STAC: A method for testing the significance of DNA copy-number aberrations across multiple array-CGH experiments

Sharon J. Diskin* Thomas Eck† Joel Greshock‡ Yael P. Mosse§ Tara Naylor¶
Christian J. Stoeckert, Jr.‖ Barbara L. Weber** John M. Maris†† Gregory R. Grant‡‡

September 13, 2005

1 Summary

Recurrent genomic DNA amplifications and deletions characterize cancer genomes and often contribute to disease evolution. Genomic copy number aberrations (CNAs) can now be detected at high resolution using microarray-based techniques. However, robust statistical methods are needed to identify non-random gains and losses across multiple experiments/samples. We have developed a method called significance testing for aberrant copy number (STAC) to address this need. STAC utilizes two complementary statistics in combination with a novel search strategy. We assess the significance of both statistics and assign p-values to each location on the genome using a multiple testing corrected permutation approach. The details of our method are described in this document. A Java version of STAC is freely available for download at `http://cbil.upenn.edu/STAC`.

2 STAC Method

Since STAC analysis currently focuses on gain and loss as separate cases, we will use the term "aberration" as the generic term for both, but the type of abberation (gain or loss) is fixed throughout this discussion. The input data consist of an aberration call for each of N samples and L genomic locations. We call the sequence of L locations the *genome*. We represent aberration with a 1 and no aberration with a 0. Therefore the data can be put in array of 0s and 1s where rows represent experiments and columns represent locations. We refer to a single row of the data array as a *profile*. A set of consecutive 1s in a row is called an (aberrant) *interval* for that profile. Therefore each profile consists of a set of intervals and their locations. Figure 1 shows an example of data consisting of $N = 37$ profiles and $L = 77$ locations. We are interested in finding those regions of aberration which occur across the samples more often than would be expected by chance.

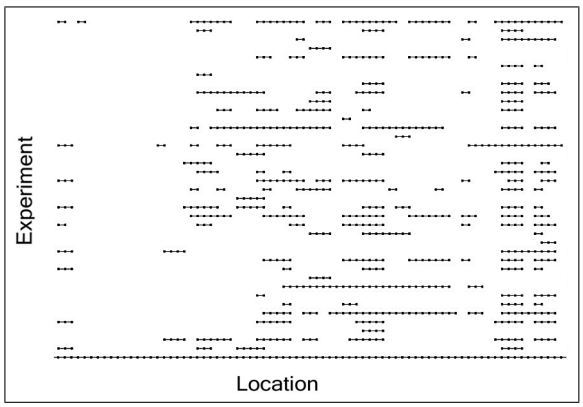

Figure 1: Example data consisting of loss calls on chromosome 11q. Rows represent samples and columns represent chromosomal locations. A black dot means there was a loss call made for that sample at that location. Consecutive black dots are connected by a line to represent an interval of aberration.

The simplest method is to set a threshold c for the frequency and then to consider for follow-up all locations which are aberrant in at least c samples. Since that method lacks statistical control over the false positives and false negatives, it is desirable to assign p-values to the different possible choices of c. Furthermore, c might not be the most powerful statistic to use for finding consistent aberration. For example, just as it increases power dramatically to replace a the simple ratio of means in differential gene expression analysis by the t-statistic, we will find the simple frequency statistic c can also be replaced by a statistic which is much more powerful for measuring concordance.

In order to calculate significance we need a reasonable null model. We don't have a model for the rate of aberration itself, instead we take the aberrations in the individual samples as given and test for the significance of *consistent* aberration across samples. The null model we use to test for this is that the observed intervals of aberration are equally likely to occur anywhere in the stretch of the genome being considered. The general approach is to choose an appro-

*Penn Center for Bioinformatics, University of Pennsylvania and Division of Oncology, Childrens Hospital of Philadelphia
†Penn Center for Bioinformatics (PCBI), University of Pennsylvania
‡Abramson Family Cancer Research Center, University of Pennsylvania
§Division of Oncology, Childrens Hospital of Philadelphia
¶Abramson Family Cancer Research Center, University of Pennsylvania
‖Deptartment of Genetics, University of Pennsylvania School of Medecine
**Abramson Family Cancer Research Center, University of Pennsylvania
††Division of Oncology, Childrens Hospital of Philadelphia and Abramson Family Cancer Research Center, University of Pennsylvania
‡‡Penn Center for Bioinformatics(PCBI), University of Pennsylvania

priate statistic, and then to apply a permutation procedure under the null model to determine the significance of the statistic. We obtain an estimate of the null distribution via permutations. A permutation consists of a random rearrangement of the intervals of each profile. In this way we preserve as much of the nature of the data as possible, except for consistency across samples. For example if a profile with M locations had only one interval of length ℓ, then there would be $L - \ell + 1$ permutations of this profile, each equally likely.

In order to avoid finding certain trivial violations of the null model, such as the fact that aberrations generally occur less frequently, if at all, in the centromeres, one should apply STAC to each chromosome arm separately.

