ZINBA: A Novel Peak Detection Algorithm for Next Generation Sequencing Data

Over the past decade, microarrays have gained much popularity in their applications, allowing the measurement of biological activity on a genome wide scale. The recently developed Next Generation Sequencing (NGS) platforms, however offer better genomic coverage, dynamic range of signal, and resolution than microarrays, allowing scientists to better understand the etiology and progression of diseases such as cancer. Combined with Chromatin Immunoprecipitation (ChIP-seq), for example, it allows for accurate determination of protein-DNA interaction sites.

Computational challenge for analyzing ChIP-seq data is to isolate statistically significant regions of biological activity, which may manifest themselves as peaks or rectangle patterns in a graphical representation of ChIP-seq data (Figure 1). Specific challenges include signal variations caused by confounding factors, which can lead to higher false positive rates and lower sensitivity if ignored, especially when control data is not present. Local copy number amplification (Figure 2, left), which is often presented in tumor samples, can lead to broad and many-fold increases in signals, and can be confused with signal-enriched region of biological activity. We have developed a novel mixture regression model that is both flexible and powerful to identify different types of signals. We include confounding factors as covariates to model background and enriched regions in parallel, and thus enhance the accuracy of our method. Compared to several existing methods, we have shown that our approach isolates a higher proportion of biologically relevant regions (Figure 1, right).

Several advantages, the primary one any set of factors to model their data as some factors may be relevant in certain samples than others. Secondly, we have developed a robust approach to quantify the copy number aberrations and control their effects in ChIP-seq data analysis. This is essential for many cancer studies since copy number aberrations often occupy a significant proportion of tumor genome. Without appropriate control for copy number effects, it is not possible to accurately dissect a large proportion of signal patterns. Thirdly, we are able to capture both peak and rectangle patterns. This flexibility allow us to study many epigenetic marks such as histone modifications, which can affect the susceptibility of different diseases, including cancer. Lastly, the regression framework allows us to estimate the relationship these factors have with signal enriched and non-signal enriched regions and test for their relevance, giving for the first time a way to statistically assess...
the role of these factors with different components of signal. We have developed a software implementation of our method that automates the entire process from one’s raw data, and is freely available to the research community.