

Statistical Analysis of Array CGH (STAC): A Novel Approach to Analyzing Multiple Experiments

Sharon J. Diskin¹, Thomas Eck², Joel Greshock³, Christian Stoeckert⁴,
Barbara L. Weber⁵, John M. Maris⁶ and Gregory R. Grant⁷

Keywords: microarray, array CGH, algorithm, statistical significance, genomic copy number, neuroblastoma, breast cancer

1 Introduction.

The development and progression of cancer is an evolutionary process marked by the accumulation of genomic and genetic alterations. Genomic copy number aberrations are frequently found in solid tumors and there is widespread belief that recurring regions of gain/loss harbor genes critical to the genesis or progression of cancer [1]. For example, tumor suppressor genes may become inactivated as a result of genomic material being lost while amplification of oncogenes may lead to over-expression at the mRNA or protein level. Array Comparative Genomic Hybridization (aCGH) is a recently developed experimental technique for detecting amplifications and deletions of DNA with high resolution on a genome-wide scale. In comparison to mRNA expression arrays for which numerous analysis methods have been developed, there are relatively few statistical methods for analyzing aCGH data. When it comes to the task of identifying regions of interest across multiple experiments, researchers often employ simple frequency cutoffs. This is followed by a tedious manual review of the regions to define boundaries. While this approach may identify some relevant locations, it is subject to investigator bias and requires considerable time. Perhaps more importantly, this approach lacks the power to uncover regions shared only within subsets of the data. These regions may be of great importance for class discovery efforts aimed at identifying new biologically or clinically relevant subtypes within heterogeneous cancers. Therefore, there is a need for rigorous but efficient computational methods to assess the statistical significance of copy number aberrations across multiple samples; these methods should also include a subset search of the sample space.

2 Methods.

Given the complex geometric nature of aCGH data, it is difficult to define statistics which address these goals with reasonable power. We introduce a new algorithm, called STAC. Our approach is to combine a frequency based statistic (the "max") with a statistic we call the "footprint" of the data. The footprint measures the span of the aberrations over a particular location and was designed to be

¹Penn Center for Bioinformatics (PCBI) and Abramson Cancer Research Center, University of Pennsylvania, Philadelphia, USA, E-mail: diskins@pcbi.upenn.edu

²Penn Center for Bioinformatics, Philadelphia, USA, E-mail: teck@optonline.net

³Abramson Family Cancer Research Center, Philadelphia, USA, E-mail: greshock@mail.med.upenn.edu

⁴Penn Center for Bioinformatics (PCBI), Philadelphia, USA, E-mail: stoeckrt@pcbi.upenn.edu

⁵Penn Center for Bioinformatics (PCBI), Philadelphia, USA, Email: stoeckrt@pcbi.upenn.edu

⁶Abramson Cancer Research Center, Philadelphia, USA, E-mail: weberb@mail.med.upenn.edu

⁷Children's Hospital of Philadelphia (CHOP) and Abramson Cancer Research Center, Philadelphia, USA, E-mail: maris@email.chop.edu

⁷Penn Center for Bioinformatics (PCBI), Philadelphia, USA, E-mail: ggrant@pcbi.upenn.edu

sensitive to tight alignments of aberrant intervals across samples. As such, the footprint provides a good complement to the frequency based statistic. We address the need to identify subtypes within the data by evaluating the footprint in a greedy and incremental manner for all subset sizes (from 2 to $n = \#$ samples). Using a permutation approach we assess the significance of the statistic and assign p-values to every 1 MB region on the genome. Graphical views of the output on both a genome-wide scale and a more detailed chromosome arm view are accessible through STACview, our supporting web based visualization tool.

3 Validation and Results

We apply our method to BAC based aCGH data from 42 neuroblastoma cell lines [2] and 55 sporadic primary breast tumors [3]. We validate our method using both well characterized regions of known biological significance in these cancers as well as previously reported regions for these datasets based on simple frequency cutoffs and manual inspection. We consider a p-value < 0.05 as statistical support for a region being significantly aberrant. Based on this criteria, our results provide statistical support for over 95% of the previously reported regions. A high level of agreement in terms of boundary placement is observed for both data sets with the majority of STAC boundaries differing by less than 1MB from those previously reported. In several cases our results narrow the previously reported regions. The algorithm reveals many novel regions of aberration that warrant further investigation. In addition, unsupervised clustering based on our STAC results reveals two distinct subgroups of high risk neuroblastoma characterized by high order associations across multiple chromosomes that have not been reported previously. We conclude that STAC is a powerful new method for the analysis of aCGH data from multiple experiments.

4 References

- [1] Albertson DG, Collins C, McCormick F, Gray JW. Chromosome Aberrations in solid tumors. *Nature Genetics*;34:3 (2003).
- [2] Mosse YP, Greshock J, Margolin A, Naylor T, Khazi D, Hii G, Winter C, Biegel JA, Maris JM, Weber BL. Detection of single copy and more complex DNA alterations in neuroblastoma with array-based comparative genomic hybridization 2005. *Cancer Research*, in press.
- [3] Naylor T, Greshock J, et al.. Genome wide copy number analysis of 55 Sporadic Breast Tumors. *Breast Cancer Research*, Submitted.