samsi
NSF ● Duke ● NCSU ● UNC ● NISS

Statistical Modeling and Analysis of Whole Genome Methylation and Chromatin Interaction (Epigenetics) Workshop
March 9-10, 2015

SPEAKER TITLES/ABSTRACTS

Deepak Nag Ayyala
Ohio State University

“Statistical Methods for Detecting Differentially Methylated Regions Based on Methylcap-Seq”

DNA methylation is an epigenetic modification known to regulate transcription and transcript splicing, and play a role in a number of other cellular mechanisms. Despite being cost-effective, capture-based methylation assays such as MethylCap-seq suffer from data resolution problem due to limited sequencing length and uncertainty of methylated CGs. To make matters worse, detection of differentially methylated regions (DMRs) needs to consider multiple potentially correlated CG sites, and furthermore, sample sizes are typically small in a genome-wide methylation study. Current practice of using univariate tests such as Fisher’s exact test ignores potential dependency between nearby CG sites, leading to erroneous Type I error rates. In this work we consider multivariate methods, as they not only allows us to model this dependency, but also results in an exponential reduction in the number of simultaneous tests needed to be performed. Specifically, we explore the applicability of several high dimensional mean vector testing procedures for detection of DMRs. These tests are an attractive alternative as they do not require any distributional assumptions including correlation structure, and are known to achieve reasonable power while controlling type I error rate even for modest sample sizes. We compare performance of the mean vector tests among themselves and to the univariate tests through simulation studies, and establish superior performance of the mean vector tests as a group. We applied these procedures to an acute myeloid leukemia(AML) dataset and an ovarian cancer dataset to detect DMRs between normal subjects and patients.

Karen Conneely
Emory University

“DNA Methylation, Gene Expression, and Aging: what can we Learn from Cross-Sectional Microarray Data?”

Epigenome-wide association studies in humans have reported thousands of age-differentially-methylated CpG sites, and recent studies show that age can be predicted from DNA methylation data with great accuracy across a wide range of cell and tissue types. However, the role of these DNA methylation changes remains unelucidated. In whole blood, the profile of associations between age and gene expression is dwarfed by the age-methylation association profile, suggesting that many of the methylation changes observed in whole blood are not directly functional, but may be marks or side effects of another process. In this work, we generate predictions for 1) competing mediation models and 2) competing evolutionary models that can potentially explain the observed relationships between DNA methylation and age. We test these predictions as a series of simple enrichment tests in integrated genomic and epigenomic data from a cross-section of whole blood samples. Our ultimate goal is to explain the widely observed pattern of age-changes in methylation. Though this goal can undoubtedly be best achieved through analysis of longitudinal and/or familial data in multiple tissues, the aim of our current work is to see how far we can get with available cross-sectional data, and to generate preliminary data and predictions for future studies.

Ming Hu
New York University School of Medicine

“A Hidden Markov Random Field Based Bayesian Method for the Detection of Long-Range Chromosomal Interactions in Hi-C Data”

Advances in chromosome conformation capture and next-generation sequencing technologies are enabling genome-wide investigation of dynamic chromatin interactions. For example, Hi-C experiments generate genome-wide contact frequencies between pairs of loci by sequencing DNA segments ligated from loci in close spatial proximity. One essential task in such studies is peak calling, that is, the identification of non-random interactions between loci from the two-dimensional contact frequency matrix. Successful fulfillment of this task has many important implications including identifying long-range interactions that assist in interpreting a sizable fraction of the results from genome-wide association studies (GWAS). The task – distinguishing biologically meaningful chromatin interactions from massive numbers of random interactions – poses great challenges both statistically and computationally. Model-based methods to address this challenge are still lacking. In particular, no statistical model exists that takes the underlying dependency structure into consideration. We propose a hidden Markov random field (HMRF) based Bayesian method to rigorously model interaction probabilities in the two-dimensional space based on the contact frequency matrix. By borrowing information from neighboring loci pairs, our method demonstrates superior reproducibility and statistical power in both simulations and real data.

Miriam Huntley
Harvard University

“How the 3D Genome Folds - Now In the Loop”

We use *in situ* Hi-C to probe the three-dimensional architecture of genomes, constructing haploid and diploid maps of ten cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1-kilobase resolution. We find that genomes are partitioned into domains (median length, 185kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs ‘facing’ one another. The inactive X chromosome splits into two massive domains and contains large loops anchored at CTCF-binding repeats.

