Poster A-8

Multiple Sample Methodology for Determining Significant Within-Class Concordance of Aberration in Array CGH Data

Authors:
Mitchell Guttman (*Computational Biology and Informatics Lab, University of Pennsylvania*)
Carolyn Mies (*Department of Pathology, University of Pennsylvania*)
Katarzyna Dudycz-Sulicz (*Department of Pathology, University of Pennsylvania*)
Sharon Diskin (*Division of Oncology, Children*)
Don Baldwin (*Microarray Facility, University of Pennsylvania*)
Gregory R. Grant (*Computational Biology and Informatics Lab, University of Pennsylvania*)

Short Abstract: Genomic duplications and deletions are typical of tumor genomes. In this study we present a method for assessing the significance of concordant genomic aberrations across multiple aCGH. By exploiting the replication across multiple samples we can detect concordant aberrations at much higher resolution than single sample approaches allow.

Long Abstract:
The occurrence of genomic aberrations such as segmental deletions and duplications are common in many types of tumor cells and are important to their nature and progression. For example, the amplification of MYC-N in Neuroblastoma is an important marker for clinical outcome [1]. Similarly, the amplification of c-ERBB2 is known to occur frequently in Ductal Carcinoma in situ (DCIS) cells and is linked to clinical outcome [2]. In general aberrations can be important in modifying the action of various pathways necessary for tumor development and survival [3]. Aberrations are relevant to other genetic diseases as well. While the copy number amplification of these regions can be high, the aberrant region can be short enough that they are difficult to detect even with high-resolution arrays looking at only one sample at a time.

Array comparative genomics hybridization (aCGH) is a high throughput high-resolution microarray based method for detecting copy number aberrations. Array CGH allows scientists to investigate whole genome aberrations [4, 5, reviewed in 6, 7]. Array CGH has proven to be a powerful tool for determining genomic aberrations of interest in various cancer types [8, 9, 10]. Similarly, this technology is quickly becoming widely used to characterize the genomic aberrations in various genetic disorders [11, 12, reviewed in 13]. In order to make use of this technology one must be able to accurately determine regions of aberration within the array. There are various approaches to this problem and in this paper we propose a novel and powerful multiple sample approach.

The promise of this technology is in part due to its ability to provide information at fine resolution. Analysis of these new kinds of data has proven challenging because the technical issues present in microarray gene expression analysis are also present in aCGH, including new CGH specific challenges. The issue that has demanded the most attention up to this point is how to transform raw microarray data into the most accurate copy-number calls at the highest resolution possible [reviewed in 14]. This is known as the segmentation problem, and

there have been numerous publications suggesting approaches to this problem on the single slide level, including Hidden Markov Models [15], Circular Binary Segmentation [16], and Wavelets [17]. The common theme of these methods is that they attempt to find aberrant segments in the genome by modeling DNA in such a way that array elements which are neighboring combine to give evidence of aberration at proximal locations.

Such single sample approaches can result in significant decrease in the native resolution of the array, which results in a loss of information, as important aberrations can be short enough to be detected by only one, or very few, array elements. Since our goal is to determine regions that are significantly aberrant across multiple samples, we have for each array element, replication across arrays that can be used to test for this effect.

In order to assess the significance of the regions in a given a set of single slide aberrations, we use the STAC algorithm [18, 19], which provides permutation based concordance p-values for each location. The null hypothesis is that, given the rate of aberration for each sample, the locations of the aberrations are independent from sample to sample. However, in reality we do not know the single slide aberration or the optimal cutoff at which to determine aberration regions for each sample.

Looking at one experiment at a time it is difficult to determine the proper criteria because we do not know what is true signal and what is noise. Looking across multiple samples it becomes clearer, however any given criterion misses information. Our approach is to measure significance across a range of criteria, weight various tests and correct the resulting confidence values. This allows us to gain power in our analysis by leveraging the multiple samples, while simultaneously controlling the family-wise error rate (FWER). We can therefore avoid the segmentation issue while simultaneously controlling the false positive rate for determining concordant aberration across multiple samples. Those locations that are concordant are assigned p-values based on the probability of the concordance occurring purely by chance.

We analyzed 20 DCIS samples to generate gain and loss confidence values for each position. We were able to detect amplification in the locus containing the ERBB2 oncogene [20]. This amplification is a small highly concordant amplification, usually limited to a region of 1-2Mb, an area covered by approximately 1-2 clones on this array and was unable to be localized by any of the available single slide methods. Using our multiple sample method we were able to identify the amplification with high confidence (p'=.011) over this small region. Furthermore, we were able to identify the single slide values at which this aberration occurs, as not all of the samples contained this aberration. We verified the presence of the aberration in all of our samples using immunohistochemistry assays and found that only those found by our method contained this amplification.

In DCIS and IDC, Chromosome 8 has been shown to contain a large deletion on the 8p arm and many gains on the 8q arm [2, 5, 8, 9, 20]. Using our method, we found that many positions on the 8p arms were in fact deleted and we were able to detect differences in regions on the arm. Rather than characterizing the entire arm as deleted, we localize and characterize the more conserved regions of deletion. Furthermore, we are able to characterize a 2Mb deletion on the 8q arm from positions 88-90Mb (p'=0.008) as well as other smaller 1Mb regions (p'=0.016-.008) of deletion which were previously

uncharacterized. Recently, a study examining chromosome 8 in IDC cell lines using high resolution Chromosome 8 specific tiling array corroborate these regions of deletion on the 8q arm [21]. The ability to map these regions in higher resolution provides us with a more accurate map of aberrations and their location along the genome.