META-ANALYSIS FRAMEWORK FOR PEAK CALLING BY COMBINING MULTIPLE CHIP-SEQ ALGORITHMS AND GENE CLUSTERING BY COMBINING MULTIPLE TRANSCRIPTOMIC STUDIES

by

Rui Chen

BS, Nanjing University of Science and Technology, China, 2010

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH GRADUATE
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Rui Chen

It was defended on

December 17th 2014

and approved by

Dissertation Advisor
George C. Tseng, ScD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Members
Gong Tang, PhD
Associate Professor Department of Biostatistics Graduate School of Public Health University of Pittsburgh

Michael Barmada, PhD
Associate Professor Department of Human Genetics Graduate School of Public Health University of Pittsburgh

Yongseok Park, PhD
Assistant Professor Department of Biostatistics Graduate School of Public Health University of Pittsburgh

META-ANALYSIS FRAMEWORK FOR PEAK CALLING BY COMBINING MULTIPLE CHIP-SEQ ALGORITHMS AND GENE CLUSTERING BY COMBINING MULTIPLE TRANSCRIPTOMIC STUDIES

Rui Chen, PhD

University of Pittsburgh, 2015

Abstract

With the availability of large amount of genomics studies, integrating information from multiple sources improves knowledge discovery. To address the complexity of genome and numerous genetic features, meta-analysis that aggregate information achieves higher statistical power for the measure of interest, and identify patterns among study results, sources of disagreement among those results.

As Next-Generation Sequencing (NGS) technologies are becoming affordable and can provide per-base resolution, NGS data serves as an appealing tool to analyze genomic features. Among various applications of NGS technologies, chromatin immunoprecipitation followed by high-throughput sequencing(ChIP-seq) is primarily used to provide quantitative, genome-wide mapping of target protein and DNA interaction events. Signal peak calling algorithms identified target regions of interest enriched *in vitro*. Despite the existing programs for previous ChIP-Chip platforms, peak calling of putative protein binding sites from large, sequencing based data-sets presents a bioinformatic challenge that has required considerable computational innovation. Popular peak calling algorithms, such as MACS, SPP, CisGenome, SISSRs, USeq, and PeakSeq, are widely applied but each of them has different emphasis on sensitivity, specificity or different size and shape selection of peaks. In the first project of this dissertation, we propose a meta-analysis framework, ChIP-MetaCaller, to combine multiple top-performing algorithms to identify and reprioritize the peaks. We

provide a forward selection algorithm to decide best combination of algorithms' output to perform meta-analysis and showed that the result improves motif enrichment and sensitivity. The results are more trackable by biologists for further validation and hypothesis generation.

The mechanisms of complex diseases like cancers involve changes in multiple genes, each conferring small and incremental risk that potentially converge in deregulated biological pathways, cellular functions and local circuit changes. To understand this complex network requires discovery of co-expression gene modules. Literature shows using meta-analysis can improve performance of identifying these modules from machine learning techniques in some pilot studies. In the second project of this dissertation, we proposed approach which is based on the clustering results of each individual study. Combining standardized distances from genes to the medoids lead to an integrated distance matrix and perform the meta-clustering. We compared the performance of proposed approach and Meta Clustering combining distance under three simulation settings and three real data sets and provide guidance for practitioners.

Two projects included in this dissertation tackles different biological questions based on genomics data. Both of them improve performance from existing methods by information integration applying meta-analysis frameworks, and provide comprehensive biomarker detection.This work could improve public health by providing more effective methodologies for biomarker detection in the integration of multiple genomic studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 HIGH-THROUGHPUT EXPERIMENTAL DATA

Genomics is a recent term that describes the study of all of a person's genes (the genome), including interactions of genes with each other and with the person's environment. Genomics includes the scientific study of complex diseases such as heart disease, asthma, diabetes, and cancers because these diseases are typically caused more by a combination of genetic and environmental factors than by individual genes. Genomics is offering new possibilities for therapies and treatments for many complex diseases, as well as new diagnostic methods. Due to the tremendous size of genome, exploration of whole genome requires high through-put experimental data. The generation of these mainly rely on two platforms: microarray and massively parallel sequencing.

1.1.1 Microarray

Microarray technology allows measurement of the levels of thousands of different RNA or DNA molecules at a given point in the life of an organism, tissue or cell. With the comparison of levels of RNA or DNA molecules can be used to decipher the complex processes going on simultaneously. Comparison between different genomic information under different biological conditions can yield vital information related to diseases. Microarray has been widely applied to almost all fields of biological researches, as arrays become more easily applicable (i.e. cheaper and reproducible). With the different designs of gene probes or experiment, microarray technology can be widely applied to various purposes of biomarkers detection including, profiling of expression levels, SNP detection, copy number alternation detection,

transcription factors binding detection, etc. With decades of improvement of both the technology and data analysis methods, microarray has been recognized as a mature platform to provide reliable high throughput data for genomics researchers.

1.1.1.1 Microarray General Procedures Based on hybridization of two DNA strands (Southern blotting) or RNA strand with DNA strand(Northern blotting), microarray, a multiplex "lab-on-the-chip", is developed to capture large amount of DNA samples. It is a 2D array on a solid substrate (usually a glass slide or silicon thin-film cell) that assays large amounts of biological material using high-throughput screening miniaturized, multiplexed and parallel processing detection methods. With the known bases of the material on the chip, each DNA spot contains picomoles (10^{-12} moles) of a specific DNA sequence, known as probes (or reporters or oligos). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA (also called anti-sense RNA) sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target.

1.1.1.2 Array Types With the rapid development of microarray technology and different research purposes, biotechnology companies have development various microarray product including, DNA microarray, (such as cDNA microarrays, oligonucleotide microarrays, BAC microarrays and SNP microarrays), MMChips(for surveillance of microRNA populations), Protein microarrays, Peptide microarrays(for detailed analyses or optimization of protein-protein interactions), Tissue microarrays, Cellular microarrays (also called transfection microarrays), Chemical compound microarrays, Antibody microarrays, Carbohydrate arrays (glycoarrays), Phenotype microarrays and etc.

Major application of DNA microarray can be further categorized by detection purposes.

- Expression profiling: In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression.

- SNP detection: Identifying single nucleotide polymorphism among alleles within or between populations. Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.
- ChIP on chip: DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome.

With the well-established statistical model to analyze and mature technology, the application listed above is still popular in majority biological lab. Other application including alternative splicing and fusion gene detection can be accomplished by specific design of microarray experiment and chips as well. However, as the emerging of NGS technology, the limitation of these application with microarray is diminished the occupation of market.

1.1.1.3 Data Analysis Proper data processing and quality control are critical to the validity and interpretability of microarray analysis. As majority analysis is aiming at comparison of difference under conditions, normalization is required to standardize data to focus on biologically relevant changes. There are many sources of systematic variation in microarray experiments that affect the measured gene expression levels such as dye bias, heat and light sensitivity, efficiency of dye incorporation, differences in the labeled cDNA hybridization conditions, scanning conditions, and unequal quantities of starting RNA, etc. Normalization is an important step in adjusting the data set for technical variation and removing relative abundance of gene expression profiles. The basic idea behind all the normalization methods is that the expected mean intensity ratio between the two channels should be one. If the observed mean intensity ratio deviates from one, the data is mathematically processed in such a way that the final observed mean intensity ratio becomes one. With the mean intensity ratio adjusted to one, the distribution of the gene expression is centered so that genuine differentials can be identified.

1.1.2 Next-Generation Sequencing

As automated Sanger sequencing method is recognized as the first generation of sequencing technology, massively parallel sequencing techniques are referred to next-generation sequencing (NGS) [20]. According to different procedure of template preparation, sequencing can be applied to various clinical researches. For example, bisulfite treated DNA sequencing can be used to detect methylation sites; chromatin immunoprecipitation captured reads can be applied to transcription factor and histone modification; complementary DNA sequencing are applied to quantify gene expression profiles, and etc. Compared with microarrays, NGS can identify and quantify genomic features without prior knowledge of genetic event locations but scanning whole genome of organisms. This unique feature provides flexibility of genetic study design.

1.1.2.1 NGS General Procedures NGS platforms produce enormous DNA molecular templates and sequence the ends of these templates simultaneously to achieve collecting hundreds of thousands base pairs information with a relatively low cost within a short time period. However, billions of sequenced short reads and relative base calling quality (in fact, total count and length of reads depend on the prepared library sizes of sample genome and sequencing platform) are written into text files with size can be up to terabytes with unknown genome location. To determine these numerous reads genome location, there are two general approaches: *de novo* assembly and alignment. *De novo* assembly relies on only reads pipe-up information to build up genome information. Due to the complexity of genome and algorithm, *de novo* assembly involves extremely expensive computation and is usually only applied to explore unknown genome or discover innovational genome structure. The most common used algorithm performing *de novo* assembly is *de brujin* graph [4]. Alignment is referring to algorithms that map short reads towards a known genome by allowing certain mismatches due to the individual biological variation. Among existing alignment software, the most popular ones of them, such as BWA and Bowtie, are based on Burrows-Wheelers transform algorithm [2]. For the background information, the section below gave a general procedure of NGS data preparation and analysis work flow.

1.1.2.2 Template Preparation Template is referred to the recombinant DNA molecule made up of a known region, usually an adapter sequence to which a universal primer can bind, and a target sequence which is an unknown portion to be sequenced. The emergence of shot-gun method during the Human Genome Project suggested to randomly breaking genomic DNA into smaller sized fragments. These fragments can be separated into two categories according to sequencers, *template fragments* which sequencer only sequence one side of fragments (also known as single-end fragments), *mate-pair templates*, also known as paired-end fragments, which sequencer sequence both ends of fragments and using the distance between two ends to provide a more accurate alignment in later steps. Then these reads is attached or immobilized to a solid surface of supports to allow thousands to billions of sequencing reactions to be performed simultaneously.

Generally there are two different ways to prepare templates according to whether amplifying templates.The first approach is *clonally amplified templates* which uses emulsion Polymerase Chain Reaction (emPCR) [6] or solid-phase amplification [11]. This approach can avoid bacterial cloning, however, require a large amount of genomic DNA material [20]. The second approach is *single-molecule templates* which is more straightforward and require less starting materials. Moreover, this approach avoids PCR, which creates mutations. Quantitative applications, such as RNA-seq [29], perform more effectively with non-amplified template source, otherwise, removal of duplicate reads is required in a later step of data analysis.

1.1.2.3 Sequencing and Imaging According to different sequencing platforms, there are four different imaging methods: Illumina/Solexa and Helicos BioScences uses *Cyclic Reversible Terminators*(CRT); Life/APG performs *Sequencing by Ligation*; Roche/454 applies *Single-nucleotitde addition: Pyrosequencing*; and Pacific Biosciences uses *Real-time sequencing*. There are essential differences in sequencing clonally amplified and single-molecule templates. Clonal amplification results in a population of identical templates, each of which has undergone the sequencing reaction. Upon imaging, the observed signal is a consensus of the nucleotides or probes added to the identical templates for a given cycle. This places a greater demand on the efficiency of the addition process, and incomplete extension of the template

ensemble results in lagging-strand dephasing. The addition of multiple nucleotides or probes can also occur in a given cycle, resulting in leading-strand dephasing. Signal dephasing increases fluorescence noise, causing base-calling errors and shorter reads [9]. Because dephasing is not an issue with single-molecule templates, the requirement for cycle efficiency is relaxed. Single molecules, however, are susceptible to multiple nucleotide or probe additions in any given cycle. Here, deletion errors will occur owing to quenching effects between adjacent dye molecules or no signal will be detected because of the incorporation of dark nucleotides or probes.

1.1.2.4 Genome Alignment and Assembly After sequencing and imaging, the sequencer machine generates numerous number of base pair information and corresponding per base calling quality. The general format is FastQ, sometime with a suffix .fastq or .fq. The nucleotide information is recorded as either A,T,C,G, or N (which means unknown which nucleotide). Corresponding quality score is commonly recorded by ACSII transformed phred score. Different platforms produces different scores and accuracy estimates.

To locate the short reads genomic location requires either align to a known reference sequence or assembled *de novo*. The decision depends on not only the study purposes but also computational workload and cost as mentioned above.

The alignment reference genome is sequenced in former study and can be download from multiple databases, such UCSC (University of California at Santa Cruz), NCBI (National Center for Biotechnology Information), and Ensembl (European Bioinformatics Institute and the Wellcome Trust Sanger Institute).

De novo assembly is subjected to explore and discover genomic regions that do not exist in the reference genome. Structural variants are shown highly correlated with cancers. Cancer patients' genome may be highly disordered where *de novo* assembly has been recommended compared to alignment. However, due to the complexity of its algorithm, assemblies have been reported mostly for bacterial genomes and mammalian bacterial artificial chromosomes [30], [32], but still challenging to apply in human genomes.

1.1.2.5 Downstream Applications of NGS

The purpose of generating these large numbers of reads with NGS is to explore genomic features. The applications of NGS data include small region variants or structural variants discovery, quantifying transcriptomes regulation, alternative splicing discovery, epigenetic biomarkers profiling and etc. Variants discovery is achieved by resequencing target regions or whole genome to address differences with reference genome or genome from matched normal samples usually based on Exome-seq or Whole Genome Sequencing. Transcription level studies are commonly based on RNA-seq data. Methylation site detection can be discovered with sequencing a bisulfite treated DNA-seq data while using a converted reference. Transcription factor and histone modifications biomarkers are analyzed based on ChIP-seq data.

1.2 "OMICS" DATA INTEGRATION

The term data integration refers to the situation where, for a given system, multiple sources and types of data are available and we want to study them integrative to improve knowledge discovery. The integration of different sources of data can improve the limitation of sample size from single studies; the integration of different types can help researchers to better understand the influence of genetic effects. For example, in the study of prostate cancer [31], we have two datasets describing the system, one containing information about gene expression at the mRNA level and the other describing the CpG DNA methylation profile. In several studies [19] [24] where gene expression and DNA methylation data were available, the genome-wide relationships between DNA methylation and gene expression have been investigated in order to infer generic rules to questions such as: "Does DNA methylation regulation occurs at CpG islands and/or shores?", or "How does DNA methylation in promoters/gene-bodies/enhancers regulate gene expression?". These kinds of analyses have advanced our understanding of gene regulation by providing "generic rules yet with several exceptions" that associate epigenetic modifications with transcription.

1.2.1 Microarray Meta-analysis

As the microarray technology has been widely applied to almost all biomedical fields, generation of various kinds of this high-throughtput technology genomic data is common for biological research. Tremendous amount of studies based on microarray platform has been conducted. There are several pulibcly available data depositories collected majority transcriptome data from microarray experiments generated from past decade, such as Gene Expression Omnibus [21] and ArrayExpress [8] . However, the ability to manage and compare the resulting data can be problematic. It is very common that transcriptomic studies are focused on same or related diseases. Combining information from such studies can achieve the goal of increase sample size and further lead to more desirable statistical power. Classical meta-analysis in statistics that combining multiple studies which of similar hypothesis can be modified and applied to the genomic information integration.

General genomic data integration can be categorized into two major settings according to the term used in the review paper (Tseng et al 2012) [28]. Genomic information integration that combine result from multiple transcriptomic studies are termed as horizontal genomic meta-analysis. On the other hand, combining multiple sources of genomics information is termed as vertical genomic meta-analysis. These kinds of multi-dimentional integration usually include but not restrict to, transcriptome expression profile, genotypes, and copy number variation of DNA fragments, methylation, microRNA and phenotype. In this dissertation, we will meta-analysis in gene co-expression module detection in Chapter 3 and we will emphasize more on horizontal genomic meta-analysis.

Among horizontal genomic meta-analysis, there are multiple different purposes of integration of information. Differentially Expressed (DE) gene detection is a commonly used downstream analysis. Meta-analysis frameworks have been widely applied to combine results of DE genes detection, and achieved a significantly improve of sensitivity. Pathway analysis is another statistical tool to infer correlation of DE evidence in the data with pathway knowledge from established databases. Shen and Tseng, et al. [26] developed a systematic framework of Meta-Analysis for Pathway Enrichment (MAPE) which integrates information at gene level, at pathway level and a hybrid of the two. Another analysis commonly applied

to microarray data is prediction analysis, also known as classification analysis or supervised machine learning. The purpose of this analysis is to predict category membership of new observation based on the discriminant model build with training data set. Combining of this kind of analysis is aiming to improve discrimination with two or more study populations. Network and gene co-expression module detection is another application of transcriptomic data analysis. When multiple transcriptomic studies are combined, most methods have been developed to improve differential analysis (candidate marker detection) and pathway analysis. These methods mostly extend from traditional meta-analysis by combination effect sizes or p-values of multiple studies to a genome-wide scale [28]. But in the field of co-expression module detection, no systematic study of integrative methods for combining multiple transcriptomic studies is available, to the best of our knowledge. This leads to the development of methods in Chapter 3. Details of this type of analysis will be further described in section 3.2.

1.3 CHIP-SEQ TECHNIQUE FOR GENOME-WIDE MAPPING OF PROTEIN-DNA INTERACTIONS

Biomarkers of protein-DNA interactions have been proved essential for understanding transciptional regulation [10]. The combination of nucleosome positioning and dynamic modification of DNA and histones play an important role in gene regulation and lead to differentiation [23]. ChIP (Chromatin ImmunoPrecipitation) is a powerful method to selectively enrich for DNA sequences bound by a particular protein in living cells. To identify all the enriched DNA sequences, downstream techniques including old techniques tilling array and massively parrallel sequencing can be applied. The ChIP process enriches specific crosslinked DNA-protein complexes using an antibody against the protein of interest. Oligonucleotide adaptors are then added to the small stretches of DNA that were bound to the protein of interest to enable massively parallel sequencing. Antibody captured DNA sample sequencing result compared with no antibody effect chromatin input is used to decide reads signal enrichment region. The summit of the signal peak is highly likely the DNA-protein binding

site. Compared to ChIP-chip, ChIP-seq technology provides multiple outstanding features. The resolution of ChIP-seq can be specified to single nucleotide while ChIP-chip relies on the size of array probes usually 16-100bps. ChIP-chip's coverage of genome is limited by sequences on the array, and repetitive regions are usually masked out. The genome coverage of ChIP-seq depends on the mappability of reads and repetitive regions can be covered. The coverage can be increased by include more lane run of sequencer or use longer reads, however, these properties requires higher cost. For example, Illumina Hi-seq 2000 platform generate approximately 4 million reads with length of 100bps with cost of $1000-$2000 per Illumina lane and depends on the design, one lane can process multiple samples.

Nearly all ChIP-Seq data so far have been generated on the Illumina Genome Analyzer, although other platforms such as Applied Biosystems SOLiD and the Helicos platform are also available for ChIP-Seq. The Illumina and the SOLiD platforms currently generate 100400 million reads in a single run, typically with 60 to 80% of reads that can be aligned uniquely to the genome.

Downstream data analysis of sequencing data requires careful statistical modeling. Algorithms that convert reads aggregation into peaks signals are recognized as peak callers. Various callers use Poisson or negative binomial distribution to address the count data of reads. The output of different callers are vary on reporting different lengths of binding regions or different measurement of significance levels, and raises difficulty to compare results from each others. In Chapter 2, we proposed a ChIP-MetaCaller framework provides practical output combining result from multiple callers.

1.4 GENE CO-EXPRESSION MODULE

Global gene expression profile studies using microarray are widely applied. Gene expression profile can be used to identify gene co-expression modules. Sets of genes expression patterns which are highly correlated across samples may reflect sharing similar function and common regulatory pathways [18], [13]. Literatures has shown using gene co-expression analysis to build up networks, identify gene communities, and gene shared functions [5], [7].

Gene co-expression networks offer genome wide information associated with specific molecular mechanisms in diseases. In Gaiteri et al. (2014) [12], they mentioned multiple co-expression network analyses in mental diseases studies. The differential co-expression analysis insights into testable predictions due to gene-gene co-expression correlations relate to core features of brain activity and structure, including spatial patterns, inter-tissue communication, epigenetic changes and other non-coding features of regulatory networks.

Multiple factors can influence the co-expression pattern of gene-gene links, such as transcription factor targets, microRNA targets, inter-tissue communications and etc. However, technical, such as batch effect, and cell-type variability can also generate similar pattern of gene co-expression which is potential confounders for the biologically meaningful regulation pattern.

As gradually increasing numbers of gene co-expression studies, a framework of combining single studies is proposed by Chang et al. (2014) [3]. In Chapter 3, we will propose a meta-clustering method to combine several gene expression profile studies, and use Pearson correlation to measure similarity between genes. Based on averaging distance matrices, which are calculated by one minus correlation matrices, to construct a combined distance of studies and applied penalized K-Medoids clustering algorithm [27] to determine the gene co-expression modules. In Chapter 3, we proposed another approach which combine studies at clusters levels and compare with Changs methods.

1.4.1 Network and Co-expression Analysis

Most complex diseases include dysfunction at the levels of genes, cells, tissues, organ regions and feedback between these networks at multiple biological scales. The overlapping effect of all these levels may lead to pathogenic mechanism of diseases obscure when taking measurement at single level. Gene co-expression network analysis provides genome-scale information and also has the potential to highlight specific molecular mechanisms in disease particularly if the biophysical basis of co-expression is integrated into network analysis and if researchers examine network properties beyond modules and hubs. In a recent review paper, Gaiteri and Ying et al. 2014 [12] summarized the multi-scale mapping of gene expression traits

and co-expression networks into the following steps: 1) Globel co-expression networks; 2) Network decomposition to modules; 3) Within module disease traits; 4) Local co-expression; 5) Additional molecular networks. The building of whole networks is extremely complicated and including computational and biological efforts. In this dissertation, we focused on the first two steps.

