31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Analysis of Epigenetic Modifications by Next Generation Sequencing

Shoudan Liang, Yue Lu, Jaroslav Jelinek, Marcos Estecio, Hao Li, Jean-Pierre Issa

Abstract— **In plants and animals, gene expression can be altered by changes not to DNA itself but rather chemical modifications either to DNA or to histones that interact with DNA. These so called epigenetic modifications persist through cell cycle. Rapidly advancing technologies, such next generation DNA sequencing, have dramatically increased our ability to survey epigenetic markers genomewide. These techniques are revealing in great details massive epigenetic changes in cancer. Analysis of next generation sequencing data present a formidable computational challenge. We will discuss methods to address these challenges in the context of analyzing histone modifications and DNA methylation data. Several techniques useful in epigenetic data analysis will be discussed, mapping tags to reference genome incorporating all known SNPs, analysis of chIP-seq data, as well as restriction enzyme-based DNA methylation analysis.**

I. Introduction

The cost of DNA sequencing has dropped for at least a million folds in the last ten years due to the emergence of next generation sequencing technologies. In the next few years, It is expected that genomic research will transition from analog technologies such as hybridization by microarray to sequencing-based digital technologies for molecular profiling. Bioinformatic challeges are at three levels. Because the large volume of data–typically telabytes of image data is generated in each run–basic processing of data can be computationally intensive. Digital nature of the data allows extraction of additional information. The second challenge is to make the digital information useful. Furthermore, the digital technology makes it possible to design new types of assay that is not possible with analog technology.

II. Mapping Tag to Genome Taking into Account of Known SNP

We will present our work in each of the three areas. Mapping the sequence tag to reference genome is a computationally intensive problem. Current mapping software[1] treats sequencing error and natural sequence variations (single nucleotide polymorphisms or SNPs) equivalently. Therefore the current methods systematically map worse near a SNP than a region that does not contain a SNP. By including all known SNPs from dbSNP database, we have developed a method that maps all regions equally.

III. Shape Function in chIP-seq Analysis

Currently, reflecting cost advantage, next generation sequencing is most commonly used to perform ChIP-seq assay, chromotin immunoprecipitation followed by massive sequencing. This technique uses antibody pulldown to surveys the DNA binding pattern of a protein of interest in a genomewide fashion. It is often used to measure histone modifications. To determine the binding sites of for the protein of interests from the chIP-seq data, we developed a method refining on previous techniques. Current bioinformatics methods[2][3] for the chIP-seq data rely on increase in the accumulations of tags in the chIP sample as compared to the same genomic region in the control experiments. We developed a method to compute the shape of the distribution function of for the accumulation of tags mapped to the vicinity of a binding event borrowing techniques from processing of electronic signals. We show how the distribution function reduce false positives in identifying binding events.

IV. Digital Restriction Enzyme Analysis of Methylation

Next generation sequencing can be employed to develop assays and performing measurements that are only possible with digital technologies. Bioinformatics plays a useful role in designing of these assays. We provide an example in measuring DNA methylation. A new method, called digital restriction enzyme analysis of methylation, uses a combination of methylation sensitive and methylation insensitive restriction enzymes to measure methylation status. The digital information, *i.e.* the nucleotides from the tags at the beginning of each read supply methylation information: tags begin with CCGGG are from methylated DNA while ones begin with GGG are from unmethylated DNA. This method therefore measures methylations ratios on restriction enzyme cutting sites from accumulation of tags. The method provides much more accurate and more reliable readings of DNA methylation than microarray technology.

References

[1] Li, H, Ruan, J. and Durbin, R. "Mapping short DNA sequencing reads and calling variants using mapping quality scores", *Genome Res.* 2008 18: 1851-1858.

[2] Yong Zhang, Tao Liu, Clifford A Meyer, J. Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li and X Shirley Liu, "Model-based Analysis of ChIP-Seq (MACS)", *Genome Biology* 2008, 9:R137.

[3] Rozowsky J, Euskirchen G, Auerbach R, Zhang Z, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein M: PeakSeq: Systematic Scoring of ChIP-Seq Experiments Relative to Controls, *Nat Biotech.* 2009, 27:66-75.

This work was supported by grants from NIH/NCI.

Shoudan Liang is with the Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

Yue Lu, Jaroslav Jelinek, Marcos Estecio, Jean-Pierre Issa are with the Department of Leukemia, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

Hao Li is with the Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA.

978-1-4244-3296-7/09/$25.00 ©2009 IEEE