Peng Jin
Emory University School of Medicine

"Dynamic Cytosine Modifications in Human Diseases"

Epigenetic information encoded by 5mC has a profound influence on mammalian development and affects various human diseases. Recent studies have shown that 5mC can be oxidized by TET family dioxygenases stepwise to 5-hydroxymethylcytosine (5hmC) first, then 5-formylcytosine (5fC), and finally 5-carboxylcytosine (5caC). The later oxidation products 5fC and 5caC can be recognized and excised by mammalian DNA glycosylase TDG and subsequently converted to cytosine through base excision repair (BER), resulting in active DNA demethylation in mammals. To understand the roles of these new cytosine modifications in gene regulation, we have developed a series of tools to profile the genome-wide distribution for each modification. Our analyses suggest a dynamic regulation of these new DNA modifications during neuronal differentiation, neurodevelopment and aging. I will discuss our most recent findings on the roles of these cytosine modifications in gene regulation and human diseases.

Victor Jin
UTHSCSA

"Genomic Analysis of Three-Dimensional Data Identifies Functional Enhancer-Mediated Gene Looping"

Although recent studies have comprehensively examined the relationship of higher order chromatin organization and the effects of chromatin conformation on transcriptional regulation, many important questions remain to be answered: 1) how many types of chromatin interacting loci exist across the genome besides promoter-enhancer interaction, *i.e.*, gene loop; 2) how these gene loops are associated with different histone marks, such as enhancer-mediated gene loops, repressor-associated gene loops and insulator-associated gene loops; 3) if these gene loops are functional relevant, i.e. what are expression levels of the genes in the loops; and 4) if the genes in these functional loops contribute in controlling the disease development and progression. To gain insight into chromatin structures and gene loops that may be linked to the cell type specificity and cancer diseases, in this study, we conducted a TCC analysis, a modified Hi-C protocol, on MCF7 and PANC1 cells. We used a newly re-implemented HiCPeak, initially developed in our laboratory for the analysis of Hi-C data, and identified more than 110,000 inter/intra-interacting loci pairs (ILPs) in both cancer cells. Of them, ~10-15% is involved in a promoter, i.e., P-centered long range looping events. After integrating them with histone modifications and CTCF data, we identified a subset of enhancer-associated looping genes which are enriched with common cancer oncogenesis pathways as well as cancer cell-type-specific signaling from the GO and IPA analyses. We plan further conduct RNA-seq analysis before and after inhibition of an enhancer mediator and identify globally enhancer-mediated expression of genes through chromatin looping events in the vicinity of the neighborhoods. In summary, our study provides genome-wide evidence that the interdependence of 3D chromatin architecture and transcriptional control.

Inkyung Jung
Ludwig Institute for Cancer Research

"Deciphering Dynamic Chromatin 3D Organization: from Structure to Gene Regulation"

Higher order chromatin structure is emerging as an important regulator of gene expression. However, the chromatin 3-dimensioanl dynamics in mammalian development and cell-type specification have yet to be fully explored. To address this question, we generated the first genome-wide chromatin interaction maps across a broad set of human tissues, human embryonic stem cells (hESC) and hESC-derived multiple lineages. From these data sets we uncovered extensive chromatin reorganization between distinct cell/tissue-types through alteration of active and inactive chromosomal compartment. Furthermore we delineate allelic chromatin interactions, chromatin modifications, and transcriptomes amongst a broad set of human tissues and cell lines to explore allele biased gene expression and the mechanisms underlying it, enabled by a chromosome-spanning haplotype reconstruction strategy. The haplotype-resolved transcriptomes reveal extensive allelic biases in the transcription of human genes, which appear to be primarily driven by individual specific genetic variations. The extensive allele biased gene expression correlates with allele biased chromatin states of linked promoters and enhancers and allelically biased higher-order chromatin structures. The integrative analyses of chromatin interaction maps and haplotype-resolved epigenome and transcriptome data sets sheds light on the regulatory effect of chromatin dynamics resulting from global chromatin 3D reorganization and long-range gene regulation controlled by local higher-order chromatin structure.

Eric Lock
University of Minnesota

"Bayesian Screening for Group Differences in Methylation Array Data"

In modern biomedical research, it is common to screen for differences between groups in many variables that are measured using the same technology. Motivated by DNA methylation data, this talk focuses on screening for equality of group distributions for many variables with shared distributional features such as common support, common modes and common patterns of skewness. We propose a Bayesian nonparametric testing methodology, which improves performance by borrowing information across the different variables and groups through shared kernels and a common probability of group differences. The inclusion of shared kernels in a finite mixture, with Dirichlet priors on the different weight vectors, leads to a simple framework for testing and we describe an implementation that scales well for high-dimensional data. We provide some theoretical results, compare with existing frequentist and Bayesian nonparametric testing methods, and describe an application to breast cancer methylation data from the Cancer Genome Atlas.