When the mRNA expression of two or more genes is correlated across multiple samples, these genes are recognized to be coexpressed. These coexpression links are generally inferred from large microarray or RNA sequencing studies with no reference to the mechanisms behind these correlations. Studies in multiples species, tissues and platforms have shown that coexpressed genes tend to be functionally related (Obayashi et al. 2008; Oldham et al 2006). Gene sets that are densely interconnected by coexpression links within the global gene network are commonly known as clusters or modules [12]. If a significant fraction of genes in a module relate to a gene ontology category or canonical pathway, through guilt-by-association (GBA) the remaining genes in the module are assumed to be related to that function (Gillis and Pavlidis 2012, Wolfe et al. 2005). Understanding the mechanisms of gene regulation during certain biological condition is one of the most difficult problems among oncologists because this regulation is likely comprised of complex genetic interactions. Building network of gene sets are typically based on detection of gene-gene co-expression modules. The underlying assumption is that the magnitude of co-expression between any pair of genes is associated with greater likelihood that two genes interact.

1.4.2 Clustering Analysis

Unsupervised machine learning (also known as cluster analysis) is a common method to discover gene co-expression modules by assign genes into groups under a predefined distance measure when these modules have unknown labels. The application of clustering analysis in categorizing genomic data is not restricted to identify co-expression module. Complex diseases may be caused by different mechanisms. Clustering analysis is applied to identify groups of patients into subtypes of diseases according to expression profiles. Representative diseases include leukemia (Golub et al., 1999), lymphoma (Rosenwald et al., 2002),

glioblastoma (Parsons et al., 2008; Verhaak et al., 2010), breast cancer (Lehmann et al., 2011; Parker et al., 2009), colorectal cancer (Sadanandam et al., 2013) and ovarian cancer (Tothill et al., 2008). In this dissertation, we focus our discussion on cluster analysis on high-throughput biological data for identifying gene modules, especially in the analysis of microarray expression profiles.

There are large variety of clustering methods, including , hierarchical clustering (Eisen et al 1998), K-means and its variants, mixture Gaussian model-based clustering (Yeung et al 2001), graph-theoretical method (Sharan et al. 2003), and tight clustering (Tseng 2005). Comparative study for gene clustering in expression profiles (Thalamuthu et al., 2006) suggests that clustering methods allowing scattered objects not being clustered, with explicit or implicit model assumptions, and with resampling evaluations seem to perform better. Penalized Weighted K-mediods (Tseng 2007) extended from K-mediods by adding penalty term to allow a set of scattered objects without being clustered. Weights are introduced to account for prior information of preferred or prohibited cluster patterns to be identified as well. In Chapter 3, we extend the Penalized Partition Around Mediods(PPAM) with a meta-analysis framework to integrate multiple study targeting the same disease or biological conditions to detect potential gene co-expression modules.

2.0 CHIP-METACALLER: AN ASSEMBLY METHOD TO COMBINE MULTIPLE CHIP-SEQ PEAK CALLERS TO IDENTIFY AND REPRIORITIZE THE PEAKS

2.1 BACKGROUND

Genome-wide ChIP experiment is used to detect binding site of a target protein's DNA interaction events. The procedure detecting these interactions is based on searching maximal signal-to-noise ratio through whole genome of organism.(Landt et. al. 2012). Target DNA cross-linked protein antigen which captured by beads attached antibody can be purified by precipitation. Unlinked DNA fragments are later enriched and mapped to genome locations indicates the interaction event regions. In ChIP-chip analysis, enriched DNA fragments is fluorescently labeled and hybridized to microarray chip. Thus, the detection of signals highly depend on design of probes on the chip. The ChIP-seq analysis which unlinked DNA fragments are analysed by high-throughput DNA sequencing platform. The detection of ChIP-seq analysis can provide better genome coverage and higher sensitivity for organisms with large genomes. There are many factors that are potential obstacles and leads to bias in ChIP-seq analyzes. Antibody deficiencies which either lack of reactivity to the intended target or can be reactive to other DNA interact proteins are general problem for both ChIP-chip and ChIP-seq. Genomic coverage due to local genome structure such as GC contents may also include the mappability of sequencing reads which lead to bias to true signals. Pair-end sequencing is well suited for providing additional information in detecting structural rearrangement and adjusting multiple mapping issues in other genomic studies. However, in ChIP-seq analysis, the design of only sequencing two ends of DNA fragments can lead to fragment bias and bring difficulty to estimate the true binding sites. Other design issues, for

example, read depth or imbalanced depths for ChIP and chromatin input samples which is used as control for background noise is discussed in review papers. (Shirley Liu et al. 2012, Marc T. Facciotti et al. 2010). Beside the obstacles from sequencing procedure and design of experiments, the influence oriented from choice of computational algorithm and relative parameters should not be underestimated.

The ChIP-seq procedure generate library of DNA fragments which are significantly enriched in regions that target protein binded with. Afterward alignment software (e.g. bowtie, bwa, and etc.) determine the genome locations of these billions of reads by mapping these sequence 'tags' against a known genome reference. In the downstream analysis, ChIP-seq peak callers based on statistical models scan through whole genome to search for regions significantly enriched in the ChIP sample compared with chromatin background or a universal background. There are more than 30 published peak calling algorithms until present. Generally, most peak callers are based on either Poisson model or negative binomial model to address the count data. For instance, MACS [33] uses a localized Poisson parameter to test significance of tag enrichment adjusting the local mapping structure; CisGenome [15] applies a negative binomial model for overdispersion. Other calling algorithms use tags on different strand information to construct the score by considering the enrichment changing of different strands (e.g. SPP [17]). The performance of different algorithms are well studied by multiple review papers in aspects of sensitivity, specificity, motif enrichment analysis, dependency of read depth and reproducibility (e.g. Shirley Liu et al. 2012). However, there is no universal best performed software in all aspects so far. Facciotti et al. (2010) compared pair-wise overlap of peak list by 11 programs. Although there are 75-80% overlap of smallest peak list across programs, there are only 45-55% of peaks shared with smaller peak lists with a large peak list, which indicates among the top ranked peaks across different callers are similar while relative lower ranked peak lists can be very different.

As senarios showed in figure 1, different peak callers apply different algorithms and report various peak regions. The top one track in each figure (1A, 1B and 1C) is the chromatin background, while the second tracks are reads pilp-up from the Su(Hw) antibody treated

sample. Track 3 to 6 are detected peak regions reported by MACS, SISSRs, CisGenome, and SPP respectively. Data generating this figure are from pair-end reads. As track two shows, we can see due to the nature of pair-end reads, there are two major peaks in this figure. However, the true motif binding site should be located in the middle of the two peaks due to fragment bias. MACS has relatively accurate predict for the enrichment region containing the true binding site. SISSRs tend to capture narrow regions; on the other hand, SISSRs reported two small regions of local reads enrichment. SPP reported a very long region (reported region comes from SPP output named with suffix "narrow_peaks"), which will certainly cover the true signal but also include non-significant area.

Motivated by the cases discussed above, we propose a meta-analysis based approach to combine sigficant peaks detected by different callers, reprioritize the peak rank and evaluate the results in terms of accuracy compared to ChIP-chip and motif enrichment analysis in multiple datasets. The purpose of this study is to integrate information from multiple callers and provide a robust list of biologically meaningful region detection. The major difficulties encountered come from three aspects. Firstly, unlike ChIP-chip can that can map events by probes, there is no unified matching matching of peak region identity by different callers. To address this issue we combine peaks by measuring genomic distance and construct candidate peak identities (CPIs). Secondly, the measurement of significant level provide by different callers are variables, ranging from p-value, FDR, folder enrichment or standardized score. We propose to use quantile normalization and use rank information mapped to a reference distribution for the top ranked peaks of each callers. Finally, significant level of non-top ranked peaks will be treated as missing values and be addressed by evidence aggregation adjusted for truncated p-values. The detailed procedure is discussed in the methods section.

2.2 METHOD AND MATERIALS

2.2.1 Data Sets

Our analysis are based on 3 datasets, 2 datasets generated from *Drosophila* S2 cells, are kindly provided by Dr. Shirley Liu's lab. The other human ChIPseq dataset is based on STAT1 binding in K562 cells performed by Yale University, which is downloaded from ENCODE project website, we use interferon-α and interferon-γ stimulated cells data as cases while non-antibody treated cells sample as background control. For each stimulated group there are two subgroups with different interferon treated time, either 30mins or 6h for both case and control groups. And there are 2 biological replicates in each subgroups.

Drosophila S2 cells ChIP-seq datasets were generated for a site-specific transcription factor (Suppressor of Hairy-wing) and a histone modification (H3K36me3). For both experiment, there existed two datasets including single-end and pair-end reads. Additional chromatin input experiment are also performed with both single-end and pair-end respectively as background information. ChIP-seq data is generated by Illumina Genome Analyzer IIx following the manufacturer's protocols. ChIP-chip analysis was using the MAT algorithm, which is among the best peak-calling algorithms for ChIP-chip data from Affymetrix data with a band width of 250 bp, a p-value cutoff of 10^{-5} and a false discovery rate cutoff of 5%.

2.2.2 Meta-Caller Method

As discussed in the introduction, the potential obstacles of combine results from multiple peak callers are discordant definition within each callers, various measurements of significant levels of signal, and missingness of peak detection in a subset of callers. Thus, we propose a general workflow shown in Figure 2. We generate common candidate peak identities (CPIs) by exhaustively searching the union of peak signals called by all software. Quantile normalization with the reference p-value distribution is applied to top ranked peaks of each callers. Non-significant peaks with no mormalized p-values are treated as missing values.

Meta-analysis by combining p-values using Fisher's method is performed and finally the CPIs are reprioritized.

1. **Individual Analysis** Raw FastQ reads are aligned with tools Bowtie version 0.12.7 with options that allowing 2 mismatch per read. Human ChIP-seq reads are aligned against UCSC hg19 reference and Drosophila reads are aligned against Ensembl BDGP5 reference. We chose 6 algorithms that are capable of using chromatin input data, are not restricted to only TF or histone marks, are supportive for analyzing ChIP-seq data, and are among the most cited algorithms. These peak callers are MACS (version 2), SISSRs [16] and PeakSeq(version 1.4), SPP, CisGenome (version 2.0), USeq [22] (version 8.7.9) and PeakSeq [25] (version 1.25). The output BAM files from Bowtie are applied to these algorithms respectively. The output data matrix of each caller contains peak chromosome location, peak summit, and index of peak calling strength p_{ij} for the jth peak in the ith caller. For MACS, SISSRs, and PeakSeq, data matrices report p-values; for CisGenome and USeq, they report FDR; and SPP report S_{wtd} scores, which are based on the count of tags both upstream and downstream.

   We use c_i, where i is from 1 to $K = 6$, to denote the ith caller. By setting default thresholds in each caller, we observed N_i reported peaks for the ith caller. For the peak reported by the ith caller denoted by m_{ij}, we observed a peak location $S_{ij} = [l_{ij}, u_{ij}]$, and an observed p-value or significant score p_{ij}.

2. **Construct Candidate Peak Identities**

   Below we construct Candidate Peak Identities(CPIs) using a graphical algorithm. Nodes represent different peak events from callers, m_{ij}. Nodes will be connected by edges when the pair-wise distance are below a threshold d. Define pair-wise distance function $D(\cdot, \cdot)$ between any two peaks is defined as:

$$D(S_{ij}, S_{i'j'}) = \begin{cases} 0, \text{when } S_{ij} \text{ and } S_{i'j'} \text{ overlap;} \\ min(|u_{ij} - l_{i'j'}|, |l_{ij} - u_{i'j'}|), \text{non-overlapping and on same chr;} \\ \infty, \text{when } S_{ij} \text{ and } S_{i'j'} \text{ are on different chromosomes.} \end{cases} \tag{2.1}$$

Two nodes, $m_{ij}, m_{i'j'}$, which are the jth and j'th observed peak event from callers $c_i, c_{i'}$respectively, will be connected with an edge, if and only if, $D(S_{ij}, S_{i'j'}) \leq d$, where d is a distance threshold that we set at $d = 50bps$ in this paper. The CPIs are identified by searching the maximal connected subgraphs. We define CPIs with chromosome coordinate interval $[L_{ij}, U_{ij}]$, where L_{ij} is minimum of l_{ij}, and U_{ij} is maximum of u_{ij} of all peaks in the CPI. Figure 3 shows two example CPIs.

Suppose we obtain M CPIs. Denote by $\tilde{S}_j$ $(1 \leq j \leq M)$ the jth CPI and $\tilde{p}_{ij}$ significant score of the jth CPI in the ith caller. When a CPI contians multiple peaks in a caller (e.g. caller c_2 in $Peak_2$ of Figure 3), the significance score is summarized by geometric mean. When a CPI contans no peak for a specific caller (e.g. caller c_4 in $Peak_2$ of Figure 3), the significance score is treated as missing at this step. Finally, we obtain a signifiance score matrix $U = \{\tilde{p}_{ij}; i \in [1, 6], j \in [1, M]\}$.

3. **Quantile Normalization and Missing Value Imputation**

To compare significant scores across different caller outputs, we applied quantile-normalization method which is widly applied to eliminate batch effect in microarray data analysis. Bolstad et. al. (2003) [1] proposed the method is to make identical distribution of probe intensities for each array. We applied this idea and use significant score vectors from three callers (MACS,SISSRs, and PeakSeq) which report p-value to build a reference p-value distribution. Then normalized p-values of CPIs from each caller vectors is generated according to the quantile of CPIs from the CDF of reference p-value distribution. The detailed steps is described as below:

We denote significant score vectors as $\tilde{P}_i = \{\tilde{p}_{i1}, ..., \tilde{p}_{iM}\}^T$, note that $U = \{\tilde{P}_1, ..., \tilde{P}_6\}$.

$$U = \begin{pmatrix} \tilde{p}_{11} & \cdots & \tilde{p}_{61} \\ \vdots & \ddots & \vdots \\ \tilde{p}_{1M} & \cdots & \tilde{p}_{6M} \end{pmatrix} \tag{2.2}$$

- Allocate truncated p-values: To destinguish $\tilde{p}_{ij}$ is whether missing or not, we introduce the censoring indicator τ as below:

$$\tau_{ij} = \begin{cases} 0, & if\ \tilde{p}_{ij}\ is\ observed; \\ 1, & if\ \tilde{p}_{ij}\ is\ missing. \end{cases}$$

Since each caller only reports top significant peaks, we assuming the missing p-values truncated due to non-significant and uniformly distributed from the threshold α_i to 1.

- Building reference p-value distribution: We assume the caller generating p-values are with caller index 1 to 3. The minimum count of non-missing p-values is m. After sorting, we have working matrix U^*. We form matrix U_m by column bind non-missing part of $\tilde{P}_1, \tilde{P}_2$, and $\tilde{P}_3$ (i.e. the left upper m by 3 element of matrix U^*). Row-wise average of U_m is taken as a vector of p-value, $(\tilde{p}^*_1, \cdots, \tilde{p}^*_m)$. Based on this p-value vector, function that maps rank to p-value, denoted as $F_p(r)$ is generated as reference p-value distribution.

$$U^* = \begin{array}{c} \text{state} \\ 1 \\ 2 \\ \\ m \\ m{+}1 \\ \\ M \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \left[\begin{array}{cccccc} \tilde{p}^*_{11} & \tilde{p}^*_{(21)} & \tilde{p}^*_{(31)} & \tilde{p}^*_{(41)} & \tilde{p}^*_{(51)} & \tilde{p}^*_{(61)} \\ \tilde{p}^*_{12} & \tilde{p}^*_{(22)} & \tilde{p}^*_{(32)} & \tilde{p}^*_{(42)} & \tilde{p}^*_{(52)} & \tilde{p}^*_{(62)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \tilde{p}^*_{(1m)} & \tilde{p}^*_{(2m)} & \tilde{p}^*_{(3m)} & \tilde{p}^*_{(4m)} & NA & \tilde{p}^*_{(6m)} \\ \tilde{p}^*_{(1m+1)} & \tilde{p}^*_{(2m+1)} & NA & \tilde{p}^*_{(4m)} & NA & NA \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ NA & \tilde{p}^*_{(2M)} & NA & \tilde{p}^*_{(4M)} & NA & NA \end{array}\right] \end{array} \tag{2.3}$$

- Truncated p-value imputation and normalization: As we described above, we imputed missing p-value by mean imputation and normalized non-missing p-values as the fomula below, the resulting p-value is denoted as ρ_{ij}:

$$\rho_{ij} = \begin{cases} \frac{1+\alpha_i}{2}, & if\ \tau_{ij} = 1; \\ F_p(r_{ij}), & if\ \tau ij = 0. \end{cases}$$

where r_{ij} is the rank of jth CPI in ith caller.

Since mathematics form of quantile normalization is not that straight forward, we give a real data example shown in Figure ??

4. **Meta-Analysis Procedure to Reprioritize CPIs**

In the quantile normalization step, we intentionally normalize all significant scores into a p-value based reference distribution. To reprioritize the CPIs, we apply a conventional Fisher's method to combine the normalized p-value of the six callers. Under the circumstance of non-missing value, meta-analysis that combining p-value can be achieved by directly applying Fisher's method by calculating evidence aggregation:

$$T = \sum_{i=1}^{K} -2ln(\rho_{ij}) \tag{2.4}$$

Under null hypothesis, $T \sim \chi^2_{2K}$, where K is number of combined studies. However, in our case, truncated p-value which is mean imputed is combined here. The Fisher's original assumption does not hold.

Tang et. al. (2014) proposed evidence aggregation mean imputation method to impute truncated p-value and derived the mean imputed evidence aggregation distributions. We applied this method and calculated mean imputed evidence aggregation as:

$$\tilde{\rho}_{ij} = \rho_{ij} I_{\{\tau_{ij}=0\}} + \frac{1+\alpha}{2} I_{\{\tau_{ij}=1\}} \tag{2.5}$$

$$\tilde{T} = \sum_{i=1}^{K_{i1}} F_X^{-1}(\rho_{ij}) I_{\{\tau_{ij}=0\}} + \sum_{i=K_{i1}+1}^{K} F_X^{-1}(\rho_{ij}) I_{\{\tau_{ij}=1\}} \tag{2.6}$$

Define

$$A = \sum_{i=1}^{K_{i1}} F_X^{-1}(\rho_{ij}) \tag{2.7}$$

$$b_i = F_X^{-1}(\frac{\alpha_i}{2}) + F_X^{-1}(\frac{1+\alpha_i}{2}) \tag{2.8}$$

$$c = \sum_{i=K_{i1}+1}^{K} F_X^{-1}(\frac{1+\alpha_i}{2}) \tag{2.9}$$

The the CDF of evidence aggregation is followed:

$$Pr(T < t) = \sum_{l=0}^{K_{i2}} f(l; K_{i2}, \alpha_i) F_A^{-1}(t - c - \sum_{i=1}^{K_{i2}} i b_i) \tag{2.10}$$

where K_{i2} is number of censored p-value for caller i, $f(l; K_{i2}, \alpha_i)$ is binomial density and F_A^{-1} is the CDF of A which follows $N(0, K_{i1})$. From this CDF of evidence aggregation we can get the combined p-value for all peak clusters with truncated p-value taken into consideration. Table 1 is an example of reported list by ChIP-MetaCaller with the information of CPIs combined peak ID and significant measure from each callers and a combined p-value in the last column.

2.2.3 Evaluation Criteria

2.2.3.1 Motif Enrichment Analysis We use transcription factor related motif enrichment to measure the biological meaning of the reported list. Motif enrichment analysis is measuring among total number of basepair from reported top ranked list of callers, the number of relative motifs count.

2.2.3.2 Accuracy comapre with ChIP-chip We use the enriched region identified by ChIP-chip analysis as a subset of true positive peaks to evaluate the overall sensitivity among all peak calling sofwares. The sensitivity is defined as the number of overlapped ChIP-chip peaks among the number of top ranked ChIPseq caller identified peaks.

2.2.4 Caller Selection

Combining all the callers' result into the meta-analysis is not always the best approach, for two reasons: first, combining more information will not only introduce true positive but also false positive; secound, combining more callers result may lead to CPI be junction of multiple signal related peaks into one CPI which is extremely long. Then, based on motif enrichment analysis, we performed forward caller selection on ENCODE data to give

a general guidance of caller selection in this meta-analysis framework. Based on single caller results' performance, we start meta-analysis combining only best two callers, and at each step add in another caller. Choose the best addition caller to perform next step forward addition until add more caller will not improve the motif enrichment. We start from the combining two callers and choose the best performanced two callers to start according to the Motif Enrichment Analysis result. Then we add in one of the other caller at a time. The forward selection will be stopped when the motif enrichment will not improve.

2.3 RESULT

2.3.1 Caller Selection

For the forward caller selection, we started from combining two best performaned callers MACS and SISSRs, then we add one callers into meta-analysis. Among all combining three callers meta-analysis result, combining MACS, SISSRs, and Useq out performaned others. Then based on MACS, SISSRs, and Useq, we add another callers into meta-analysis. However, as Figure 5 shows, none of other combination performs better than combining MACS, SISSRs, and Useq. Thus, for a motif pattern unknown dataset, the MetaCaller framework may based on combining MACS, SISSRs, and Useq.

2.3.2 Motif Enrichment Analysis and Sensitivity

Motif associated ranked peaks was performed to evaluate the true discovery rate only on datasets (ENCODE data and Su(Hw) dataset) with known motif pattern of capturing antibody. We scan ChIP-seq peak sequences with the corresponding position-weight matrix (PWM) to check the coverage of peaks contain the known motif of Su(Hw) or STAT1 transcription factor. We use JASPAR database motif PWMs as preference and compare the performance between different callers. Motif enrichment score (MES) is calculated as ratio of number motif included in the peaks to number of top-ranked peaks.