2.1 The frequency statistic

The first statistic we consider associates to each location m the frequency of aberration over that location, denoted $\mathcal{F}(m)$. The statistic $\mathcal{F}(m)$ is an indicator of the deviation from the null model at location m. If p is a permutation of the data, we denote the frequency of aberration in the permuted data over location m by $\mathcal{F}_p(m)$, $m = 1, \ldots, L$. Let D be the distribution (over all permutations p) of $\max_m \mathcal{F}_p(m)$. In other words, D is the permutation distribution of the maximum frequency. We define a frequency based p-value at location m_0 by the right-hand tail probability of D at m_0. I.e.

$$p_{\mathcal{F}}(m_0) = \frac{\#p \text{ such that } \max_m \mathcal{F}_p(m) < \mathcal{F}(m_0)}{\text{total number of permutations}}.$$

Since we are comparing $\mathcal{F}(m_0)$ to the distribution of the maximum frequency over all m, the resulting p-value $p_{\mathcal{F}}(m_0)$ is a multiple testing corrected confidence measure (over $m = 1, \ldots, L$) for rejection of the null model. Since our statistic is an indicator of behavior at location m_0, we prioritize the locations by the $p_{\mathcal{F}}(m_0)$. If location m_0 is significant to level α (usually .05 or .01), then location m_0 has a frequency which is unlikely under the null model, indicating a likelyihood of biological significance. The regions can then be followed up to determine which are true biological effects and which are experimental artifacts (perhaps due to faulty array elements). We define the *confidence* at location m as $1 - p_{\mathcal{F}}(m_0)$. Figure 2 shows the data from Figure 1 with the confidences overlayed at each location as grey bars. The red line graphs the actual frequency. One can see from these data how the frequency is not significant at the three leftmost locations, however there is clearly something going on there. The fact that the frequency is nine with tightly aligned short intervals is more striking than, for example, the frequency of 10 we see at location marked with an asterisk ("$*$"). Therefore we would like a statistic which is also sensitive to these regions of tight alignment, even if they are not significantly frequent.

2.2 The footprint and the normalized footprint

To overcome the shortcoming of the frequency based statistic outlined at the end of the previous section, we utilize

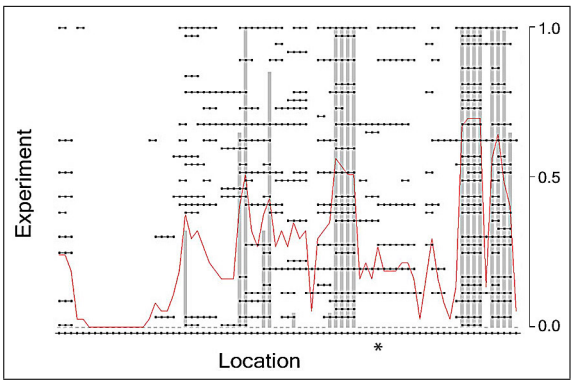

Figure 2: Display showing the frequency confidences only, indicated by grey bars. The red line graphs the actual frequencies.

a statistic and methodology originally introduced by Grant *et al.* in the context of direct Identity-By-Descent (IBD) mapping [1]. The biological question here is quite different, however the data and statistical problems which arise from them are quite similar.

A *stack* is defined as a set of intervals which contains at most one interval per profile and where there is at least one common location to every interval in the set. Note that in Grant *et al.* the second requirement is not imposed. We define the *footprint* $\mathcal{F}(\mathcal{S})$ of the stack $\mathcal{S}$ to be the number of locations c such that c is contained in some interval in the stack $\mathcal{S}$ (see Figure 3).

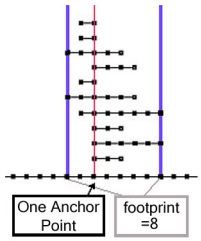

Figure 3: The footprint of a stack is the number of locations contained in some interval of the stack. The anchor points of a stack are the locations contained in every interval of the stack.

Suppose there are five intervals from four different profiles each of length four. There are many ways to position the intervals relative to each other to have a maximum frequency of five. However, there is only one way to position them relative to each other to have a footprint of four (see Figure 4(a)). Therefore the footprint is measuring tight alignment as a much more significant case than the frequency is.

The footprint alone, however, is not effective as a statistic. One issue can be seen from Figure 4(b). The longer intervals in the stack on the right are more tightly aligned than the shorter intervals in the stack on the left. However, each stack has footprint equal to seven. To make the footprint comparable, regardless of the interval sizes involved, we must first normalize it by dividing by its expected value. $\mathcal{NF}(\mathcal{S}) = \mathcal{F}/E(\mathcal{F})$. The expected value is calculated under the usual null model. In this way the normalized footprint of a perfectly aligned stack of five intervals would be ap-

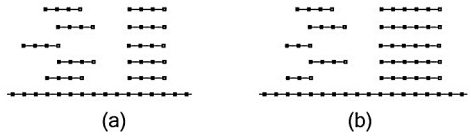

Figure 4: In example (a) on the left, there are many configurations which achieve the maximum frequency of 5, the stack on the left is one example. The stack on the right shows the only configuration which achieves a footprint of 4. In example (b) on the right there are two stacks, each with footprint equal to seven. The stack on the left has much looser alignment than the stack on the right, however, the stack on the right has much smaller normalized footprint.

proximately 1/5 for interval which are small relative to the length of the genome being considered, while for a stack of loosely aligned intervals it would be larger than 1/5 regardless of their lengths.

