BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 8 2009, pages 1078–1079
doi:10.1093/bioinformatics/btp091

Genome analysis

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

Björn Nilsson[1,2,3,*], Mikael Johansson[4], Fatima Al-Shahrour[1,3], Anne E. Carpenter[1] and Benjamin L. Ebert[1,3]

[1]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA, [2]Department of Hematology and Transfusion Medicine, Lund University Hospital, SE-221 85 Lund, Sweden, [3]Hematology Division, Brigham and Women's Hospital, Harvard Medical School, One Blackfan Circle, Boston, MA 02115, USA and [4]Department 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

Advance Access publication February 19, 2009

Associate Editor: Alfonso Valencia

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.*, 2008), 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.

*To whom correspondence should be addressed.

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

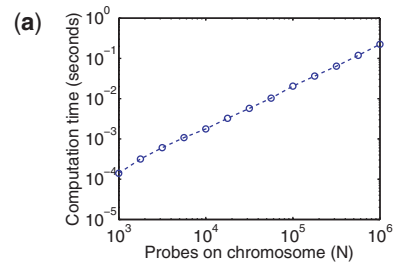
$$\sum_{i=1}^{M}\sum_{j\in I_i}(f_j-\mu_i)^2+\lambda M, \quad (1)$$

where $f_1,\ldots,f_N$ 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 λ, 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 that leads towards the minimum. By requiring that successive 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

 © The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article/25/8/1078/323999 by guest on 18 June 2024



(b)

	Glioblastoma	HCC1143
Ultrasome	0.40	0.32
CBS (Venkatraman and Olshen, 2007) (fast version)	8511	9315
CGHclassify (Engler *et al*, 2006)	36 550	26 810
GLAD (Hupé *et al*, 2004)	57 457	34 762
CGHseg (Picard *et al*, 2005)	n/a†	n/a†
SW-ARRAY (Price *et al*, 2005)	n/a†	n/a†
CLAC (Wang *et al*, 2005)	n/a†	n/a†
ACE (Lingjaerde *et al*, 2005)	n/a†	n/a†
BioHMM (Marioni *et al*, 2006)	n/a†	n/a†

Computation time (seconds per Affymetrix SNP6.0 array, 2.5 GHz CPU, 3 GB RAM). File access not included.
†Failed for large data or was terminated after several days wait for one array.

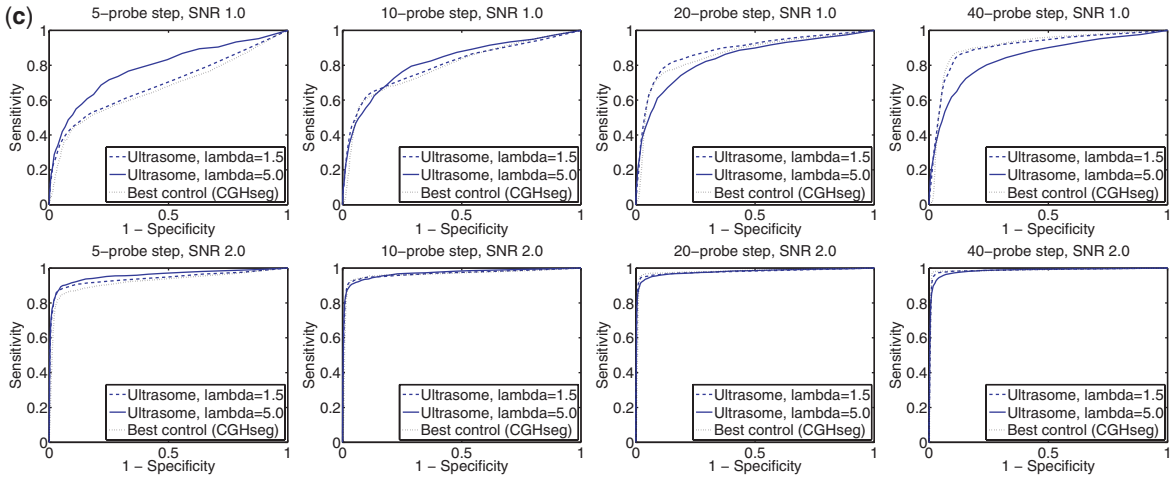


Fig. 1. **(a)** Computation time by number of probes per chromosome (average across 1000 random locations in the HCC1143 data). As a reference, Affymetrix SNP6.0 measures 1.43×10^5 copy numbers on chromosome 1. **(b)** Computation times for glioblastoma and HCC1143 data (Affymetrix SNP6.0; 1.85×10^6 pointwise copy number estimates). Some reference methods failed when applied to large data (n/a), potentially for algorithmic reasons or because of suboptimal implementation. Either way, our results reflect the expected performance of the tools that are directly available to users. **(c)** 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 λ to optimize the detection of small aberrations (low λ, solid blue) or large unbroken aberrations (high λ, dashed blue).

$O(\log N)$ per-iteration complexity and near-linear empirical overall complexity (Fig. 1a).

We tested Ultrasome on Affymetrix SNP6.0 profiles (1.85×10^6 probes) of the breast cancer cell line HCC1143 (from our lab) and glioblastoma multiforme (TCGA Network, 2008). Remarkably, the computation time was <1 s per array, orders of magnitudes faster than current standard methods (Fig. 1b). For a study with hundreds of samples, this translates to a reduction in wait time from days to minutes. Such increased efficiency is not only a matter of convenience, but also facilitates the tuning of technical parameters to the needs of particular studies (Supplementary Material). To verify that known aberrations are detected, we constructed a set of artificial 'chromosomes' with spiked-in aberrations of varying widths and noise levels and computed receiver operating characteristics for each case (Supplementary Material). In this test, an established benchmark (Lai *et al*., 2005), Ultrasome performed on par with current methods (Fig. 1c).

In conclusion, Ultrasome is a high-performance tool designed to facilitate the detection of chromosomal aberrations in copy number data of multimillion-probe or higher resolution. The program is available in a command-line version (Windows and Linux) and a graphical user interface version (Windows), accepts data in standard formats, and interfaces with Integrative Genomics Viewer (`www.broad.mit.edu/igv`) to allow data visualization.

Conflict of Interest: none declared.

REFERENCES

Engler,D. *et al*. (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, **7**, 399–421.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Lai,W. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lingjaerde,O. *et al*. (2005) CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**, 821–822.

Marioni,J. *et al*. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

McCarroll,S. *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174 .

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Price,T. *et al*. (2005) SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res.*, **33**, 3455–3464.

TCGA Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Venkatraman,E. and Olshen,A. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

Downloaded from https://academic.oup.com/bioinformatics/article/25/8/1078/323999 by guest on 18 June 2024