We use the enriched region identified by ChIP-chip analysis as a subset of true positive

peaks to evaluate the overall sensitivity among all peak calling sofwares. The sensitivity is defined as the number of overlapped ChIP-chip peaks among the total number of ChIP-chip analysis detected regions. To adjust the bias oriented from different peak region lengths from callers, we use relative CPI defined regions as peak regions from single callers to overlap with ChIP-chip peak regions.

2.3.3 Accuracy Comparison

2.3.3.1 *Drosophila* S2 cells Data In the Su(Hw) dataset, we use both the motif enrichment and ChIP-chip accuracy criteria to benchmark the performance. In Figure 6, results of peak calling from six individual callers are presented for the single-end dataset (left) and paired-end dataset (right). Since some callers tend to generate narrow peaks (e.g. SISSRs) while others may generate wide peaks, we calculate the total number of base pairs included among the top peaks on the x-axis. On the y-axis, the number of motifs contained in the top peaks is shown. Since SISSRs particularly identified narrow and smaller number of peaks, we observe that it identifies a large number of motifs in relatively small number of base pairs (green curve). We applied a forward selection algorithm by adding the best next caller at a time and generated a meta-caller that combines MACS, SISSRs and Useq using our meta-analysis algorithm (red line in Figure 5). This meta-caller clearly detects more motifs than other callers for a given amount of base pairs in the top peaks. Figure 6 shows the result of adding an additional caller to the MACS-SISSRs-Useq meta-caller.

Next, we consider the high confident peaks detected by ChIP-chip as the (pseudo-) gold standard and perform an evaluation of the accuracy on both Su(Hw) transcription factor data and H3K36me3 antibody data. In Figure 7, x-axis represents the number of top-ranked peaks selected by each individual or meta callers. On the y-axis, we calculate the sensitivity of peak detection (i.e. among peaks detected by ChIP-chip, the percentage of peaks detected by the callers). Meta-caller clearly recover more ChIP-chip signals than other callers for a given amount of base pairs in the top peaks. The result demonstrates meta-caller perform equally with the best performance caller with very topped rank and have higher sensitivity than other callers when rank higher.

2.3.3.2 ENCODE STAT1 Data Similarly, We performed motif enrichment analysis on ENCODE STAT1 dataset with same strategy which used in analyzing *Drosophila* data. Among 8 datasets, according to motif enrichment, ChIP-MetaCaller out-performed in 5 datasets (Infa6hrep1, Infg30rep1, Infg30rep2, Infg6hrep1, and Infg6hrep2). In other datasets, either SISSRs or PeakSeq have a outstanding performance that ChIP-MetaCaller cannot achieve. When SISSRs is already reach a very high sensitivity, combining other callers can not increase more true positive addition to SISSRs detection. And for specific dataset of Infa6hrep2, PeakSeq out perform all the other callers, and the MetaCaller conclude from forward selection based on all 8 datasets that does not include result from PeakSeq which lead to missing true positive detections. The result demonstrates mostly inferior or equal performance when adding one more caller.

2.4 CONCLUSION AND DISCUSSION

In this paper, we extensively compared six ChIP-seq calling algorithms (MACS, SPP, SISSRs, CisGenome, Useq and PeakSeq) in two datasets (SuHw and ENCODE STAT1). We used the corresponding motif binding sites retrieved from an existing database JASPAR as the first benchmark. When ChIP-chip data are available (in the first SuHw dataset), we treat the confident peaks called from ChIP-chip as the gold standard and use the sensitivity calculation as the second benchmark. Since no caller dominantly outperforms other methods, we proposed a meta-calling algorithm by combining multiple callers. In the two sets of data evaluated, the meta-caller outperformed individual callers in most situations. Combining MACS, SISSRs and Useq seem to provide the best performance except for one dataset in ENCODE where PeakSeq outperforms all other methods by a large margin. In this case, including PeakSeq in the meta-caller will improve the performance.

In a routine application when the motif pattern and binding sites are known, we suggest to use the information and apply the forward selection algorithm to select the best subset of callers for constructing a meta-caller. There are a few limitations of our study. Firstly, the motif binding patterns and binding sites from the database may not always be available

or accurate. This may affect the application to future datasets. Secondly, we treat the confident peaks called from ChIP-chip as the gold standard in the evaluation of SuHw. The logic sounds contradictory to the intuition and common belief that ChIP-seq is more accurate thatn ChIP-chip. But since we use a conservative threshold for ChIP-chip to obtain confident peaks and we use this criterion as the secondary benchmark, the result seems to be consistent with the motif enrichment evaluation.

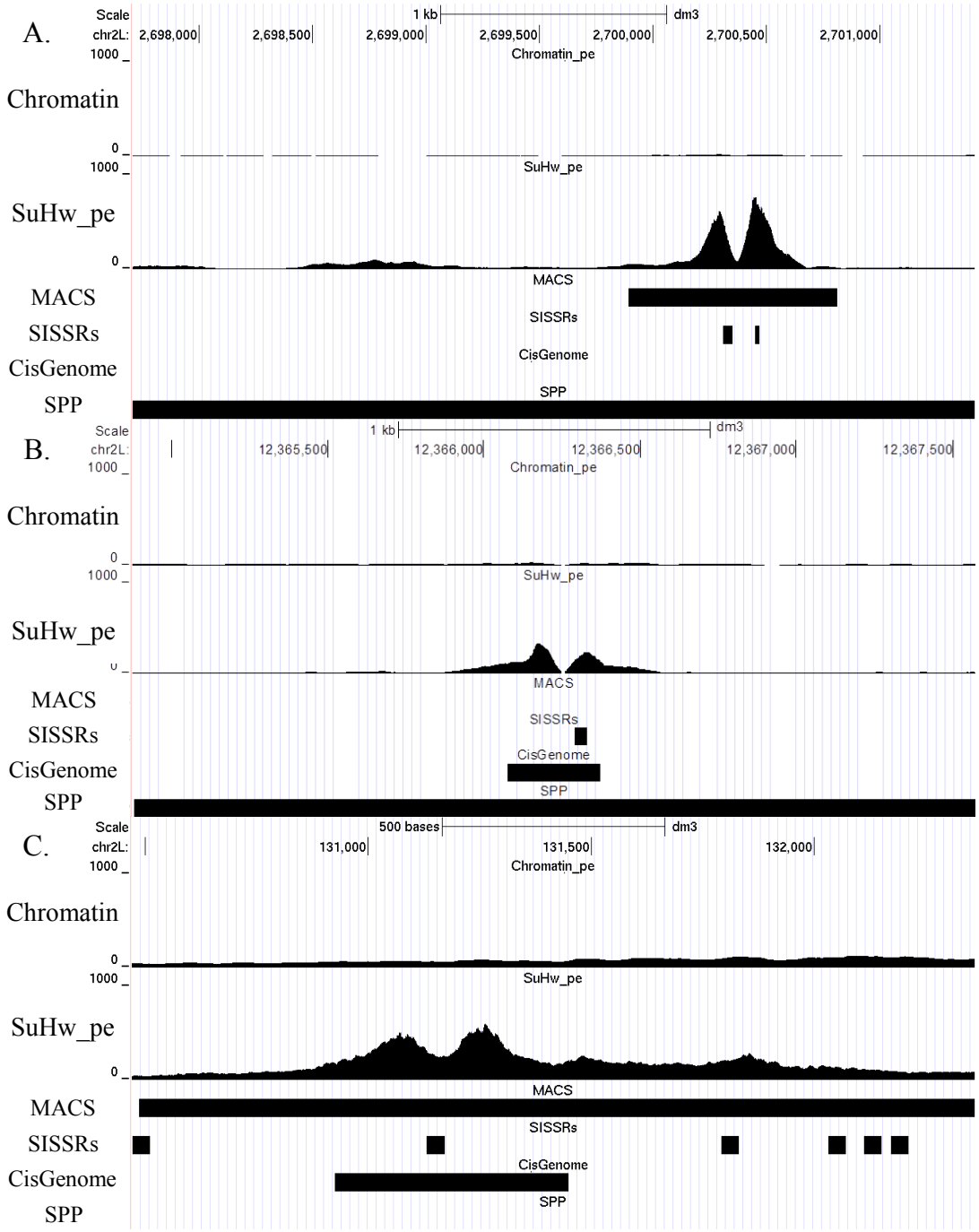

Figure 1: View of ChIP-seq raw reads pile-up and software defined peak regions

Table 1: An Example of ChIP-MetaCaller Report.

CPI	Chr	Start	End	MACS_ID	MACS_m	SISSRs_ID	SISSRs_m	CisGenome_ID	CisGenome_m	SPP_ID	SPP_m	p_value
174	chr2L	3107975	3109032	286	1.226e-27	6159	0.0043	466	0.310493	160	3884.0077	4.595e-49
...	...	...	...	...	...	...	...	...	...	...	...	...
2240	chr2L	8514846	8515328	714	2.616e-03	9375	0.0780	NA	NA	419	250.4781	1.248e-08
...	...	...	...	...	...	...	...	...	...	...	...	...
48641	chr2L	12366075	12366374	NA	NA	1167	0.0096	2141	0.982061	573,574	456.655	3.922e-12

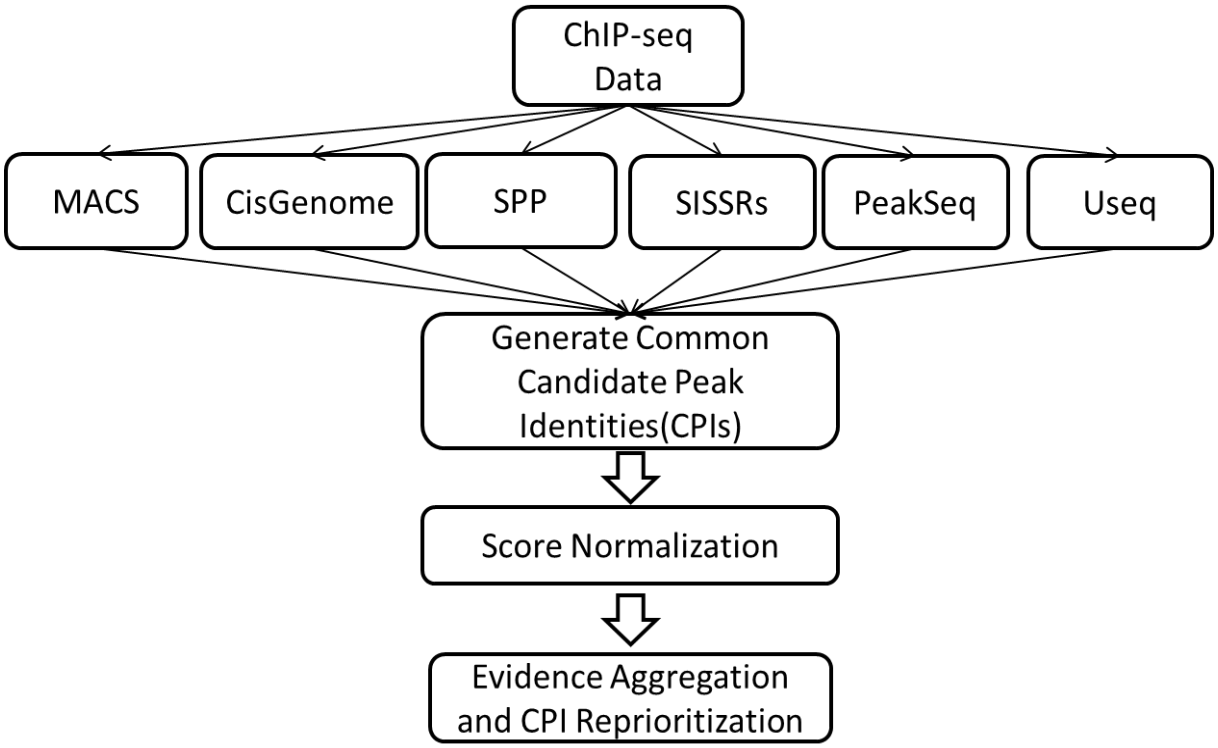

Figure 2: General workflow of ChIP-MetaCaller.

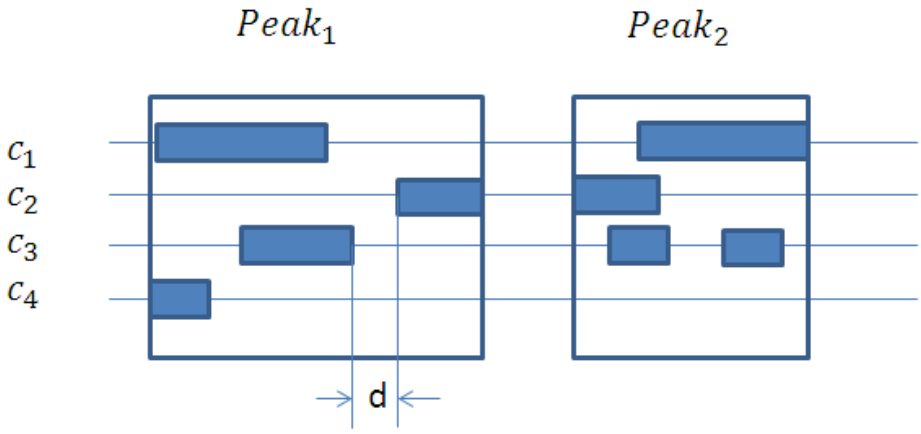

Figure 3: Definition of CPIs.

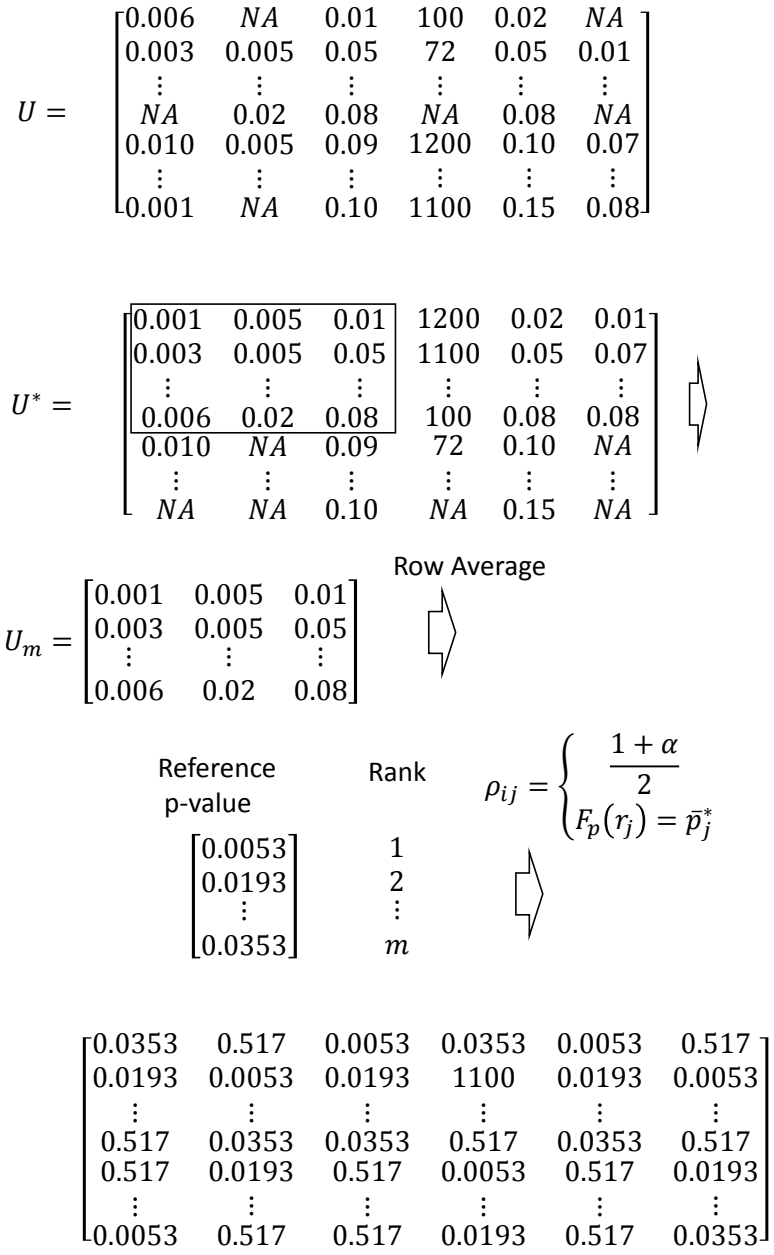

Figure 4: Real data example of quantile normalization.

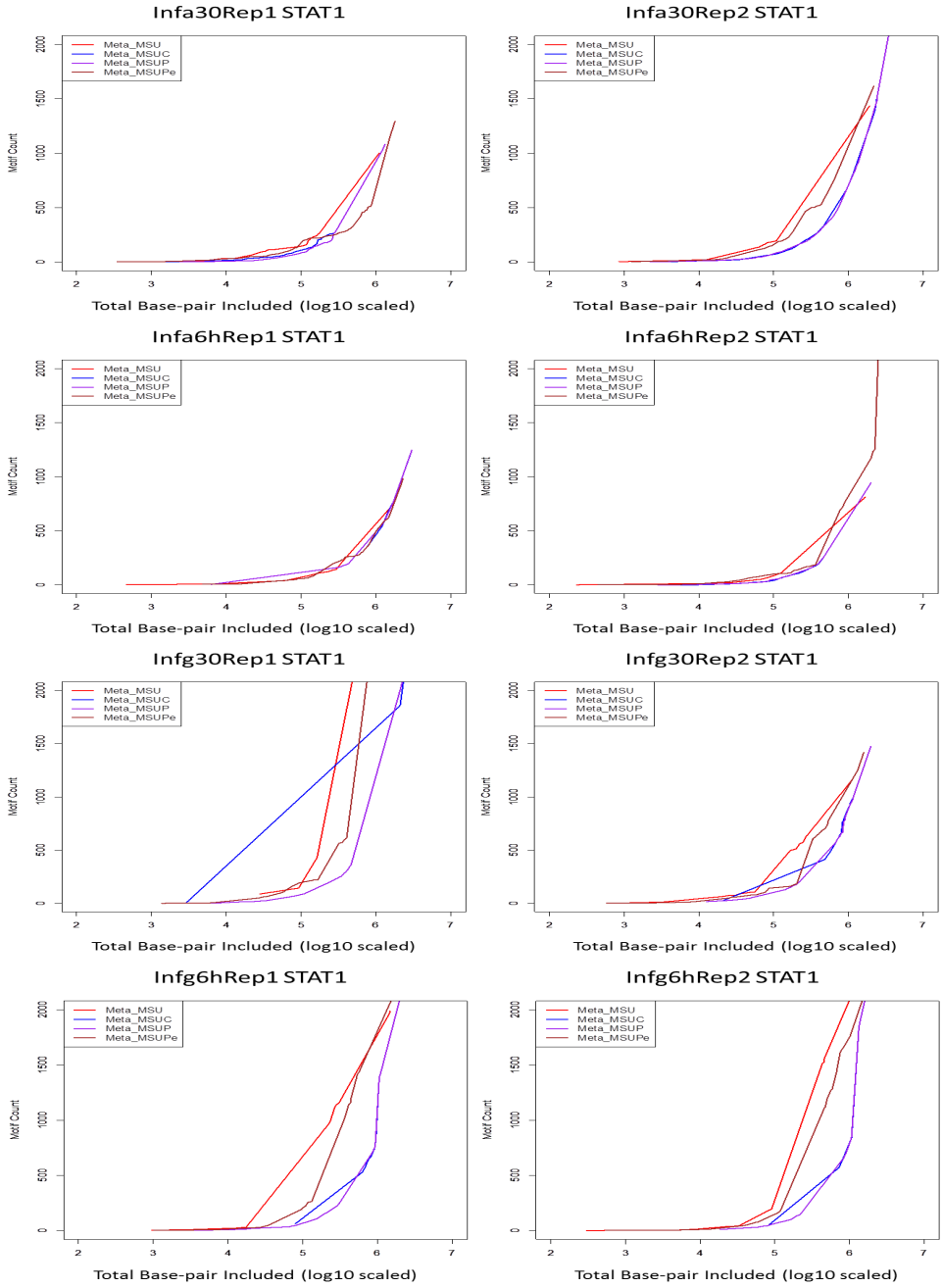

Figure 5: Motif Enrichment Analysis based forward selection of combination of callers. This figure shows comparison of combine MACS, SISSRs, and Useq against combining these three callers plus one more caller.

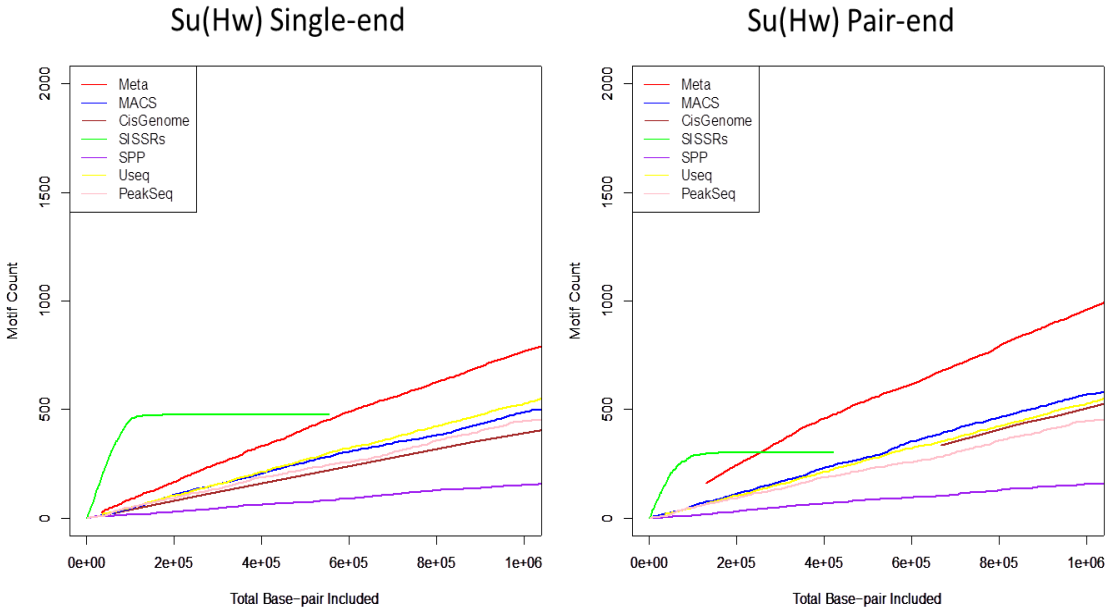

Figure 6: Motif enrichment score based on Su(Hw) dataset, by comparing fraction of containing a motif by total base pair included from top ranked peaks.