Oswaldo A. Lozoya
National Institute of Environmental Health Sciences, NIH

"The Impact of Mitochondrial Dysfunction on the Epigenome"

The classical scientific consensus maintains that the primary physiological role of mitochondria in eukaryotic cells is ATP production through oxidative phosphorylation (OXPHOS). During OXPHOS, reducing equivalents from the TCA cycle (e.g. NADH, FADH) pair with membrane-bound multimeric complexes in mitochondria, prompting the electron transport chain (ETC) that drives ATP synthesis; metabolites from the TCA cycle also elicit rate-limited enzymatic functions throughout the cell, including transcriptional control and regulation of chromatin conformation. Therefore, these physiological functions depend on an effective crosstalk between the nuclear and mitochondrial genome – each encoding different components of the ETC machinery. We are interested in addressing the potential role of mitochondria in modulating epigenetic regulation of genes within the nucleus. The hypothesis driving our experiments is that changes in mitochondrial metabolism can alter epigenetic reactions in the nucleus and as such regulate gene expression of both DNA coding and non-coding regions. Consistent with this hypothesis, our preliminary data indicate that loss of mitochondrial oxidative phosphorylation induces decrease in histone acetyltransferase (HAT) activity and histone acetylation, and increase in repetitive element expression. To further test our hypothesis, we are currently determining the mechanism through which mitochondrial function impacts HAT activity and histone acetylation (such as measuring different metabolites). We will also perform bioinformatics analysis of gene expression and nucleosome regulation, including RNA-seq and ChIP-seq analyses. Together, our experiments will allow us to determine whether broad rate-limiting OXPHOS derivatives indeed change gene expression by altering the epigenetic landscape of the cells at a genome-wide scale and, if so, where on the genome.

Co-authors: Janine H. Santos, Richard P. Woychik

Liang Niu
University of Cincinnati

“Statistical Modeling and Analysis of ChIA -PET Data”

Chromatin interactions mediated by a particular protein are of interest for studying gene regulation. A recent molecular technique, Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), that uses chromatin immunoprecipitation (ChIP) and high throughput paired-end sequencing, is able to detect such chromatin interactions genomewide. However, ChIA-PET may generate noise (i.e., pairings of DNA fragments by random chance) in addition to true signal (i.e., pairings of DNA fragments by interactions). In this talk, we present MC_DIST based on a Bayesian mixture modeling framework to identify true chromatin interactions from ChIA-PET count data (counts of DNA fragment pairs). A simulation study showed that MC_DIST outperforms the previously proposed hypergeometric model in terms of both power and type I error rate. A real data study showed that MC_DIST may identify potential chromatin interactions between protein binding sites and gene promoters that may be missed by the hypergeometric model.

Yongseok Park
University of Pittsburgh

“Statistical Analysis of DNA Methylation using Within-Fragment Information”

DNA methylation plays important roles in genomic imprinting, genome stability and regulation of gene expression. Next-generation sequencing following bisulfite treatment of DNA (BS-seq) is gold standard to study DNA methylation. The application of whole genome bisulfite sequencing has created enormous large scale modeling and data mining challenges. Such “big” data also provide great opportunities for researchers to develop statistical methods to decipher underlying biological information. Unlike DNA sequencing, DNA methylation patterns are different within cells even from the same normal tissues. Therefore, studying methylation patterns using sequencing fragments may provide further insights of biological processes. Furthermore, current statistical methods rely on the sufficient sequencing depths (i.e. the number of coverage reads) for a given CpG site. Due to uneven distribution of sequencing reads and high cost of BS-seq, lacking of sequencing depths is a common problem of current BS-seq experiment. In this talk, we will also discuss a direction of improving statistical methods using within-fragment information.

Maureen Sartor
University of Michigan

“Analysis Tool for Combined DNA Methylation and 5-Hydroxymethylcytosine Data”

The wide application of reduced representation bisulfite sequencing (RRBS) and whole genome bisulfite sequencing (bis-seq) now allows the study of genome-wide DNA methylation at single CpG site resolution. In addition, 5-hydroxymethylcytosine (5hmC), once thought of as simply a transient step in the demethylation process, is now known to be functionally important in distinguishing tissues, and is aberrant with functional consequences in multiple cancers. Because bisulfite sequencing cannot distinguish DNA methylation (5mC) from 5hmC, new approaches such as hmeDIP-seq have been developed for genome-wide assessment of 5hmC.

Currently, there is a lack of software for analyzing and interpreting these data types, especially in conjunction with each other and with gene expression data. Recently, we developed a method for the analysis and annotation of RRBS or bis-seq data, called methylSig. At its core, methylSig uses a beta-binomial model to test for differentially methylated CpG sites or regions, and can incorporate local information to improve group methylation level and/or variance estimation for experiments with small sample size. We also recently developed a method for consistent or differential peak detection in ChIP-seq data, which can be used for whole genome DNA methylation and 5hmC pull-down methods, such as meDIP-seq and hmeDIP-seq. This method, called PePr, uses a negative binomial model along with several pre- and post-processing steps, and also incorporates local chromosomal information to improve variance estimation across samples. We are now building upon these methods to develop software within the Galaxy Project that will allow integrative analysis, visualization, and annotation of 5mC and 5hmC data, and integration with gene expression results.

