Alongside probe-level copy number estimation, the central step in copy number data analysis is to partition the genome into contiguous regions that share the same copy number on average. With increasing resolution, this has become challenging as current standard methods, originally developed for lower resolution microarrays, are associated with computational requirements that grow steeply with the number of probes (Lai et al., 2005). This leads to long wait times, increases the need for extraordinary computing resources, and complicates analysis.

To address this issue, we developed a new aberration caller, Ultrasome, based on an efficient computational strategy that exploits the structure of the delineation problem to process copy number data in near-linear time. As illustrated here, Ultrasome is capable of processing latest generation copy number arrays, including Affymetrix SNP6.0, about 10 000 times faster than standard approaches while retaining comparable analytic accuracy.

2 RESULTS

The mathematical details are described in Supplementary Material. In short, partitioning a chromosome amounts to fitting a piece-wise constant function to the data, in our case by minimizing

$$
\sum_{i=1}^{M} \sum_{j \in I_i} |f_j - \mu_i|^2 + \lambda M,
$$

where $f_1, \ldots, f_M$ are DNA copy numbers indexed by chromosomal position, $I_1, \ldots, I_M$ a set of ordered subintervals (segments) covering the interval $[1,N]$, $\mu_1, \ldots, \mu_M$ the corresponding segmental copy numbers and $M$ the number of segments (true value unknown a priori). The first term imposes consistency with the original data; the second term imposes regularity by penalizing the number of breakpoints. By adjusting the parameter $\lambda$, the balance between consistency and regularity can be set and the method optimized for the detection of small or large aberrations (details and guidelines in Supplementary Material).

To minimize (1), we exploit that, for any partitioning, the optimal segmental copy numbers are the averages of the point-wise copy numbers over the subintervals, allowing the solution space to be re-parameterized as an $N$-dimensional binary space where each coordinate indicates whether a point is a breakpoint (a starting point of a segment) or a non-breakpoint (an interior point). In this space, we aim to find a sequence of increasingly better partitionings in the sequence be related by toggling the breakpoint status of exactly one point, we can proceed by repeatedly identifying the site on the chromosome whose change in state from breakpoint to non-breakpoint (or vice versa) reduces the value of Equation (1) maximally, toggling the state of that point, and repeating until no improvement can be found. By use of a special data structure, a heap-sorted queue with backpointers, our approach can be accelerated to

*To whom correspondence should be addressed.

1078 © The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
Ultrasome shows receiver operating characteristics as strong as those of current standard methods. The figure also exemplifies the effect of changing the implementation. Either way, our results reflect the expected performance of the tools that are directly available to users.

Despite the increased efficiency, Ultrasome shows receiver operating characteristics as strong as those of current standard methods. The figure also exemplifies the effect of changing the breakpoint penalty \( \lambda \) to optimize the detection of small aberrations (low \( \lambda \), solid blue) or large unbroken aberrations (high \( \lambda \), dashed blue).

**REFERENCES**