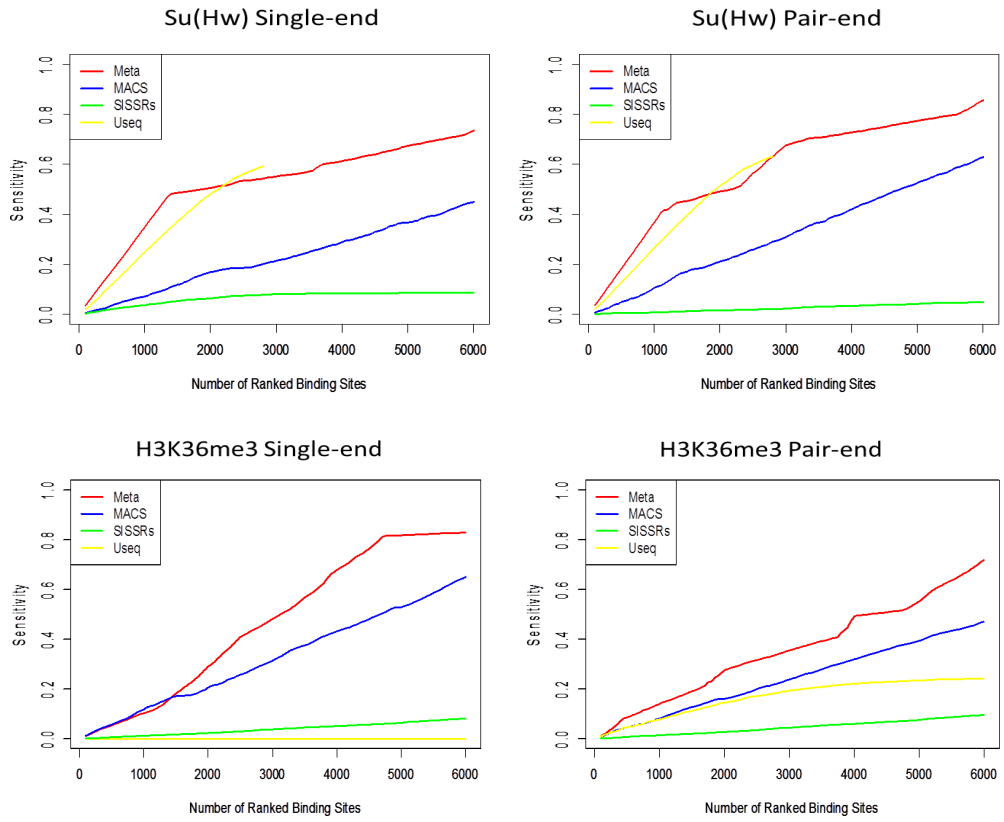

Figure 7: Sensitivity by ChIP-chip peak overlap among top ranked peaks. Su(Hw) single-end (topleft), and pair-end (topright);H3K36me3 single-end (bottomleft), and pair-end (bottomright)

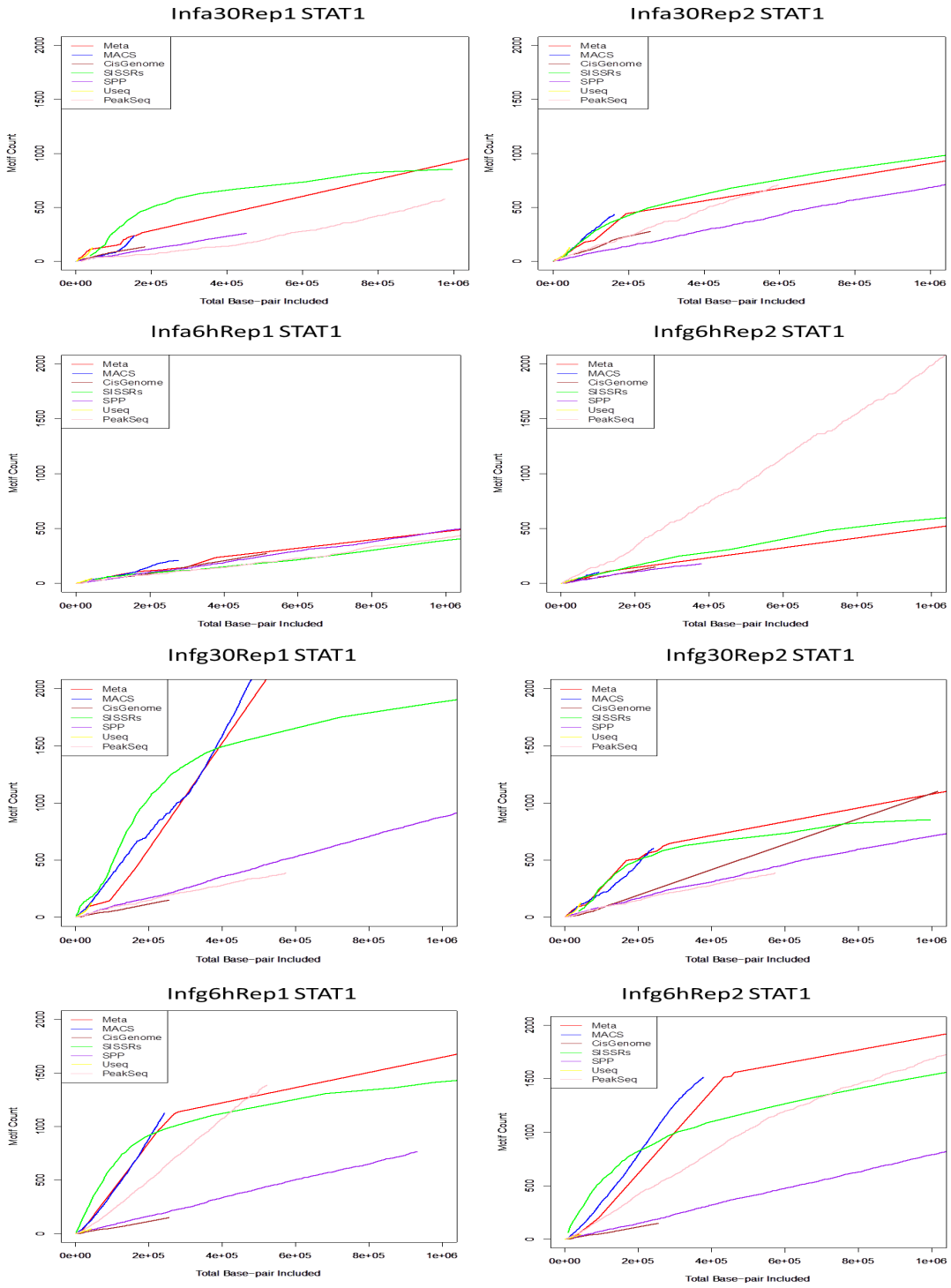

Figure 8: Motif enrichment analysis for the 8 datasets from ENCODE project.

3.0 GENE META CLUSTERING

3.1 BACKGROUND

Many genetic diseases are complex in their etiology and caused by a combination of multiple genetic abnormalities. Researchers investigated same biological problem based on data generated by tissue sample provided by patients and donors. Combining multiple studies in co-expression analysis can integrate information and achieve larger sample size. Many previous studies have extended cluster analysis on gene expression profile to integrative studies. Some studies (Mabbott,N.A. et.al. 2009, Carrera,J. et. al. 2009,J upiter,D.C. and VanBuren,V. (2008)) directly merged individual finding in single studies into a network. Others combine pairwised gene interactions across studies by either vote counting (Niida,A. et. al. 2009) or Fisher's method (Srivastava,G.P. 2010) which is similar to meta-analysis for DE detection. Segal et al. conducted probably the first large-scale microarray meta-analysis for network or co-expression analysis and developed module map by combining 1975 arrays in 26 cancer studies to characterize expression behavior of 2849 modules collected from various sources (e.g. Gene Ontology, KEGG pathways and gene expression clusters). Wang et al. formulated a regularized approach to combine multiple time-course microarray studies for inferring gene regulatory networks. Zhou et al. proposed a 2nd-order correlation analysis to construct network and functional annotation by combining 39 yeast data sets. Huttenhower et al. used a scalable Bayesian framework to combine studies for pairwise meta-correlation and predicted functional relationship. Wang et al. developed a semi-parametric meta-analysis approach for combining co-expression relationships from multiple expression profile data sets to evaluate similarity and dissimilarity of gene network across species. Steele et al. proposed a weighted meta-analysis Bayesian network based on combining statistical confidences attached to net-

work edges and a consensus Bayesian network to identify consistent network features across all studies. However, none of the methods above provides a general framework of expression profile integration to detect gene co-expression modules. A proposed meta-clustering based on co-expression analysis is performed by Chang et al. (2014) [3] at combining studies at distance measure levels. In this chapter, we will also propose to perform meta-clustering at clusters level and perform comprehensive evaluations. Instead of detecting co-expression pattern of all studies, the new proposed method is aiming to detect co-expression gene modules at combining cluster result from single clustering studies. A general workflow of two meta-clustering analyses is shown in Figure 9.

3.2 MATERIALS AND METHODS

3.2.1 MetaClustering by combining distances(MetaCluster.D)

The existing way to combine multiple transcriptome studies to construct co-regulated gene modules using meta-clustering algorithm is proposed by Lun-Ching Chang [3]. Denote by X_{gsk} the gene expression intensity of gene g, sample s and study k, and $X_{gk} = (X_{g1k}, \ldots, X_{gSk})$ the vector of gene expression intensities of gene g and study k. Define the dissimilarity measure as distance between gene i and gene j for a given study k is calculated by $d_{i,j}^{(k)} = 1 - |\text{cor}(X_{ik}, X_{jk})|$, where $\text{cor}(X_{ik}, X_{jk})$ is the Pearson correlation of the two expression intensity vectors. To combine the dissimilarity information of the K studies, we took mean of meta-dissimilarity measure between gene i and gene j as $d(g_i, g_j) = \text{Mean}(d_{ij}^{(1)}, d_{ij}^{(2)}, \ldots, d_{ij}^{(K)})$. Given the meta-dissimilarity measure, the Penalized K-medoids clustering algorithm was then applied to construct co-expression gene modules [27]. The target function to be minimized by Penalized K-medoids is shown below:

$$L(C) = \sum_{i=1}^{G} \sum_{g_i \in C_h} d(g_i, \overline{g}_h) + \lambda \cdot |S| \tag{3.1}$$

where the clustering result $C = (C_1, \ldots, C_H, S)$, contain H non-overlapping tight clusters and a set of scattered genes S cannot be groupped into any of H clusters. $\overline{g}_h$ denote the

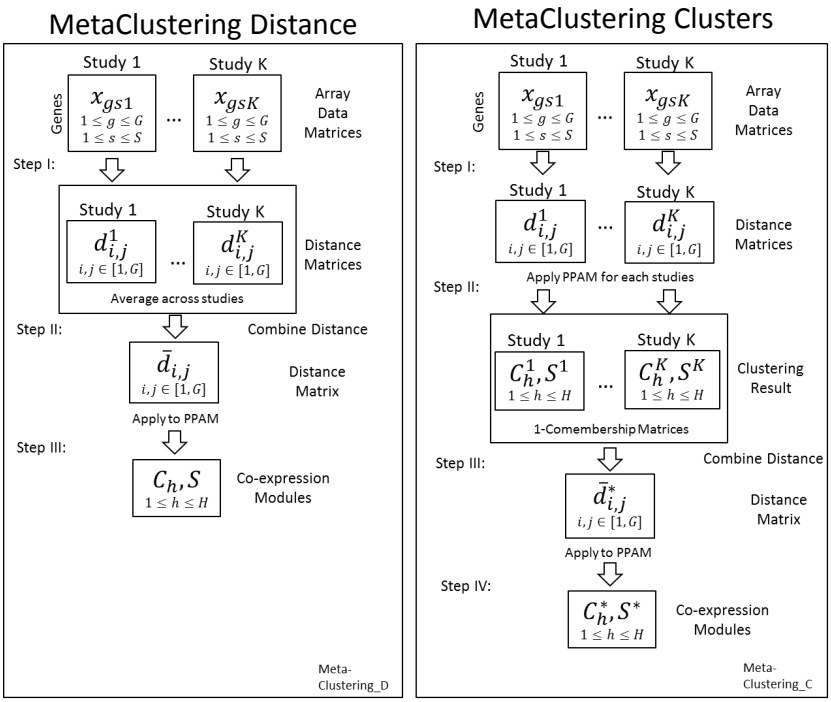

Figure 9: General workflow of meta-clustering methods to combine co-expressed genes in different approaches. A. Meta-clustering Distances; B. Meta-clustering Clusters

mediod gene of cluster h whose average distance to all other genes in the cluster is minimal.$|S|$ denoted the size of the scattered gene set S. λ is a tuning parameter controlling the tightness of gene clusters and the number of scattered genes.

3.2.2 MetaClustering by Combining clustering results(MetaCluster.C)

Unlike using average dissmilarity measure, we propose an alternative MetaCluster.D approach. To help understanding the concepts,we generate a 2D data case showed in Figure 10. We use: k as the index of studies, $k = 1, ..., K$; i as the index of genes $i = 1, ..., G$; h as the index of cluters, $h = 1, ..., H$. Yet we have data matrix $X_{gs} = (X_{gs1}, \ldots, X_{gsK})$, the dissimilarity measure for gene i and gene j is calculated the way described in section 3.2.1. Instead of averaging distance matrix across studies, we applied Penalized K-mediods clustering algorithm to each study. The clustering results in a matrix $\mathbf{C} = (C^{(1)}, \ldots, C^{(k)})$. Each vector of $\mathbf{C}$, $C^k = (C_1^k, \ldots, C_H^k, S^k)$ contains H gene clusters and a scattered gene set.

We denote O_h^k as the mediod fo cluster h in study k. d_{ih}^k is denoted as the euclidean distance of gene i to O_h^k. δ_h^k is distance threshold for cluster h. We normalize the distance across clusters by dividing the threshold.

$$u_{ih}^k = \frac{d_{ih}^k}{\delta_h^k} \tag{3.2}$$

Then we use averaging distance to the mediods from the tightest cluster as the cluster distance represented the study. In the case showed in Figure 10, we choose the v_{ij}^k with the orange color. Compare the distance between two gene vectors in Figure 10A and B, the visually nearer pairs of genes set will conclude a smaller v_{ij}^k. In comparison between 10A and C, we can see, two genes from different clusters, will conclude a larger v_{ij}^k compare to those are from same cluster.

$$v_{ij}^k = f(u_{ih}^k, u_{jh}^k) = argmin_k \sum_{(i,j)} \frac{u_{ih}^k + u_{jh}^k}{2} \tag{3.3}$$

Then we use two different combine method to generate the entry distance matrix of P-PAM. The first one is Meta.C.avg by:

$$\mathbf{D} = 1 - \frac{1}{K} \sum_{k=1}^{K} V_{ij}^k \tag{3.4}$$

The other method, Meta.C.max, is combine by using only the maximum distance:

$$\mathbf{D} = 1 - min_{k \in K} V_{ij}^{k} \tag{3.5}$$

The method of average is aim to combine the distance information of all the study with equal weight. On the other hand, the method of using the maximum distance among all the studies is aiming to be conservative when combining result across studies, which require same clustering label across all the studies.

3.2.3 Parameter Selection

Like K-means and its variants clustering methods, PPAM algorithm require a input parameter of the number of clusters. Moreover, PPAM uses parameter λ as a coefficient of penalty term to control number of scattered genes. In some situation, investigator or researcher is certain about these parameters according to prior information. Thus, user can directly set K and control λ to restrict predictive genes count. For instance, in study of yeast cell cycle, previous literature has prove that there are approximately 200 genes related to yeast cell cycle and there are four stage of cell cycle, user can directly set $K = 4$ and control λ to restrict predictive genes count around 200. However, in the case, that user have no idea of the setting of these parameters, we apply two methods to provide estimation of parameters from data.

After we introduce the method of estimating parameters, for simplicity, we would like to transform original parameters λ used in PPAM paper into a corresponding parameters β which controls the clustered gene proportion. This can be easily accomplished by R programming. This procedure is complemented by trying different λ to fit the setting of β.

3.2.3.1 Prediction Strength In this paper we follow the prediction-based resampling method proposed by Tibshirani et al. (2005) (also see Breckenridge, 1989; Dudoit and Fridlyand, 2002) for selecting k and λ. The idea of this prediction strength method is that by searching through different pairs of (k, λ) from the parameter space, find the best parameter pairs according of current the data set. The way to measure the "best", is by splitting the

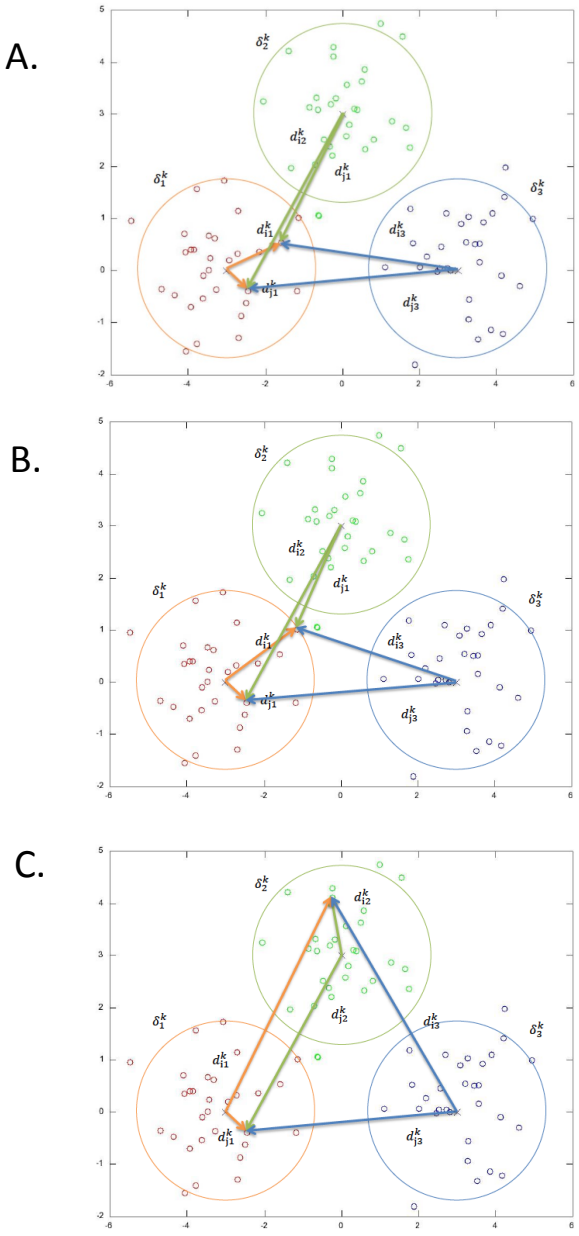

Figure 10: 2 Dimensional Model illustration of distance in MetaCluster.C.

whole data set X into two equal sized sub-data sets: training set X_{tr} and testing set X_{te}. The main idea involves three steps: (a) cluster the training data X_{tr}; (b) cluster the testing data X_{te} (c) measure how well the training set clustering result predicts co-memberships in the testing data.

The correct parameter selection should generate consistent clustering results in training and testing data and produce a good prediction in step (c). We denote by $C(X_{tr}; k, \lambda)$ the clustering operation on the training data. Following the convention in Tibshirani et al. (2001), we denote by $D[C(X_{tr}; k, \lambda), X_{te}]$ the (n/2) by (n/2) co-membership matrix in the testing data X_{te} judged by the clustering result from training data, $C(X_{tr}; k, \lambda)$. The nearest centroid criterion is used for such judgment, that is, each point in the testing data is assigned to the nearest cluster centroid of $C(X_{tr}; k, \lambda)$. For any pair of points i and i' in the testing data, the $i - i'$th element of the comembership matrix $D[C(X_{tr}; k, \lambda), X_{te}]_{ii'}$will take the value 1 if both i and i' fall into the same cluster under $C(X_{tr}; k, \lambda)$ judgment and zero otherwise. We denote by $(C_1^{te}, ..., C_k^{te}, C_{k+1}^{te} = S^{te})$ the resulting cluster indexes from clustering the test data in step (b) such that $X_{te} = \cup_{j=1}^{k+1} C_j^{te}$ and $n_1, ..., n_{k+1}$ are the number of observations in each cluster. The prediction strength of the training and testing data split is defined as:

$$ps(k, \lambda) = min_{1 \le j \le k+1} \frac{\sum_{i \neq i' \in C_j^{te}} I(D[C(X_{tr}; k, \lambda), X_{te}]_{ii'} = 1)}{n_j(n_j - 1)} \tag{3.6}$$

where $I(\bullet)$ is the indicator function which equals 1 if the statement is true and 0 otherwise. Intuitively, we compute for each cluster in the test data, the proportion of all pairs of objects that are also assigned in the same cluster by the training cluster centroids judgment. We repeat the independent samplings for the training and testing data (10 times in this dissertation) and the averaged prediction strength ps is reported. Normally $(k*, \lambda*) = argmax\ ps(k, \lambda)$ is used for final clustering in practice. However, in the context of gene clustering in microarray, we may want to select as large k and as small λ as possible with reasonably high prediction strength ($ps < 0.6 or 0.7$) so that many important tight cluster patterns are retrieved.