Robert Schmitz
University of Georgia

“Challenges and Biases Associated with Whole-Genome Bisulfilte Sequencing Data”

The development of whole-genome bisulfite sequencing (WGBS) has resulted in a number of exciting discoveries about the role of DNA methylation leading to a plethora of novel testable hypotheses. Methods for constructing sodium bisulfite-converted and amplified libraries have recently advanced to the point that the bottleneck for experiments that use WGBS has shifted to data analysis and interpretation. Here I will present empirical evidence for an over-representation of reads from ethylated DNA in WGBS. This enrichment for methylated DNA is exacerbated by higher cycles of PCR and is influenced by the type of uracil-insensitive DNA polymerase used for amplifying the sequencing library. Future efforts to computationally correct for this enrichment bias will be essential to increasing the accuracy of determining methylation levels for individual cytosines. It is especially critical for studies that seek to accurately quantify DNA methylation levels in populations that may segregate for allelic DNA methylation states.

Hao Wu
Emory University

“Differential Methylation Analysis from Whole-Genome Bisulfite Sequencing: a Matter of Spatial Correlation, Coverage Depth and Biological Variance”

DNA methylation is an important epigenetic modification involved in many biological processes and diseases. Recent developments in whole genome bisulfite sequencing (WGBS) technology have enabled genome-wide measurements of DNA methylation at single base pair resolution. Many experiments have been conducted to compare DNA methylation profiles from different biological backgrounds, with the goal of identifying differentially methylated regions (DMRs) of the genome. However, existing methods for DMR detection may lose accuracy and power by failing to account for several important characteristics of the data.

We develop a novel statistical method for detecting DMRs from WGBS data. The method is based on a hierarchical model that accounts for the spatial correlation of methylation levels, coverage depth, and biological variations among replicates. An important feature of the method is that the biological variations can be estimated using information from neighboring CG sites even when there is only one replicate per biological condition. Simulations and the analyses of several real datasets demonstrate that the proposed method performs favorably compared with existing methods. The proposed method is implemented in the Bioconductor package DSS.

Pearlly Yan
Ohio State University Comprehensive Cancer Center

“DNA Methylation Profiling in Cancer: Perspectives from a Genomics/Computation Group”

DNA methylation profiling uncovers heritable changes that do not involve a change in the DNA sequence. DNA methylation mark is stable and relatively easy to assay. With the recent advances in the detection and the quantification of DNA methylation using next-generation sequencing (NGS), researchers can now examine the effect of promoter- or gene body methylation on gene expression, integrate DNA methylation and histone modifications to elucidate transcription factor binding events, and identify methylation-based biomarkers for disease diagnosis and prognosis on an unprecedented scale. Herein, I describe an in-house QC tool to assess methylation data quality and a custom computation approach to compute nucleotide-resolution methylation values from methylome analysis. PrEMeR-CG (Probabilistic Extension of Methylated Reads at CpG resolution) is developed to harness the implicit information associated with MethylCap-seq (*in vitro* capture of methylated DNA followed by sequencing of enriched fragments) library fragment profiles to infer nucleotide-resolution methylation values in addition to read counts data. This method was used to produce the methylation data used in tandem with gene expression to produce a novel, clinically significant gene signature in Acute Myeloid Leukemia.

Michael Zhang
University of Texas, Dallas

“Computational Advances in ChIP-seq and ChIA-PET Data Analysis”

ChIP-seq and ChIA-PET have generated vast amount of genome-wide chromatin localization and interaction data at high throughput and resolution. Better analysis of such data can provide more and deeper biological insights that will help us further understanding function and mechanism of chromatin

states and epigenetic gene regulation. I will introduce our recent development of some novel computational methods in ChIP-seq and ChIA-PET Data Analysis.

Chenchen Zou
Jackson Laboratory for Genomic Medicine

“Multi-track Structure Inference Model for Genome-wide Chromatin Conformation-capturing Data”

Genome-wide conformation-capturing experiments like Hi-C and ChIA-PET reveal consensus chromatin folding features that are closely related to biological function and a growing body of researches have pointed to the existence of conserved genome structures. We have developed a statistical model that can jointly fit multiple contact matrices from different experiments to infer 3D chromatin structure. Our model treats chromatin as a hidden Markov chain under contact matrices and frozen the optimal structure under hamilton dynamics. We tested our model under various resolutions in four cell lines’ Hi-C data. All of their predicted structures fit the data better and have higher correlations with ChIA-PET or FISH than BACH and ChromSDE. Our model outperforms other existing methods in both simulation and real data, especially in sparse data.