The normalized footprint is still not sensitive enough. Suppose there is one profile which has a very long interval. Then this interval will cause the footprint (as well as the normalized footprint) to be large regardless of the other intervals involved. We therefore do not apply the normalized footprint to each clone, but rather apply it to each stack, regardless of how many profiles are involved. We then test the value of the normalized footprint on each stack for significance with respect to the usual null model.

For any stack $\mathcal{S}$, we call m an *anchor point* of the stack if m is contained in every interval. We denote the set of all anchor points of a stack $\mathcal{S}$ by $\mathcal{S}^*$. By our definition of "stack", $\mathcal{S}^* \neq \emptyset$ for all stacks $\mathcal{S}$. Let $\mathcal{D}_n$, $n = 2, \ldots, N$ be the permutation distribution of $\mathcal{NF}(\mathcal{S})$ over all stacks $\mathcal{S}$ which consist of exactly n intervals. For each stack $\mathcal{S}$ in the unpermuted data, let $\mathcal{P}(\mathcal{S})$ be the p-value given by the tail probability in $\mathcal{D}_n$ for values less or equal to $\mathcal{NF}(\mathcal{S})$. For each location m, let

$$\mathcal{R}(m) = \min_{\substack{\text{All } \mathcal{S} \text{ such} \\ \text{that } m \in \mathcal{S}^*}} \mathcal{P}(\mathcal{S})$$

$\mathcal{R}$ provides a uniform p-value based score which makes all positions comparable, regardless of the nature of the stacks over them. We cannot use the score as a meaningful p-value however, since they are not multiple testing corrected. Therefore we must perform a second permutation calculation on the $\mathcal{R}(m)$ themselves in order to assess true significance. Since $\mathcal{R}$ is a score for each location, much as the frequency is, we assess the significance of $\mathcal{R}$ in exactly the same way.

Figure 5 shows the footprint based confidences alone. Notice how the stack at the left end is now found to be significant. In addition another stack has been revealed (marked with an "$*$") that was less apparent and which we might have missed by eye. The algorithm shows that there is significantly tighter alignment at this location than would be expected by chance.

The only serious obstacle in calculating the p-values is the impossibility, except in very simple cases, of searching, for each n, the astronomically large space of all stacks $\mathcal{S}$ which consist of exactly n intervals. This problem only arises with the footprint statistic since the frequency does not require a a subset search. We use a similar search strategy as in Grant *et al.* with an additional step to remove redundancy at each level. The approach is heuristic and first finds, for B a positive integer, the best B anchored stacks involving two intervals, "best" meaning with smallest normalized footprint after removing redundant stacks. The algorithm then extends those B stacks in all possible ways to anchored stacks involving three intervals and finds the best B of those. Those B stacks of three intervals are in turn extended to all anchored stacks of four intervals and the best B of those are determined. This is continued up to the largest possible stack and the minimum normalized footprint found at each step is recorded. These values are used in the distributions $\mathcal{NF}(\mathcal{S})$ used above. We refer to B as the *search* parameter. This method has been tested as being quite accurate in that it rarely misses the most significant stacks. This does not bias our p-values however since the same search strategy is applied to both the permuted and unpermuted data.

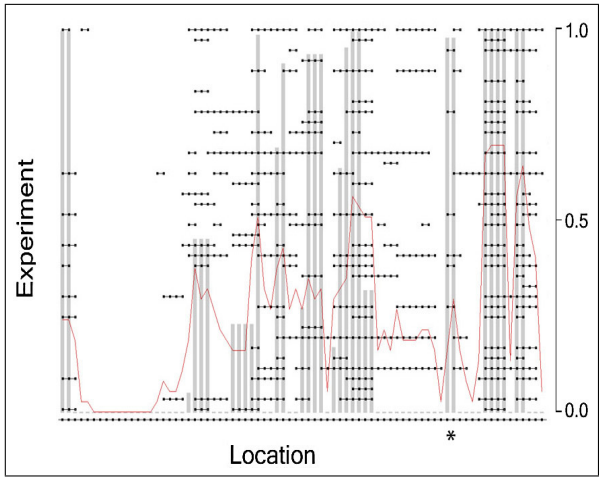

Figure 5: Display showing the footprint confidences only, indicated by grey bars.

While the footprint statistic is almost uniformly more powerful than the frequency, we find that the two statistics can convey different meaning. Therefore, we report results for both the footprint and the frequency.

References

[1] Grant G., Manduchi E., Cheung V., Ewens W. (1999) *Annals of Human Genetics* **63**, 441-454.