3.2.3.2 Consensus Clustering Although, prediction strength can theoretically provide estimation of both k and λ simultaneously, sometimes the prediction strength does not give a clear maximum across parameter space. In this scenario, we applied iterative consensus clustering (ICC) algorithm to estimate the number of clusters k and after the cluster number is fixed, then we estimate the parameter λ. The procedure of consensus clustering is (a)subsample 80% sample from original data set for 50 times; (b)use PAM algorithm to cluster these data sets with parameter k range from [2,10]; (c)based on the similarity matrix, calculate the consensus index change when k is increase. The best number of k is chosen by selecting the gain of consensus index does not exceed certain threshold. Beside of consensus index, there are heatmaps generated for each k based on the consensus matrix. The more consistent clustering result across iterations, the darker the color in the heatmap is. Clear block pattern along the diagonal of the heatmap indicates consistent clustering result across iteration. The detailed calculation and algorithm can refer to Wilkerson, M.D. 2010.

3.2.4 Evaluation Criterion

For the comparison between cluster results, we compare the predicted label with known label or other methods predicted label using adjusted rand index (Hubert and Arabie, 1985)(ARI).

3.2.4.1 Concordance across Studies For the purpose to show the consistence result of meta-analysis frame works, we compare not only the meta-analysis result of combining all the studies to the single studies results, but also shows the leave one out meta analysis result with left-out studies' result. The performance is bench-marked by adjusted rand index(ARI).

3.2.4.2 Statility To estimate the stability of each methods, we use bootstrapping method to sample same number of samples of each study with replacement, then apply each method to the bootstrapped data set. Compare the clustering result with original data sets result. Methods with good stability will show higher ARI of this comparison.

3.2.4.3 Biological Meanings To estimate the biological meanings, we apply detected co-expression modules of each methods to pathway enrichment analysis. The background

knowledge for pathway analysis ,we select database from KEGG, GO and Biocarta with relative species and relative conditions. We expect outstanding methods detected co-expression modules will have more significant pathways.

3.2.5 Data Sets

3.2.5.1 Simulated Data Set To better understand the performance of different meta-analysis framework and compare to single study clustering results, we construct simulated data set with $S(S = 4)$ and $K(K = 5)$ under three different senarios. The different senarios are related to different distributions of gene expression profiles across samples. To best mimic the nature of microarray study, we construct predictive genes with correlation structure and noise genes (e.g. housekeeping genes or unexpressed genes). Below are the detailed generative steps to create co-expression module predictive genes, and noise genes.

1. Mean expression pattern across samples: we simulated three differen expression pattern here with different mean expression level. Here we call this mean expression pattern template expression.
   - Senario 1: Samples under different subtypes: Under this senario, we want to mimic the situation that there are co-expressed gene modules can predict sample with clusters (subtypes of diseases or biological conditions related to phenotype). We simulated data set that based on ground truth of gene expression profiles within each co-expression modules, samples can categorized into $T(T = 4)$ subtypes; number of samples of each subtype t in study s are followed a poisson distribution with mean 30: $N_{st} \sim POI(30)$; number of predictive gene of module m: $n_m \sim POI(100)$; mean of expression level: $\mu_{tm} \sim UNIF(-3, 3)$.
   - Senario 2: Samples under monotone time series pattern: Under this senario, we want to mimic the situation that samples are taken from different time points and predictive gene expression value will monotone increase or decrease across time. We simualted number of samples of each study s: $N_s \sim POI(100)$; number of predictive gene of module m: $n_m \sim POI(100)$; mean of expression level $\mu_m = a_m + b_m * x^{c_m}$ with $a_m \sim N(0, 3); b_m \sim N(0, 3); c_m \sim N(1, 0.5); (c_m > 0)$.

- Senario 3: Samples under cyclic time series pattern: Under this senario, we want to mimic the situation that samples are taken from different time points and predictive gene expression value will change with a cyclic pattern across time. The difference between this with senario 2 is the mean of expression level $\mu_m = a_m + b_m * sin(\frac{\pi x}{50} + c_m)$ with $a_m \sim N(0,1); b_m \sim N(1.5, 0.5); c_m \sim N(5,2)$.

2. Correlation structure of predictive genes: with the template expression level across samples, we simulated the co-expression structure by letting expression of genes within same module followed a multi-variate normal distribution.
   - Add biological variation to template expression level of gene i: $X'_{stmi} \sim N(\mu_{tm}, \sigma_1^2)$;
   - Add in within module correlation structure:
     $(X_{stm1}, ..., X_{stmn_m})^T \sim MVN(X'_{stmi}, \rho\Sigma_{stm})$; where Σ_{stm} is standardized Σ'_{stm}(keep diagonal elements equal to 1) and:
     $\Sigma'_{stm} \sim W^{-1}(\Phi, v)$; here W^{-1} is inverse Wishart distribution, $v = 150(v > n_m - 1), \Phi = 0.5I_{n_m} + 0.5J_{n_m}, I_{n_m}$ is identity matrix and J_{n_m} is matrix with all the elements are 1.

   In this simulation study, we want to evaluate the performance of different method while the within module variation increaese, i.e. ρ increases.

3. Noise gene: in this step, we add in house keeping genes that are not related to the prediction of co-expression modules by simulating: number of scattered genes: $G_0 = 1000$; mean expression level for the scattered genes: $\mu_{tm} \sim UNIF(-5, 5)$; expression level for scattered genes: $X_{sgi} \sim N(\mu_{tg}, \sigma_2^2)$; expression variation of scattered genes is fixed: $\sigma_2^2 = 1$.

   Heatmaps of simulated data are shown in Figure 11 and Figure 12.

3.2.5.2 Yeast Cell Cycle Data Yeast cell cycle data, orignally published in Spellman's (Futcher et. al. Citation) paper, is a classical gene expression profile which studies yeast genes whose trancscipt levels in various periodically cell cycles. DNA microarrays data generated from strains with 4 different cell cycle arresting methods:α Factor, Elutriation, and *cdc15* and *cdc28* tempreture-sensitive mutant. Samples are taken periodically. More than 6000 genes microarray log2 transformed intensity ratios are recorded in the data.

Heatmap example of single study simulated under senario 1:

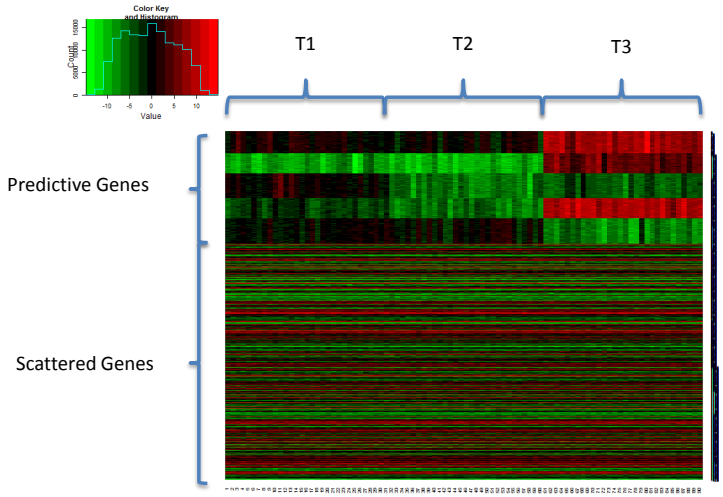

Figure 11: Simulated data visualization of senario 1

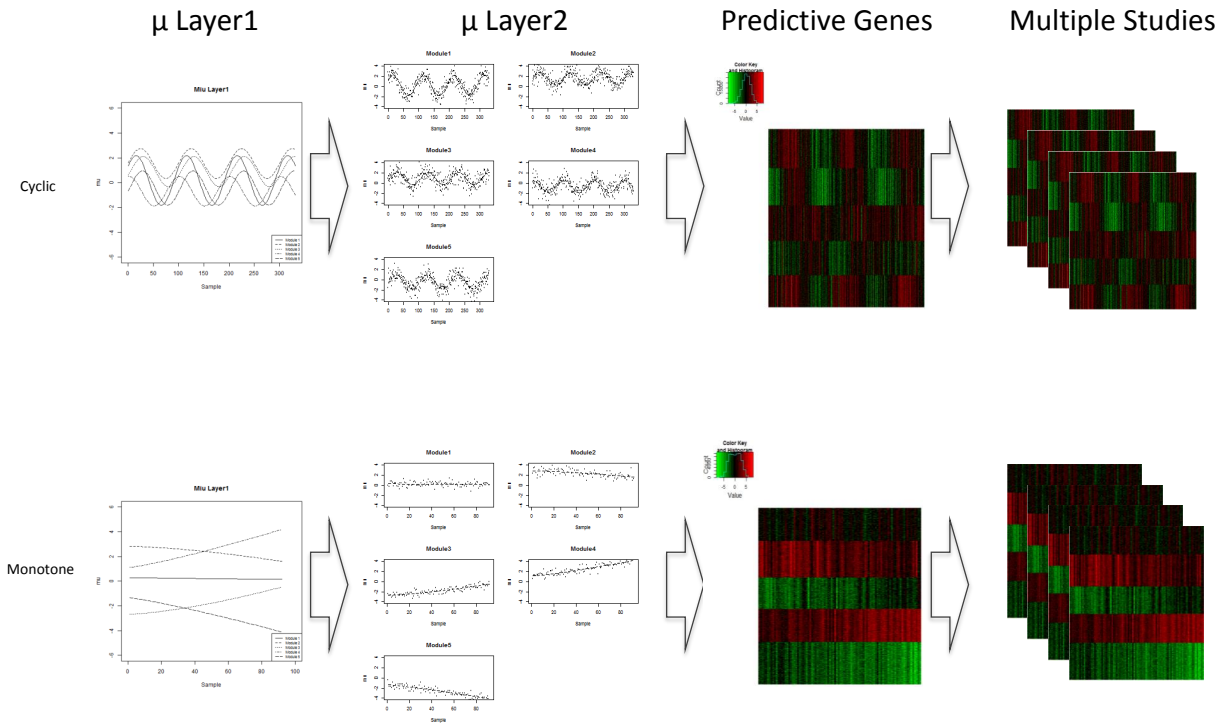

Figure 12: Simulated data visualization of senario 2 and 3

3.2.5.3 Breast Cancer Data In this data set, we combine four independent breast cancer studies: Wang (Wang et al., 2005), Desmedt (Desmedt et al., 2007), TCGA (Network, 2012) and METABRIC (Curtis et al., 2012). First three studies contain about 150-500 samples. Wang and Desmedt applied Affymetrix U133A chip that generated logintensities ranging between 2.104 and 14.389, while TCGA adopted Agilent Custom 244K array that produced log-ratio intensities ranging between -13.816 and 14.207. METABRIC (Curtis et al., 2012), which contained 1,981 samples from Illumina HT12 arrays. All probes in three studies were matched to gene symbols before meta-analysis.

3.2.5.4 Mouse Metabolism Data Mouse Metabolism Dataset: Energy metabolism in mouse model. An energy metabolism disorder in children is associated with very longchain acyl-coenzyme A dehydrogenase (VLCAD) deficiencies. In an ongoing unpublished project, two genotypes of the mouse modelwild type and VLCAD-deficient were studied for three types of tissues (brown fat, liver and heart) with 4 mice in each genotype group. Microarray experiments were applied separately to study the expression changes across genotypes.

3.3 RESULTS

3.3.1 Simulation Result

3.3.1.1 Tuning Parameter In the Figure 13, we showed iterative consensus clustering(ICC) heatmaps of $k = 4$ and $k = 5$ for the simulated data. When k approaches the best cluster number, the heatmap shows clear pattern. In addition to visualization of heatmap, the lower panel of figure 13 included increase of consensus index, and showed that when k is beyond 5, there is almost no gain of consensus index. Figure 13 is an example of ICC for scenario 1 simulated data at $\rho = 5$. In conclusion, ICC can predict cluster number $k = 5$ which is consistence with the true simulated module number. Beside of this example, the estimation of cluster number have a consistent result with above example under other simulated scenarios.

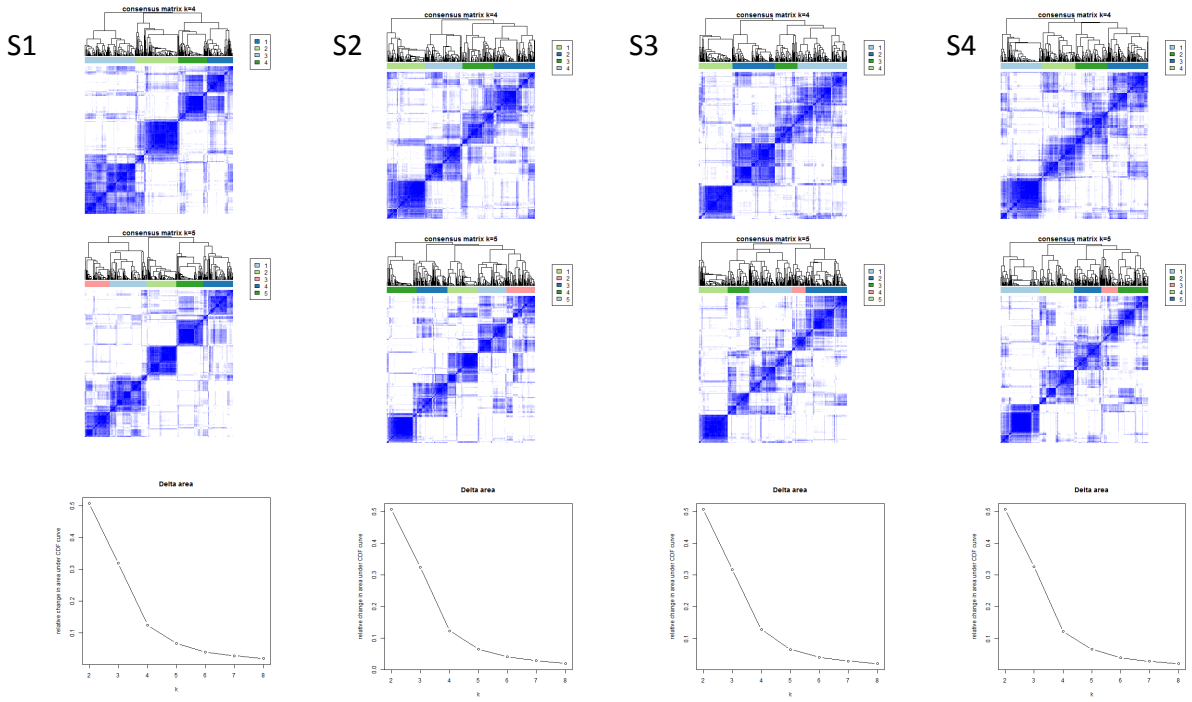

Figure 13: The Heatmaps of consensus clustering result to decide cluster number

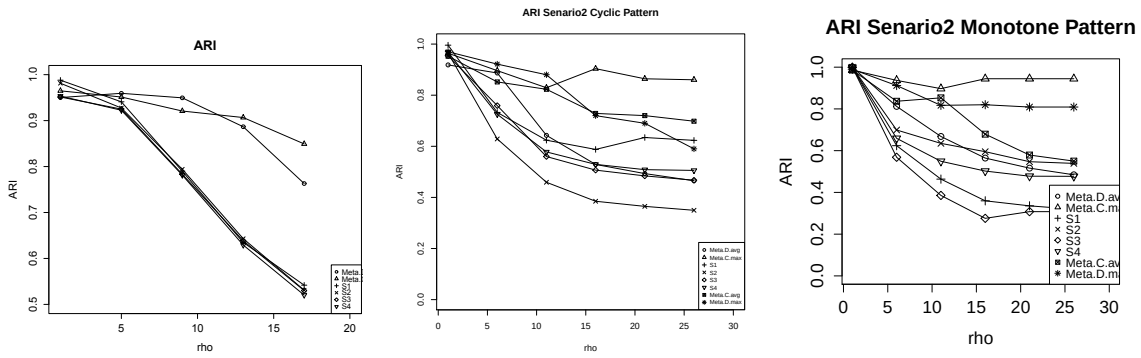

Figure 14: ARI of methods across different ρ under three different simulated scenarios(Note: When ρ exceeds 20, the ARI will not further decreases.)

3.3.1.2 Compare to Underlying Truth Since we simulated the data sets, we know the true co-expression module label of each gene. Here we compare the predicted label to the true label under different covariance matrix (ρ range from 1 to 20). Under each ρ, we simulated 100 data sets and calculated average ARIs over 100 simulations and their standard errors.

As Figure 14 shows, as with module covariance increases, ARI of all methods decreases. Both meta-analysis frameworks, MetaClust.D and MetaClust.C, outperform the single clusters result in simulated scenario 1. Under scenario 2 and 3, MetaClust.C with combining maximum distance recover better module labels than the other methods.

Why MetaClust.D perform poorly when ρ increases:

- Major Prediction Error: mislabel predictive gene into scattered gene sets;
- When within module variation increases, it is more likely one gene in certain study have similar pattern with scattered gene, especially when monotone pattern, b_m is small, the curve is flat;
- MetaClust.D averages distance over studies, lead to a larger distance, tend to cluster this kind of gene into scattered gene set;
- MetaClust.C calculate minimum standardized distance to the centroids first and then take the largest distance across study can avoid this kind of mislabel.

3.3.1.3 Concordance Across Studies Here we use leave one study out at a time perform clustering and compare to the left out study to mearsure the clustering result concordance . ARIs result listed in table2 to table 4 are based on $\rho = 9$ as according to Figure 14, this value is smallest ρ that can distinguish the performance of methods. As tables showed, similarity between single studies clustering results are relatively lower than that between both meta-analysis frameworks and single studies in all three simulated scenarios. This conclude that both meta-analysis frameworks conclude the clustering results from single study and recover the module information.

3.3.1.4 Stability Here we use bootstrapping on samples in each study, and compare bootstrapping cluster result to whole data set to measure the clustering result stability. The ARI values listed in the table 5 to 6 are average over 100 times bootstrapping and number in the brackets are the standard errors. Each bootstrap sample same numbers of samples with replacement. The result listed in the table shows that stability decreases as within module variation(ρ) increases along with standard error over bootstrapping increases for all the methods. At low within module variation situations, all the methods have good stability. However, as expression value became noisy within modules, meta-analysis frameworks outperform single studies in sense of stability under all three simulation settings.

Table 2: ARI of Concordance Across Studies measurement: Simulation Senario 1

Method	S1	S2	S3	S4
S1	1			
S2	0.852	1		
S3	0.796	0.809	1	
S4	0.877	0.861	0.816	1
MC.D(-S1)	0.912			
MC.D(-S2)		0.908		
MC.D(-S3)			0.878	
MC.D(-S4)				0.842
MC.C(-S1)	0.861			
MC.C(-S2)		0.831		
MC.C(-S3)			0.836	
MC.C(-S4)				0.814

Table 3: ARI of Concordance Across Studies measurement: Simulation Senario 2

Method	S1	S2	S3	S4
S1	1			
S2	0.681	1		
S3	0.812	0.825	1	
S4	0.710	0.834	0.616	1
MC.D(-S1)	0.791			
MC.D(-S2)		0.817		
MC.D(-S3)			0.861	
MC.D(-S4)				0.791
MC.C(-S1)	0.691			
MC.C(-S2)		0.821		
MC.C(-S3)			0.724	
MC.C(-S4)				0.800

Table 4: ARI of Concordance Across Studies measurement: Simulation Senario 3

Method	S1	S2	S3	S4
S1	1			
S2	0.621	1		
S3	0.768	0.801	1	
S4	0.613	0.787	0.794	1
MC.D(-S1)	0.786			
MC.D(-S2)		0.821		
MC.D(-S3)			0.808	
MC.D(-S4)				0.753
MC.C(-S1)	0.778			
MC.C(-S2)		0.868		
MC.C(-S3)			0.863	
MC.C(-S4)				0.879

Table 5: ARI of Stability measurement: Simulation Senario 1

Method	MC.D	MC.C	S1	S2	S3	S4
$\rho = 1$	0.99(0.01)	0.99(0.01)	0.98(0.01)	0.99(0.01)	0.99(0.01)	0.99(0.01)
$\rho = 5$	0.95(0.05)	0.97(0.02)	0.92(0.05)	0.96(0.03)	0.93(0.03)	0.89(0.02)
$\rho = 9$	0.88(0.04)	0.86(0.03)	0.72(0.04)	0.74(0.07)	0.71(0.02)	0.69(0.05)
$\rho = 13$	0.74(0.07)	0.72(0.07)	0.51(0.06)	0.56(0.12)	0.53(0.10)	0.59(0.10)
$\rho = 17$	0.58(0.08)	0.62(0.10)	0.32(0.15)	0.29(0.08)	0.43(0.12)	0.39(0.09)

Table 6: ARI of Stability measurement: Simulation Senario 2

Method	MC.D	MC.C	S1	S2	S3	S4
$\rho = 1$	0.90(0.05)	0.99(0.02)	0.99(0.01)	0.98(0.01)	0.99(0.02)	0.98(0.02)
$\rho = 6$	0.90(0.05)	0.97(0.05)	0.92(0.08)	0.96(0.03)	0.93(0.05)	0.89(0.06)
$\rho = 11$	0.80(0.08)	0.84(0.05)	0.72(0.06)	0.71(0.05)	0.73(0.06)	0.68(0.05)
$\rho = 16$	0.69(0.07)	0.72(0.07)	0.60(0.06)	0.56(0.04)	0.55(0.07)	0.59(0.08)
$\rho = 21$	0.63(0.08)	0.68(0.08)	0.36(0.09)	0.49(0.09)	0.56(0.08)	0.38(0.10)

Table 7: ARI of Stability measurement: Simulation Senario 3

Method	MC.D	MC.C	S1	S2	S3	S4
$\rho = 1$	0.92(0.01)	0.95(0.02)	0.93(0.01)	0.92(0.01)	0.92(0.02)	0.94(0.02)
$\rho = 6$	0.90(0.05)	0.91(0.05)	0.93(0.08)	0.93(0.04)	0.89(0.06)	0.81(0.08)
$\rho = 11$	0.82(0.05)	0.81(0.10)	0.70(0.05)	0.73(0.08)	0.72(0.07)	0.72(0.05)
$\rho = 16$	0.75(0.07)	0.69(0.08)	0.60(0.07)	0.56(0.07)	0.55(0.07)	0.60(0.08)
$\rho = 21$	0.62(0.10)	0.60(0.10)	0.38(0.11)	0.49(0.07)	0.47(0.09)	0.62(0.10)

3.3.2 Yeast Cell Cycle Data

3.3.2.1 Data Preprocessing This data set contain 4 studies described in section 3.2.5. Each study contains expression values for 6183 gene features and 18-24 samples. We use KNN [14] method to impute the missing values and use IQR selected genes with top rank sum of means and standard deviation.

3.3.2.2 Parameter Estimation We calculated prediction strength as described in method section. Figure 16 shows the prediction strength in the parameter space of four single studies. Then we average the *ps* value across studies and show in Figure 17. As figure shows, there is not a consistant global maximum prediction strength across all the studies.

Thus, we would like to use ICC estimate of number cluster. As figure **??** shows, when $k = 5$ there is clearly consensus clusters in heatmap and consensus index gain beyond this point does not change much. With prior information, yeast cell cycle related gene count is approximate 200. Thus we decided to choose cluster number $k = 5$ while set $\beta = 0.25$ to include around 250 genes as predictive genes.

3.3.2.3 Clustering Results Clustering result generated by each methods are shown with heatmaps in Figure 18. We can see although single studies have clearly cluster pattern, however, there is no consistancy of these clusters across studies according to results shown in Table 8. The both MC_C and MC_D generated clusters have less clear pattern compare to single study, but there still clearly clusters pattern. The reason clusters pattern in heatmaps from two meta-analysis is less clear may due to the heteogeniety of the data.

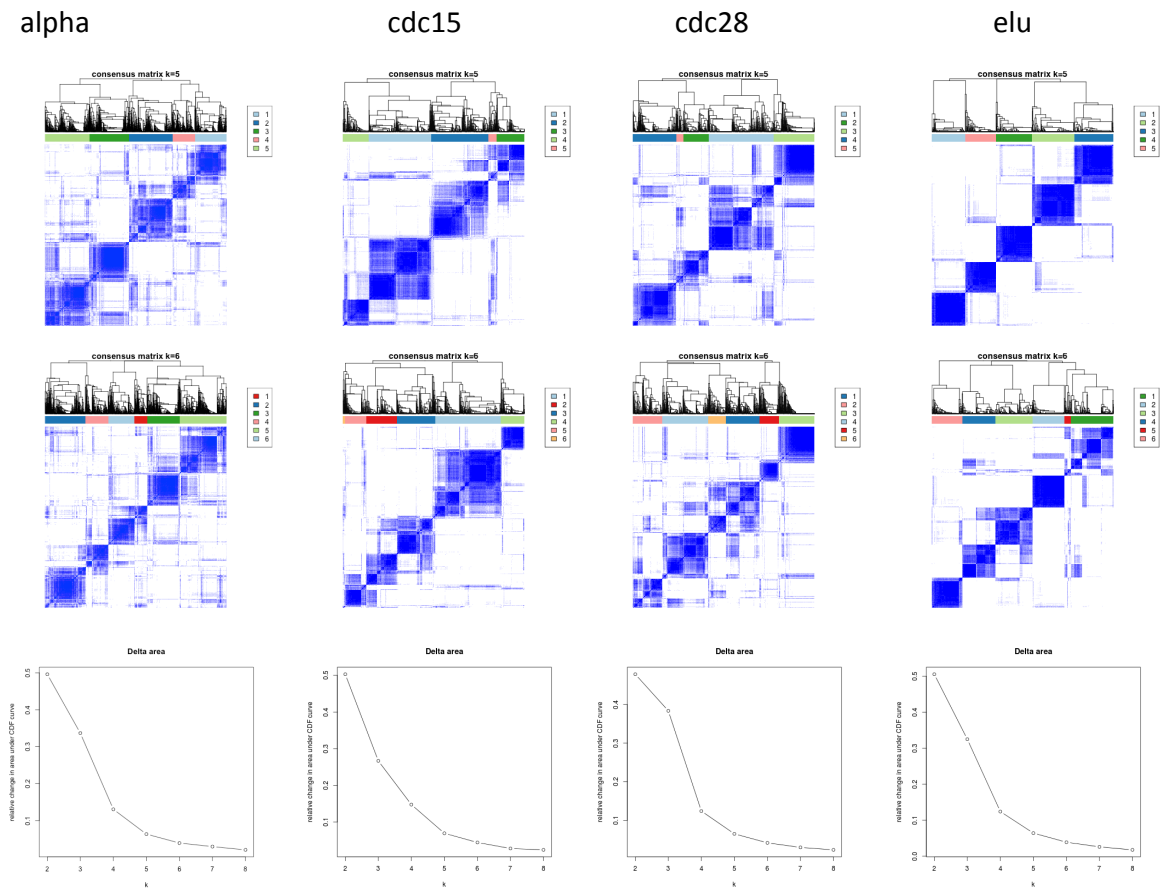

Figure 15: Consensus Clustering Result of Yeast Cell Cycle Data Set

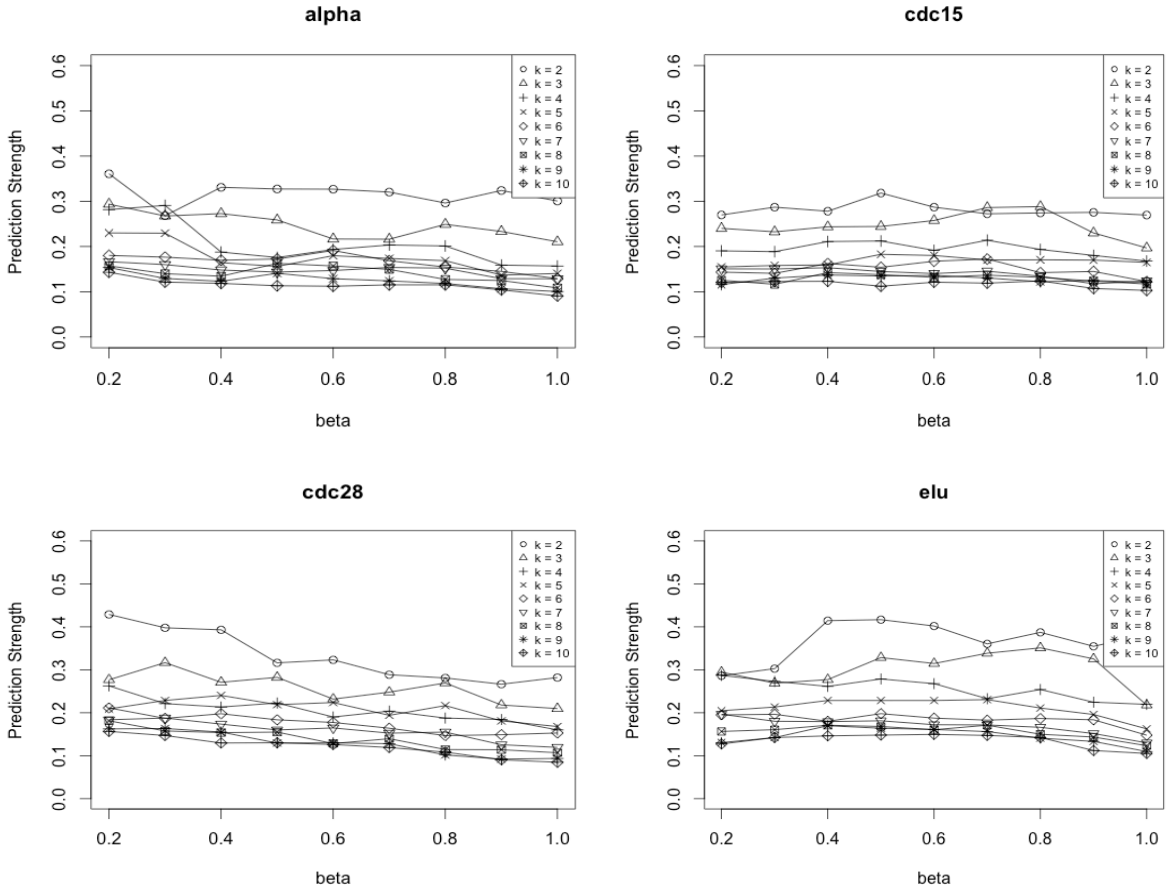

Figure 16: The Prediction Strength Estimated from Yeast Cell Cycle Dataset

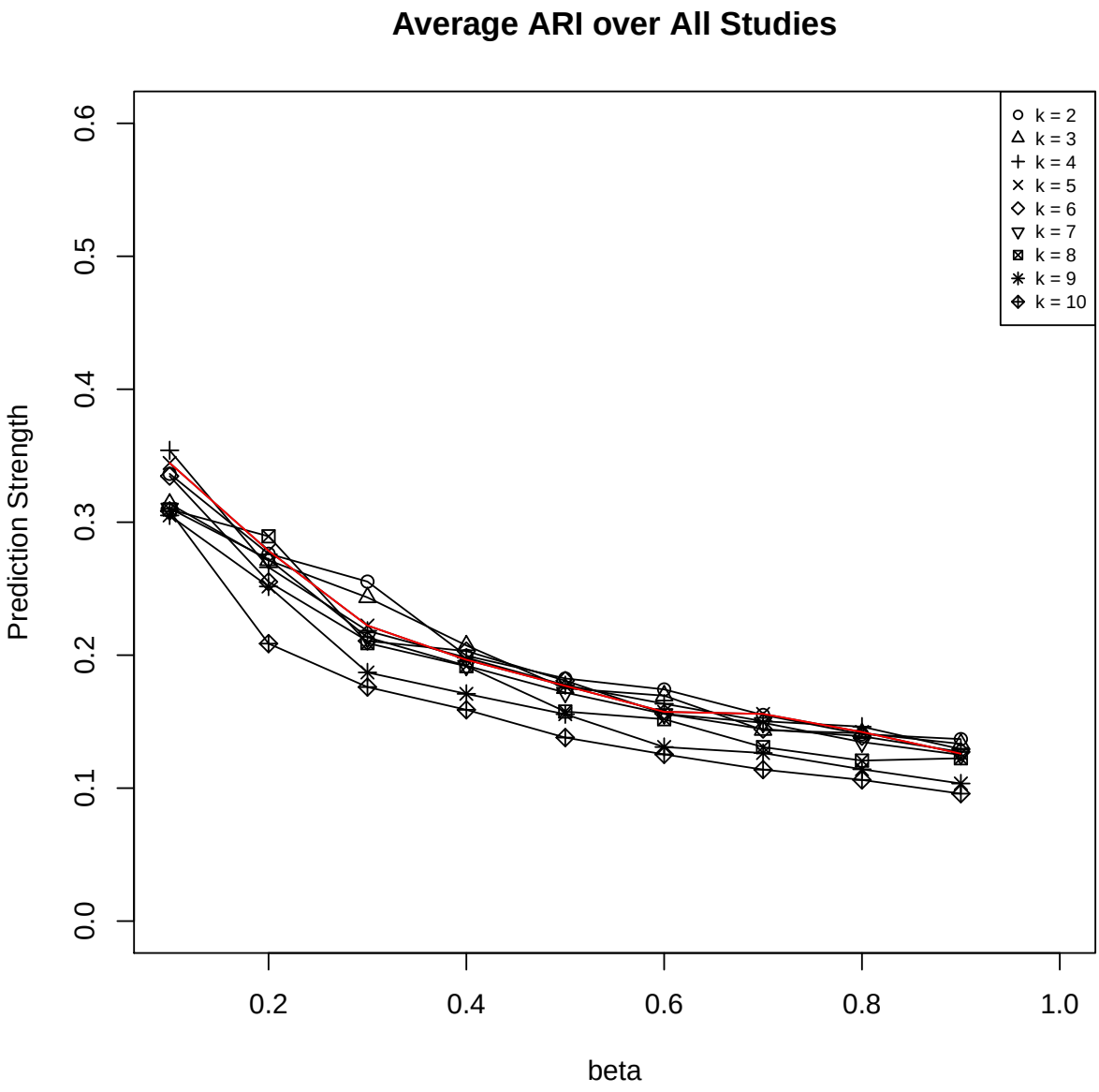

Figure 17: The Prediction Strength Estimated from Yeast Cell Cycle Dataset, averaged across studies

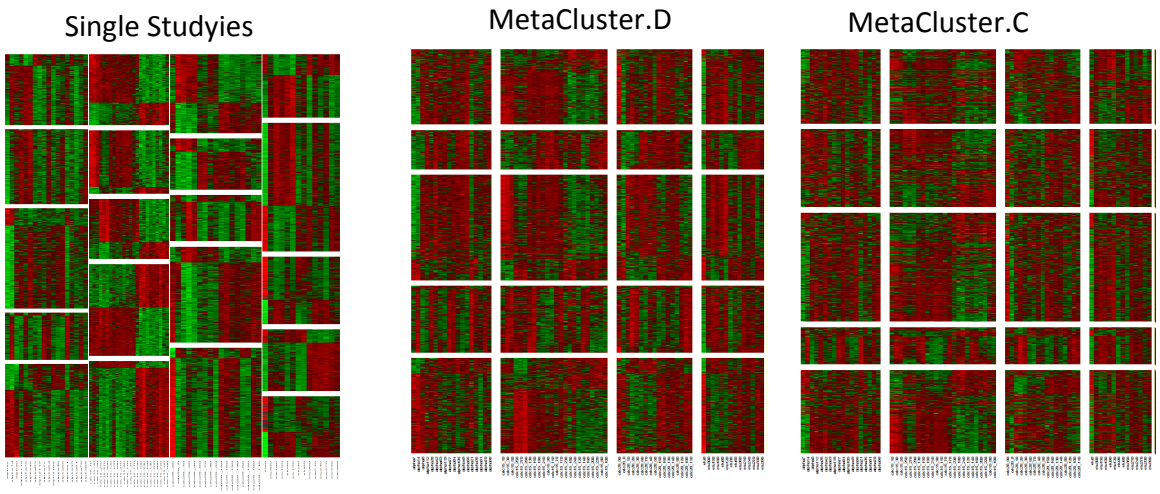

Figure 18: The Heatmaps of cluster result of Yeast Cell Cycle Dataset

3.3.2.4 Concordance across studies

Leave one study out analysis described in 3.3.1.3 is to mearsure the clustering result concordance across studies. ARIs result listed in table 8 shows gene modules detected by two meta frameworks with left one study out data sets can have more similarity to the left study result. Compare to the ARIs in simulation study and BRCA data set (result shown in section 3.3.3), the value of ARIs in this measurement is smaller. This is due to the concordance across single studies is lower which means different single studies conclude non-consistant gene co-expression modules.

3.3.2.5 Stability

Here we use bootstrapping on samples in each study, and compare bootstrapping cluster result to whole data set to mearsure the clustering result stability. The ARI value listed in the table 9 are from 100 times bootstrapping. Each bootstrap sample same numbers of samples with replacement. As table shows, MC_D outperform single studies, while MC_C have a relative high stability as well.

3.3.2.6 Biological Meanings

We conducted pathway enrichment analysis using Fisher's exact test, which set genes in the clusters as signal and 6 Yeast cell cycle related pathways as background (pathways information downloaded from KEGG database). As jitter plots in Figure 19 shows, there is very weak signals of Fisher's exact test p-value for all the methods clustered result. This indicates although we use MC_C and MC_D methods identified stable co-expression modules, the detection do not seem to associated with mechanism of yeast cell cycles.

Table 8: ARI of Concordance Across Studies measurement: Yeast Cell Cycle Data

Method	Alpha	Cdc15	Cdc28	Elu
Alpha	1			
Cdc15	0.032	1		
Cdc28	0.048	0.020	1	
Elu	0.006	0.022	0.016	1
MC.D(-alpha)	0.227			
MC.D(-cdc15)		0.288		
MC.D(-cdc28)			0.353	
MC.D(-elu)				0.243
MC.C(-alpha)	0.095			
MC.C(-cdc15)		0.173		
MC.C(-cdc28)			0.188	
MC.C(-elu)				0.191

Table 9: ARI of Stability measurement: Yeast Cell Cycle Data

Method	MC.D	MC.C	Alpha	Cdc15	Cdc28	Elu
ARI	0.385	0.368	0.257	0.363	0.239	0.299

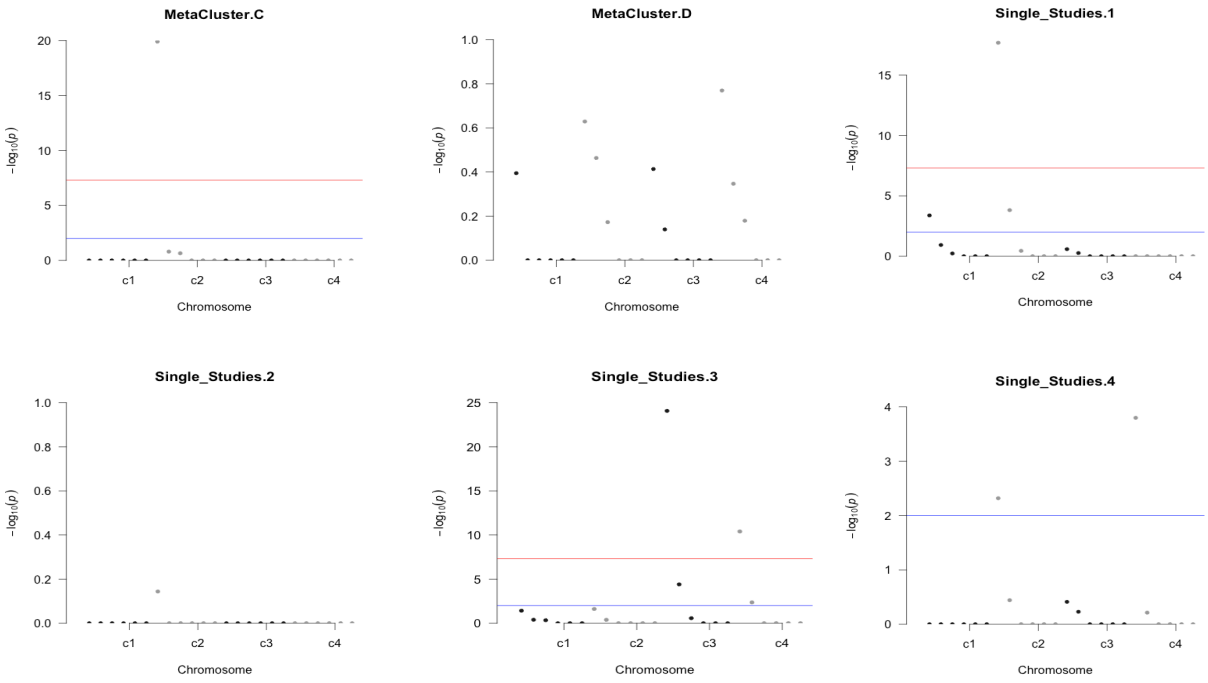

Figure 19: The Heatmaps of cluster result of Yeast Cell Cycle Dataset

3.3.3 Breast Cancer Data

3.3.3.1 Data Preprocessing Four studies of this data set is labeled as "BRCA", "GSE7390", "GSE2034",and "METABRIC" with description in section 3.2.5. Each study contains expression values for more than 22,000 gene features and 18-24 samples. We use KNN(citation) method to impute the missing values and use IQR selected top 1000 genes with means rank sum and standard deviation rank sum.

3.3.3.2 Parameter Selection Here we calculated (k, β) pairs related prediction strength. β is range from [0.1,1], k is range from [2,10]. Prediction strength value by parameters is shown in figure **??**. The general pattern of ps value decreases when k decreases and β increases. As there is no global maximum of ps in the parameter space, we followed suggestion in (Tseng 2012), to keep relativly large k and small β to achieve tight cluster. Therefore, we decided to choose $\beta = 0.2$ (200 genes as predictive) and cluster number $k = 4$.

Then we use consensus clustering to valid our estimation of number cluster. Here we calculated resample 80% samples from each study at a time, calculate consensus index and iterative for 50 times. As figure **??** shows, when $k = 4$ there is clearly consensus clusters in heatmap and consensus index gain beyond this point does not change much.

Thus, we decided to choose cluster number $k = 4$ which is consistent with prediction strength method result.

3.3.3.3 Clustering results After the clustering analysis, we matched the gene modules by the best gene overlapping through the different methods. As **??** shows, meta-analysis frame works keeps orginal module pattern within single studies. Moreoever, the meta-analysis methods, provide a consistent modules across studies.

3.3.3.4 Biological Meanings To evaluate the biological meaning of detected modules, we applied pathway enrichment analysis which perform using Fisher's exact test by testing association of selected intrinsic genes and genes in a particular pathway. We applied the Bio-

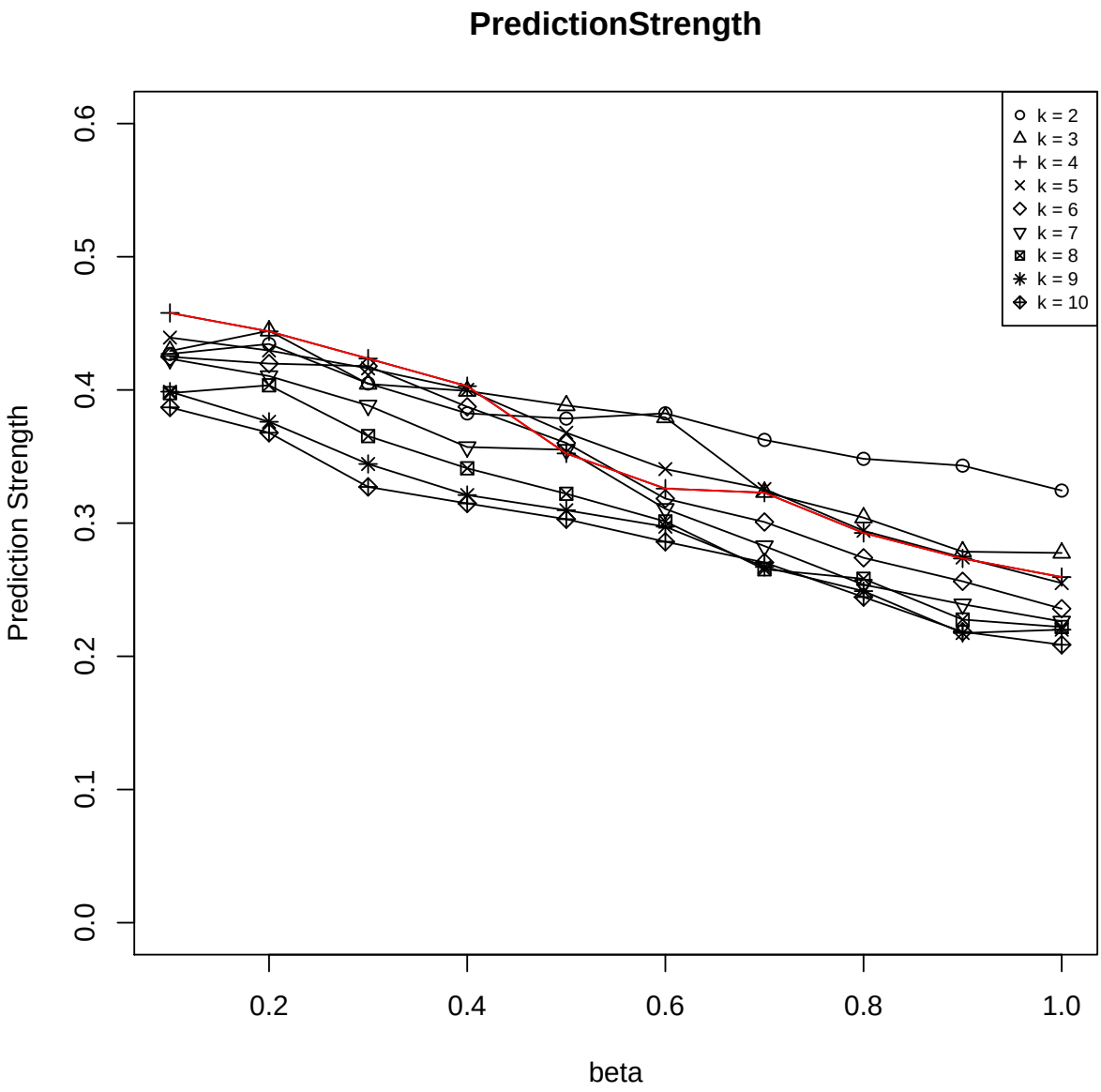

Figure 20: Prediction Strength of BRCA data set

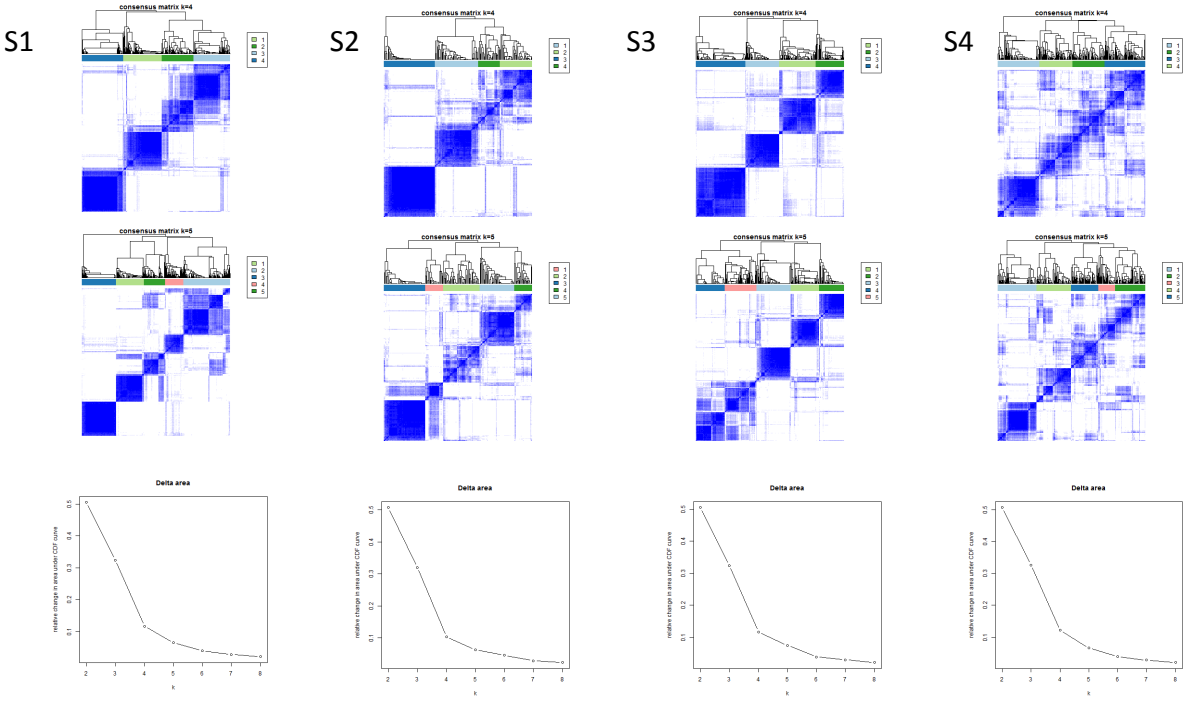

Figure 21: Consensus Clustering Result of BRCA Data Set

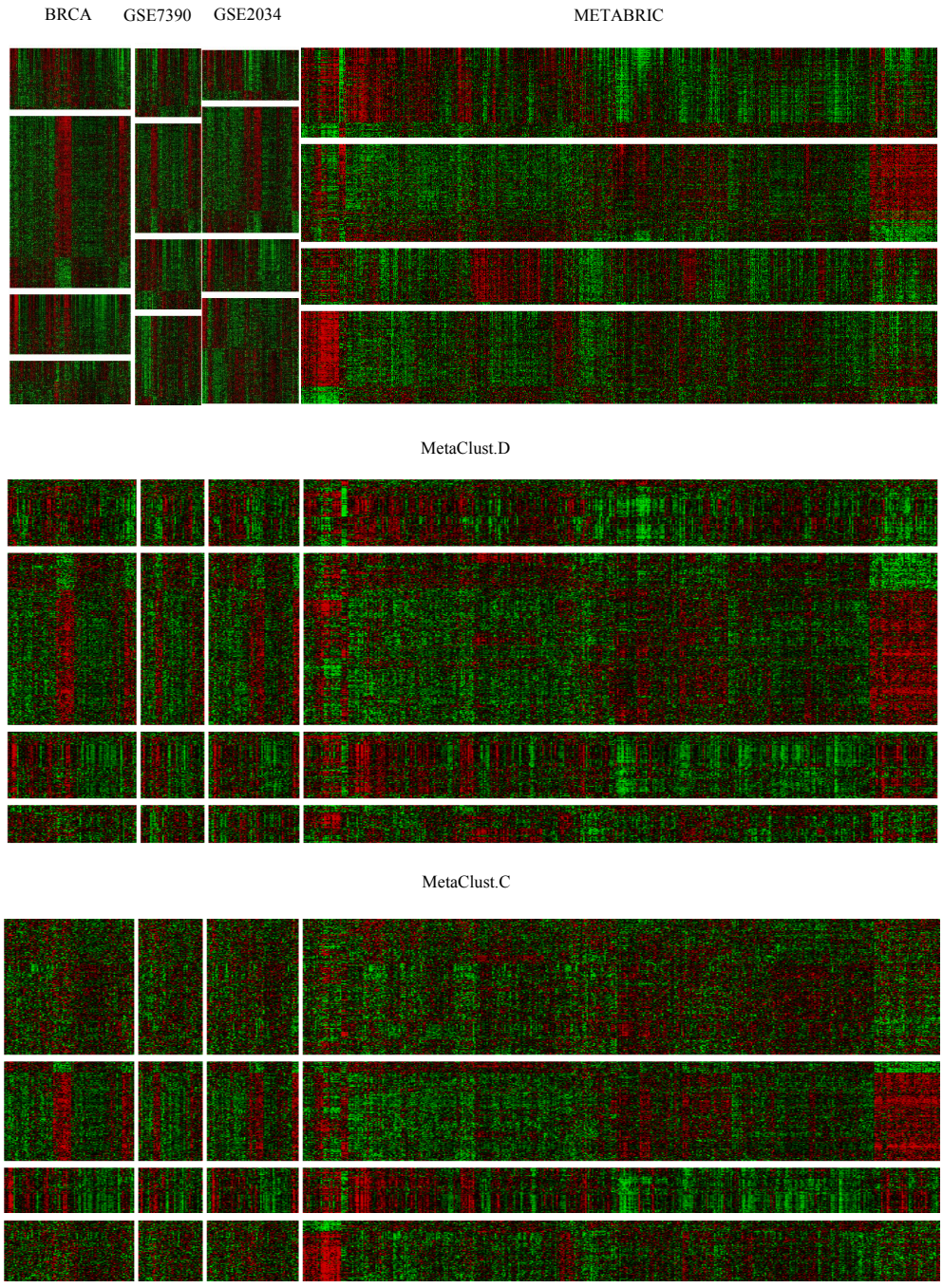

Figure 22: Heatmap of BRCA studies MetaClust.D, and MetaClust.C

Carta Database obtained from MSigDB (http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C2). This database contains 217 curated cancer related pathways and is particularly suited to evaluate the breast cancer example. Figure ?? shows the jitter plot pathway enrichment p-values at log-scale (base 10). The blue horizontal solid line corresponds to $p = 0.01$ signicant level threshold. The pathway enrichment result from meta analysis frameworks can recover most significant pathways in single studies.

Beside BioCarta database, we also applied pathway enrichment analysis on 1822 Gene Oncology (GO) pathways. Figure ?? shows the jitter plot pathway enrichment p-values at log-scale (base 10). The red horizontal solid line corresponds to the Bonferroni adjusted p = 0.05 signficant level threshold. Again, the pathway enrichment result from meta analysis frameworks can recover most significant pathways in single studies. MC_D detected cluster 4 are more significant according to Fisher's exact test.

3.3.3.5 Concordance Across Studies Here we use leave one study out at a time and compare to the single study to mearsure the clustering result concordance across studies. ARIs result listed in table 10 shows gene modules detected by two meta frameworks with left one study out data sets can have more similarity to the left study result.

3.3.3.6 Stability Here we use bootstrapping on samples in each study, and compare bootstrapping cluster result to whole data set to mearsure the clustering result stability. The ARI value listed in the table below are from 100 times bootstrapping. Each bootstrap sample same numbers of samples with replacement.Each bootstrap sample same numbers of samples with replacement. As table 11 shows, MC_C outperform single studies, while MC_D have a relative high stability as well.

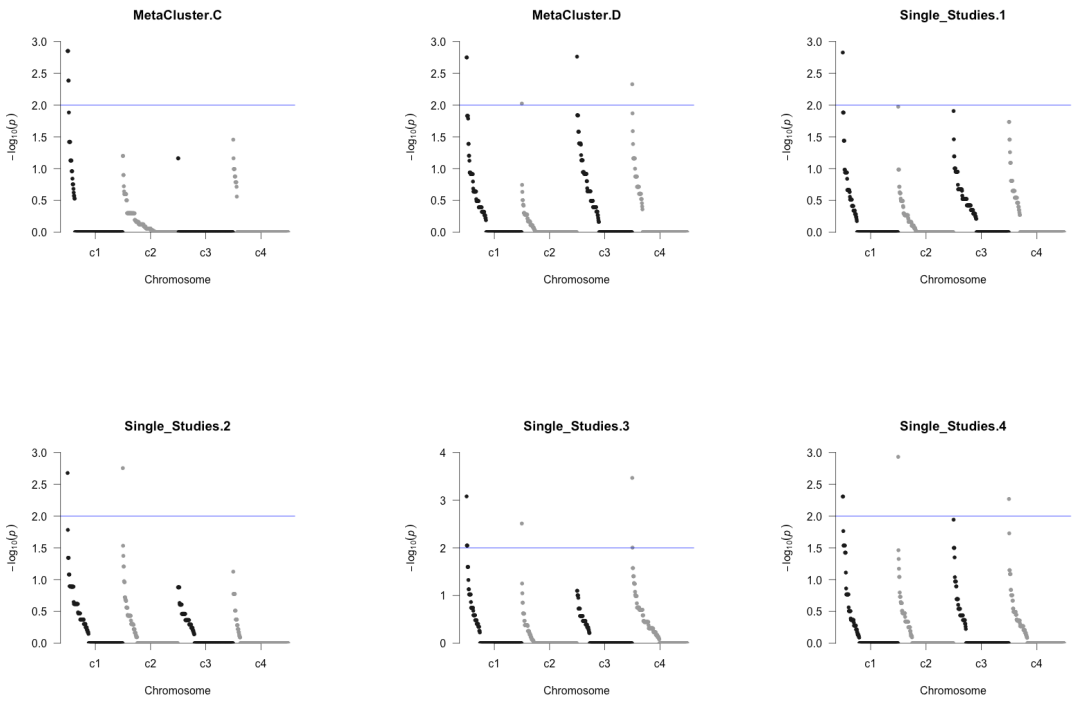

Figure 23: Jitter Plot of BioCarta Pathways in BRCA studies

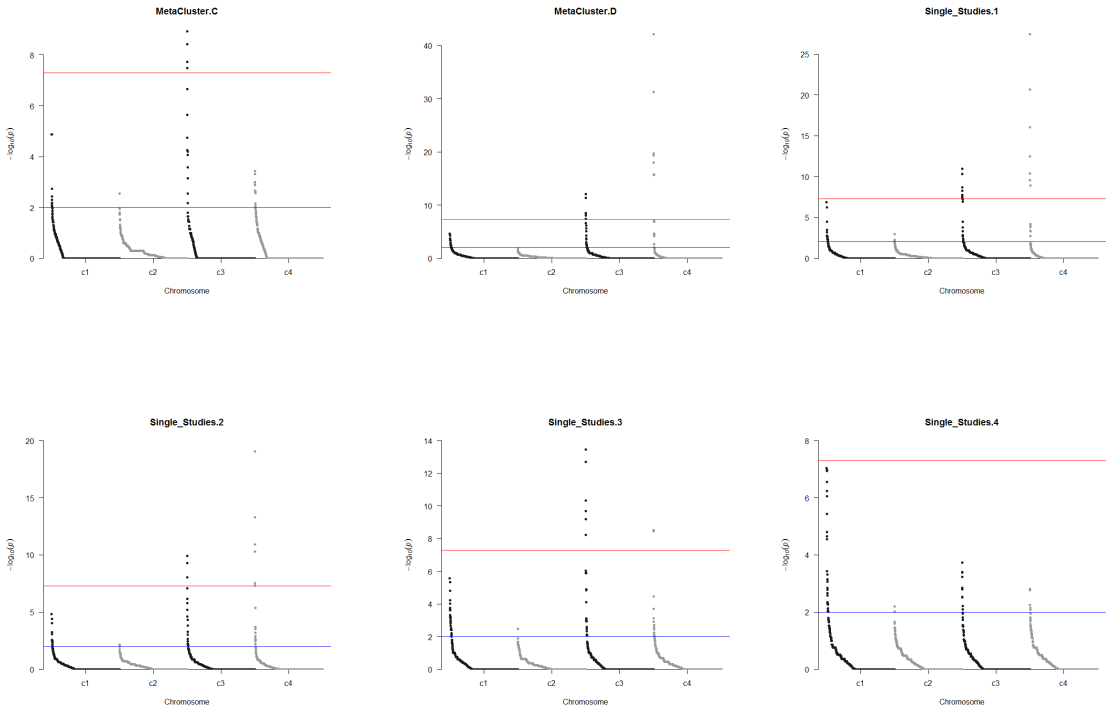

Figure 24: Jitter Plot of GO Pathways in BRCA studies

Table 10: ARI of Concordance Across Studies measurement: Breast Cancer Data

Method	BRCA	GSE7390	GSE2034	METABRIC
BRCA	1			
GSE7390	0.415	1		
GSE2034	0.560	0.423	1	
METABRIC	0.300	0.308	0.319	1
MC.D(-BRCA)	0.753			
MC.D(-GSE7390)		0.830		
MC.D(-GSE2034)			0.836	
MC.D(-METABRIC)				0.781
MC.C(-BRCA)	0.758			
MC.C(-GSE7390)		0.719		
MC.C(-GSE2034)			0.900	
MC.C(-METABRIC)				0.471

Table 11: ARI of Stability measurement: Breast Cancer Data

Method	MC.D	MC.C	BRCA	GSE7390	GSE2034	METABRIC
ARI	0.527	0.561	0.486	0.509	0.398	0.368

3.3.4 Mouse Metabolism Data

3.3.4.1 Data Preprocessing Four studies of this data set are based on sample from different organ of mouse, which are labeled as "brown", "heart", "liver",and "skeleton" with description in section 3.2.5. Each study contains expression values for more than 14,495 gene features and 9-12 samples. We use KNN(citation) method to impute the missing values and use IQR selected top 1000 genes with means rank sum and standard deviation rank sum.

3.3.4.2 Parameter Selection Here we calculated (k, β) pairs related prediction strength. β is range from [0.1,1], k is range from [2,10]. According to figure ??, there is no global maximum among parameter space across all the studies. "Liver" study with $k = 4$ shows a local maximum of prediction strength value. As we average prediction strength across study at same parameters, shown in figure ??, there is neither global maximum nor obvious drop of ps. In this case, we decide to use ICC to decide cluster number first.

Here we calculated resample 80% samples from each study at a time, calculate consensus index and iterative for 50 times.

As figure ?? suggested, we decided to choose cluster number $k = 5$. Based on this decision, we decided to choose $\beta = 0.4$ (400 genes as predictive) according to figure ??.

3.3.4.3 Clustering Result Figure ?? shows the result of single studies, MC_D, and MC_C. Since we observe that sample LCAD.3 in liver tissue study shows an outlaid pattern with other samples and this pattern dominate the prediction of modules, we decided to exclude LCAD.3 sample from study "liver". The clustering result excluding sample LCAD.3 from study "liver" is shown in figure ??.

3.3.4.4 Concordance Across Studies Again, we used same approach to estimate concordance across studies as which applied to previous data sets. As mentioned in Yeast Cell Cycle data analysis, the low between studies similarity of detected modules affect the performance of both MC_C and MC_D. But still, MC_D have higher ARI compare to that measures similarity between single studies.

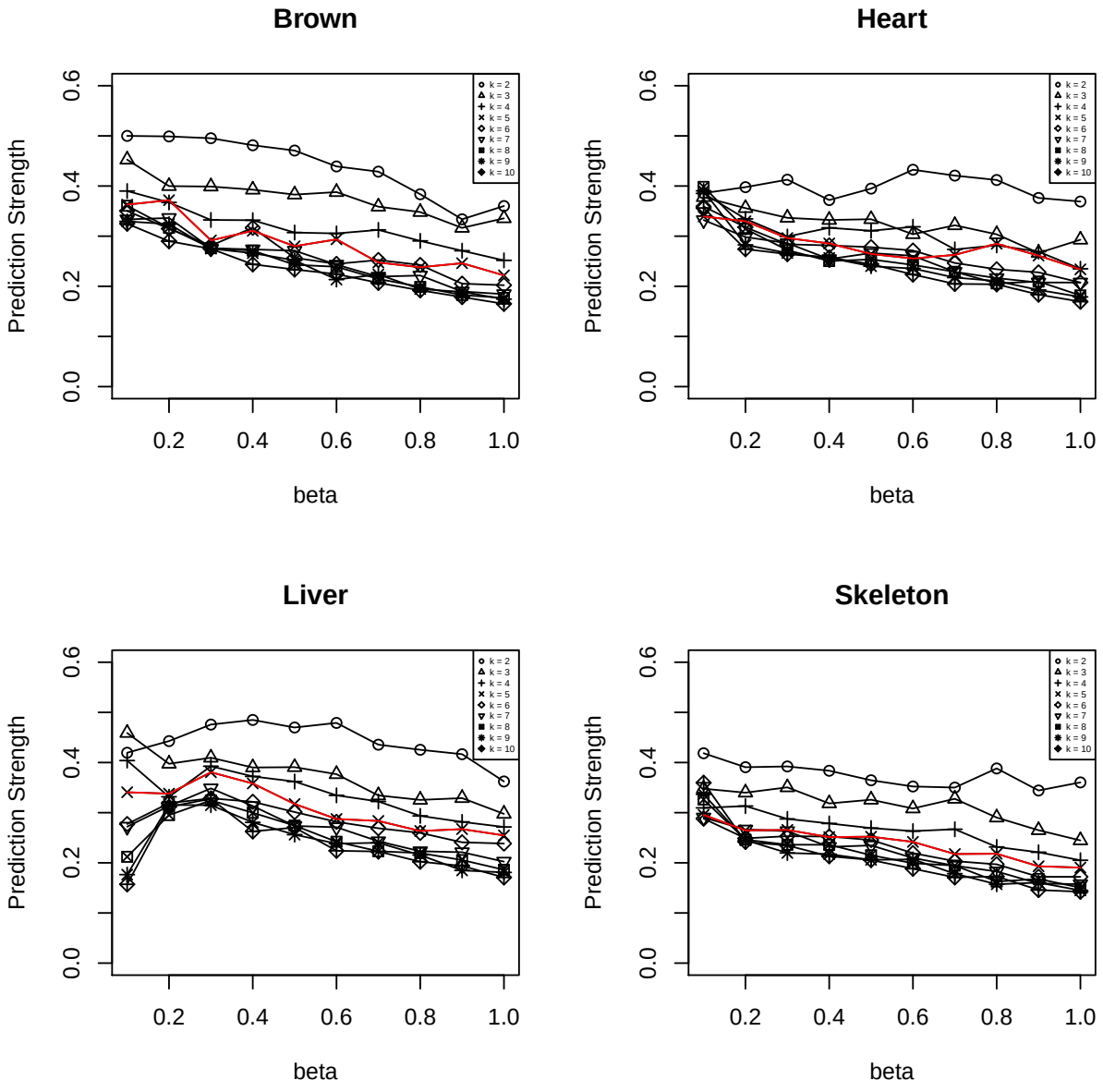

Figure 25: ARI of Stability measurement: Breast Cancer Data

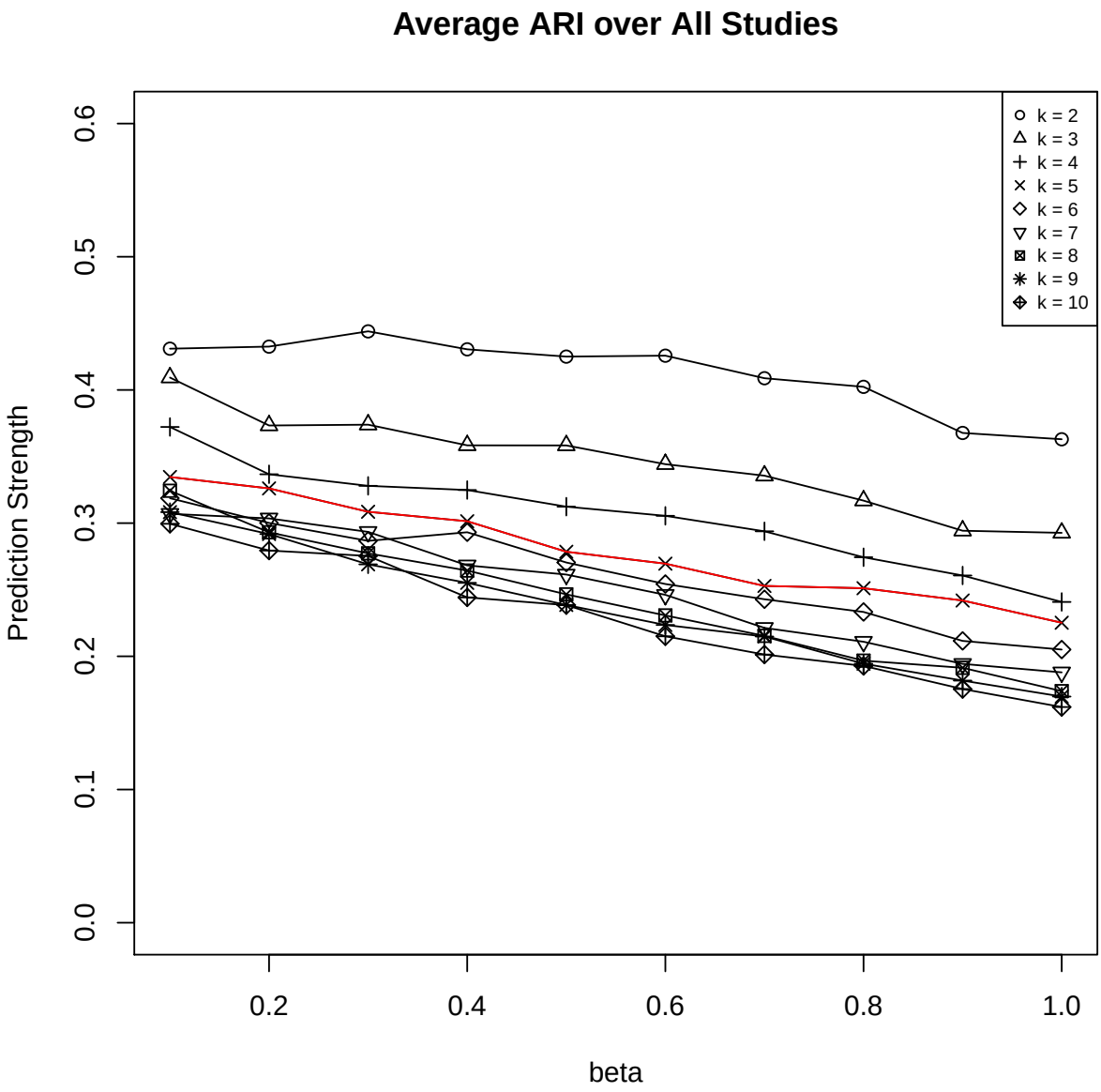

Figure 26: ARI of Stability measurement: Breast Cancer Data

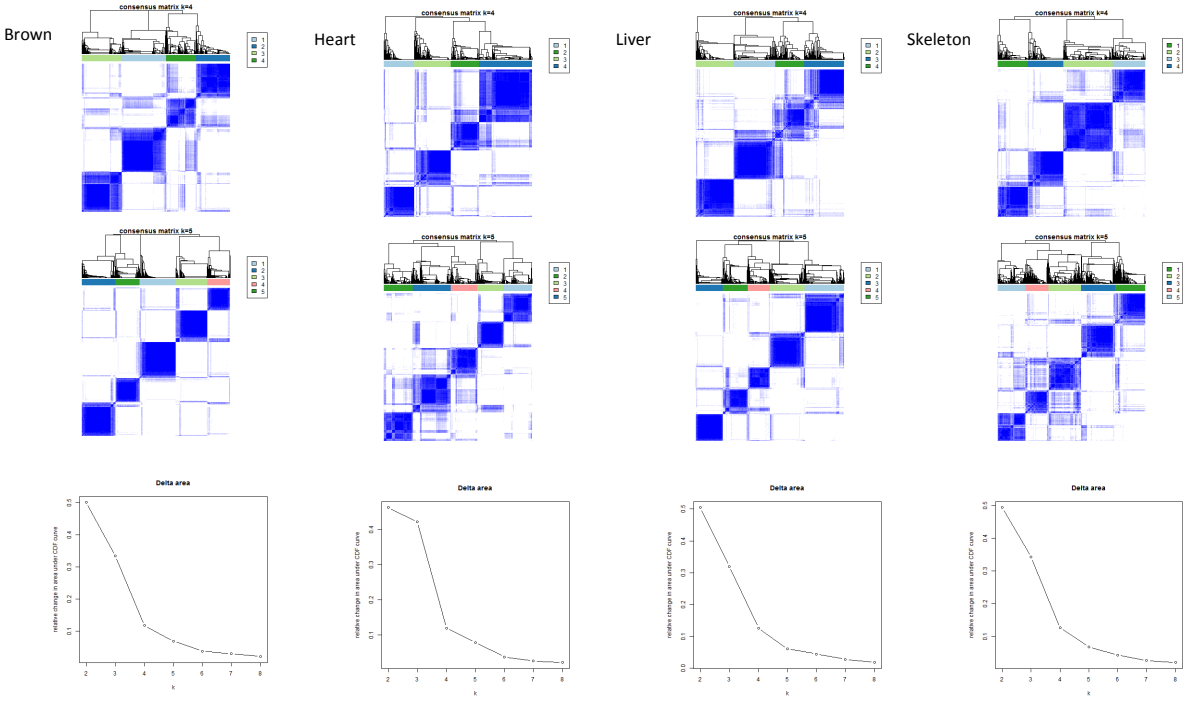

Figure 27: ARI of Stability measurement: Breast Cancer Data

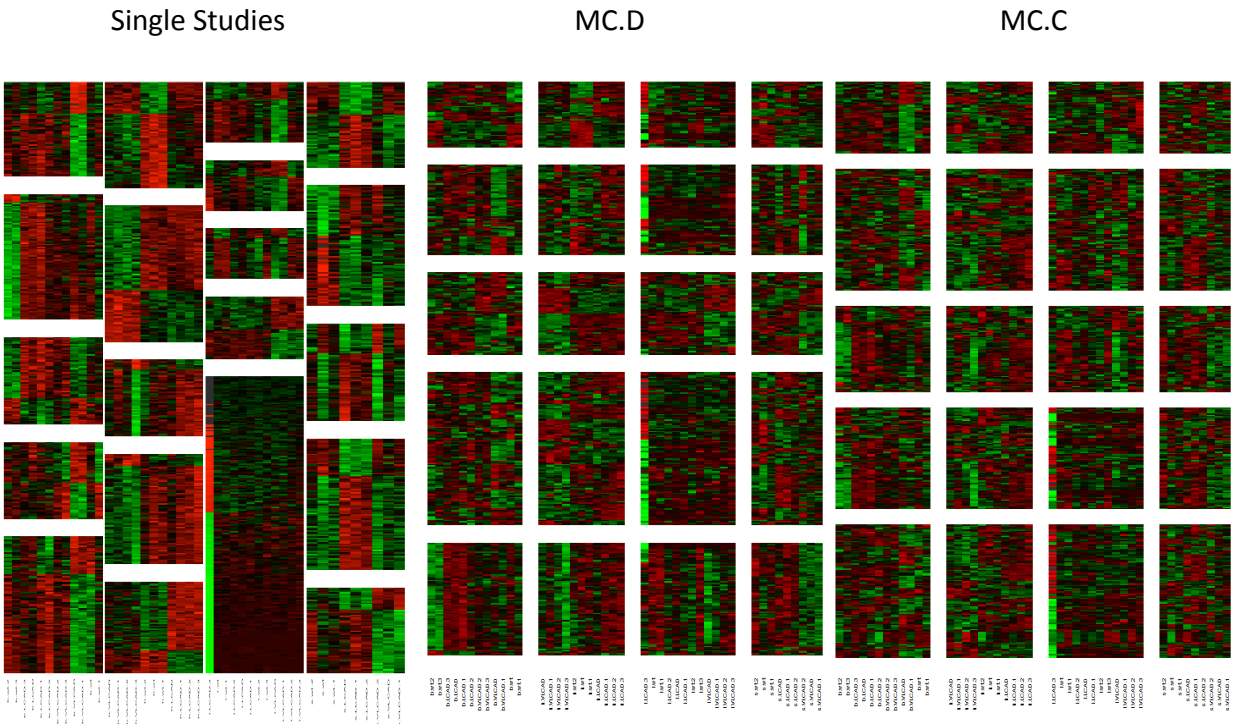

Figure 28: Heatmap of Mouse Metabolism Data

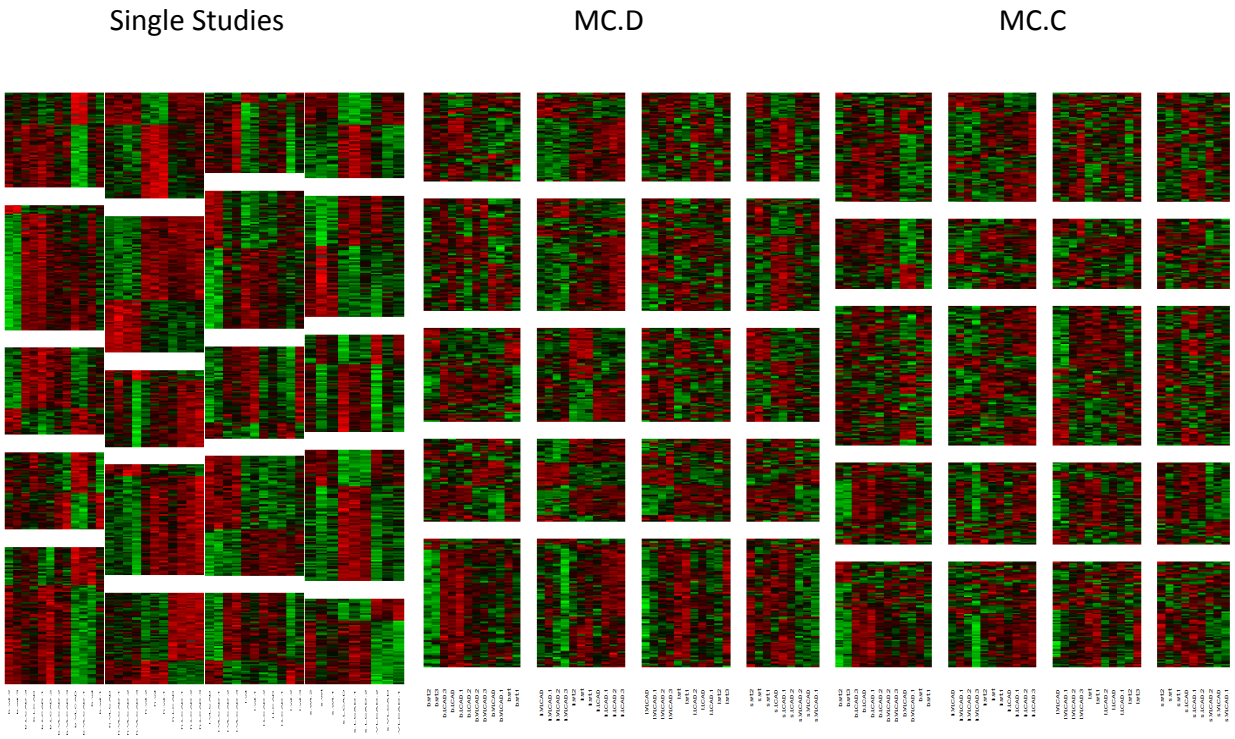

Figure 29: Heatmap of Mouse Metabolism Data: exclude sample LCAD.3

3.3.4.5 Stability Here we use bootstrapping on samples in each study, and compare bootstrapping cluster result to whole data set to mearsure the clustering result stability. The ARI value listed in the table below are from 100 times bootstrapping. Each bootstrap sample same numbers of samples with replacement. With this dataset, stability measurement among all the methods have relative small standard deviation. This is due to the small sample size in original data.

3.4 CONCLUSION

As the result we shown in 3 simulation data sets and 4 real data sets, meta-analysis frameworks proposed, MC_C and MC_D, perform better in consideration of concordance across studies and stability. Under the simulation settings, MC_C combining the maximum distance over all studies have better recovery of underline truth, while MC_D combining by average distance have better stability. Among real data analysis, MC_C and MC_D have relative same stability and concordance. Whether two proposed meta-analysis frameworks' detected co-expression modules are biological meaningful, is highly depend on the single study detected module consistance across studies.

Table 12: ARI of Concordance Across Studies Measurement on Mouse Metabolism Data: exclude sample l.LCAD.3

Method	Brown	Heart	Liver	Skeleton
Brown	1			
Heart	0.058	1		
Liver	0.148	0.057	1	
Skeleton	0.026	0.010	0.046	1
MC.D(-Brown)	0.436			
MC.D(-Heart)		0.447		
MC.D(-Liver)			0.573	
MC.D(-Skeleton)				0.551
MC.C(-Brown)	0.221			
MC.C(-Heart)		0.485		
MC.C(-Liver)			0.188	
MC.C(-Skeleton)				0.341

Table 13: ARI of Stability Measurement on Mouse Metabolism Data

Method	MC.D	MC.C	Brown	Heart	Liver	Skeleton
ARI	0.52(0.01)	0.48(0.01)	0.50(0.01)	0.42(0.02)	0.46(0.01)	0.47(0.01)

4.0 SUMMARY

Large-scale meta-analysis of genomic studies is becoming increasingly common, because it is now more feasible and there are a greater volume of data sets available. In this dissertation, I have addressed these issues as follows:

In chapter 2, I proposed ChIP-MetaCaller which provide a meta-analysis framework of combining ChIP-seq peak signals. In the sense of sensitivity, the proposed method are also recover more signals included in ChIP-chip analysis detection. In terms of enrichment of motifs, meta-analysis combining result from MACS, SISSRs, and Useq provide a higher sensitivity peak list; in specific dataset, the optimal combination of callers may differ, user may use caller selection strategies provide by this study to explore.In conclusion, we suggested to perform the ChIP-MetaCaller with combining results from MACS, SISSRs, and Useq. And search the motif pattern within top ranked peak regions. And with a known motif pattern dataset, user may use forward selection to obtain an optimal combination of callers' result to achieve better discovery.

In chapter 3, we proposed two approaches to integrate genomic expression profiles and compared these meta-analysis frameworks with single studies in sense of stabiliry, concordance across studies and biological meanings. And suggested that both MC_C with average distance and MC_D with maximum distance provide a stable and consistant way to integrate expression profiles datasets.

BIBLIOGRAPHY

[1] B.M. Bolstad, R.A Irizarry, M. strand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[2] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. 1994.

[3] LC Chang, S. Jamain, CW Lin, D. Rujescu, G. C. Tseng, and E. Sibille. A conserved BDNF, glutamate- and GABA-Enriched gene module related to human depression identified by coexpression meta-analysis and DNA variant genome-wide association studies. *PLoS ONE*, 9(3):e90980, 2014.

[4] N.G. de Bruijn. A combinatorial problem. *Nederl. Akad. Wetensch. Proc. Ser.A*, 49:pp. 758764, 1946.

[5] R. Dobrin, J. Zhu, C. Molony, C. Argman, M. L. Parrish, S. Carlson, M. F. Allan, D. Pomp, E. E. Schadt, et al. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol*, 10(5):R55, 2009.

[6] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA*, 100:8817–8822, 2003.

[7] L. L. Elo, H. Järvenpää, M. Orešič, R. Lahesmaa, and T. Aittokallio. Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process. *Bioinformatics*, 23(16):2096–2103, 2007.

[8] EMBL-EBI. Arrayexpress.

[9] Y. Erlich, P. P. Mitra, M. delaBastide, W. R. McCombie, and G. J. Hannon. Alta-cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods*, 5:679–682, 2008.

[10] P. J. Farnham. Insights from genomic profiling of transcription factors. *Nature Rev. Genet.*, 10:605–616, 2009.

[11] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34:e22, 2006.

[12] C. Gaiteri, Y. Ding, B. French, G. C. Tseng, and E. Sibille. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, Brain and Behavior*, 13(1):13–24, 2014.

[13] C. Gaiteri, JP Guilloux, D. A. Lewis, and E. Sibille. Altered gene synchrony suggests a combined hormone-mediated dysregulated state in major depression. *PloS one*, 5(4):e9970, 2010.

[14] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu. *impute: impute: Imputation for microarray data*. R package version 1.36.0.

[15] H. Ji. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotech.*, 26:1293–1300, 2008.

[16] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucleic Acids Res.*, 36:5221–5231, 2008.

[17] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotech.*, 26:1351–1359, 2008.

[18] C. K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genet.*, 36:900–905, 2004.

[19] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. 454(7205):766–770.

[20] M. L. Metzker. Sequencing technologies [mdash] the next generation. 11(1):31–46.

[21] NCBI. Geo.

[22] D. A. Nix, S. J. Courdy, and K. M. Boucher. Empirical methods for controlling false positives and estimating confidence in ChIP-seq peaks. *BMC Bioinformatics*, 9:523, 2008.

[23] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.*, 10:669–680, 2009.

[24] V. K. Rakyan, H. Beyan, T. A. Down, Mohammed I. H., S. Maslau, D. Aden, A. Daunay, F. Busato, C. A. Mein, B. Manfras, KR. M. Dias, C. G. Bell, J. Tost, B. O. Boehm, S. Beck, and R. D. Leslie. Identification of type 1 diabetesassociated dna methylation variable positions that precede disease diagnosis. *PLoS Genet*, 7(9):e1002300, 09 2011.

[25] J. Rozowsky. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotech.*, 27:66–75, 2009.

[26] K. Shen and G. C. Tseng. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323, 2010.

[27] G. C. Tseng. Penalized and weighted k-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics*, 23(17):2247–2255, August 2007.

[28] G. C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799, 2012.

[29] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.*, 10:57–63.

[30] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23:500–501, 2007.

[31] Y. P. Yu, Y. Ding, R. Chen, S. G. Liao, BG. Ren, A. Michalopoulos, G. Michalopoulos, J. Nelson, G. C. Tseng, and JH. Luo. Whole-genome methylation sequencing reveals distinct impact of differential methylations on gene transcription in prostate cancer. *The American Journal of Pathology*, 183(6):1960 – 1970, 2013.

[32] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18:821–829, 2008.

[33] Y. Zhang. Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, 9:R137, 2008.