Fall 12-11-2023

Advancements in Cancer Genomics: Graph-Based Motif Discovery and Single-Cell Analysis

Sayed Hossein Saghaeiannejad Esfahani

Advancements in Cancer Genomics: Graph-Based Motif Discovery and Single-Cell Analysis

by

Sayed Hossein Saghaeiannejad Esfahani

Under the Direction of Alexander Zelikovsky, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computer Science

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Studying biology of cancer cells enables us to understand how disease is growing and leads to new methods of diagnosing and treatment of many types of cancers. Over the past decades, researchers have surveyed multiple features of cancer cells such as genetic alteration in tumors, mutational patterns, copy number changes, and transcription factor binding sites. For this reason, scientists have employed Next Generation Sequencing (NGS) as it enables sequencing of thousands of DNA molecules. In this thesis, we aim to design and apply effective algorithms for interpreting and analysing cancer genomics data using NGS technique. In particular, we take advantage of microbiome RNA sequencing to investigate transcription factor binding sites of the genome, known as motifs. A new method for motif discovery is introduced and tested on synthetic and real data. Along with motif discovery method, a new approach for inferring haplotype-specific copy numbers among single cell sequences is presented. The inferred haplotype-specific copy numbers lead to inferring tumor clones and corresponding phylogenetic tree of these clones.

INDEX WORDS: Phage display library, Next Generation Sequencing, Microbiome, Regulatory motif, Graph theory, Single Cell Sequencing, Copy Number Aberrations, Loss of Heterozigosity, Whole Genome Duplication, Dijkstra algorithm, Integer Linear Programming

Advancements in Cancer Genomics: Graph-Based Motif Discovery and Single-Cell Analysis

by

Sayed Hossein Saghaeiannejad Esfahani

Committee Chair:	Alexander Zelikovsky
Committee:	Pavel Skums
	Murry Patterson
	Yurij Ionov

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2023

DEDICATION

To my parents,

In every page of this journey, in every challenge I faced, and in every triumph I celebrated, your love, wisdom, and unwavering belief in me have been my constant guide. This accomplishment is not just a testament to my efforts, but a reflection of your sacrifices, support, and endless encouragement. To you, I owe not just this achievement, but my enduring gratitude and love.

With all my heart,

Hossein

ACKNOWLEDGMENTS

I would like to take this opportunity to express my deepest gratitude to the individuals who have played a significant role in the completion of my PhD dissertation. Their guidance, support, and expertise have been invaluable to me throughout this academic endeavor.

First and foremost, I extend my heartfelt thanks to my advisor, Dr. Pavel Skums. Dr. Skums has been an unwavering source of guidance and mentorship. His profound knowledge, dedication, and insightful feedback have not only shaped the direction of my research but also nurtured my growth as a researcher. I am truly fortunate to have had Dr. Skums as my mentor.

I would also like to express my sincere appreciation to Dr. Alex Zelikovsky for his invaluable assistance and contributions to my research. Dr. Zelikovsky's expertise and willingness to engage in intellectual discussions have enriched my work and broadened my horizons.

Furthermore, I extend my gratitude to Dr. Yurij Ionov for his guidance throughout the dissertation process, particularly in the challenging motif discovery project. Dr. Ionov's expertise in this area has been instrumental in advancing my research, and I am thankful for his support.

Last but not least, I want to acknowledge Dr. Murry Patterson for his unwavering support and assistance. Dr. Patterson's encouragement and valuable insights have been a source of motivation and clarity during my academic journey.

I am also thankful to my colleagues, friends, and family for their encouragement and understanding during the ups and downs of this challenging journey.

This dissertation would not have been possible without the collective support and expertise of these individuals. I am humbled by their contributions, and I am grateful for the opportunity to have worked with such outstanding mentors and colleagues.

I would like to extend my heartfelt appreciation to a broader circle of individuals who have been instrumental in my academic journey.

To my friends in the department, your camaraderie, support, and collaboration have been invaluable. I am especially grateful to those who collaborated with me on various projects, sharing their expertise and insights. Your contributions have enhanced the quality of my research and made the academic experience richer.

I also owe a debt of gratitude to my family. To my parents, your unwavering support, encouragement, and belief in my abilities have been the cornerstone of my academic pursuit. Your sacrifices and dedication to my education have not gone unnoticed, and I am deeply thankful for everything you have done for me.

To my siblings, your encouragement and understanding during the challenges of my academic journey have been a constant source of strength. Your belief in me has been a driving force in my pursuit of knowledge.

I am truly fortunate to have such a wonderful network of friends and a loving family. Thank you for being a crucial part of this journey and for standing by me through every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

This dissertation explores the realm of cancer genomics, leveraging advanced computational methodologies and innovative approaches. The primary theme focuses on the integration of cutting-edge sequencing technologies, graph theory, and statistical analysis to shed light on crucial aspects of cancer biology. We have organized this chapter into two main sections: "Technology" and "Bio-informatics Challenges." Each section will include subsections to cover the relevant information about NGS sequencing, the phage display technique, and single-cell sequencing, as well as the challenges associated with these technologies and how our research addresses them.

1.1 Technology

1.1.1 Next Generation Sequencing

In the era of modern high-throughput sequencing technologies, advancements in cancer genomics have taken center stage. Techniques such as Illumina sequencing, Roche 454 sequencing, Ion Torrent, and SOLiD sequencing have revolutionized our ability to decode the complicated genetic landscapes underlying cancer progression. The NGS workflow involves a sequence of steps:

1.1.1.1 Library Preparation

DNA or RNA samples are broken into smaller pieces and equipped with adapters at their ends. These adapters contain sequences that enable the fragments to attach to a solid surface.

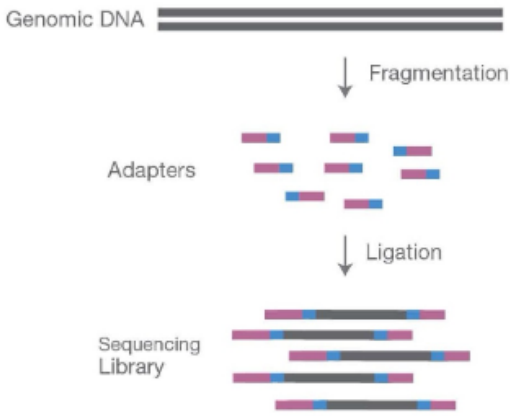

Figure 1.1 The 1st step: library preparation CD Genomics (2023)

1.1.1.2 Clonal Amplification

The fragments are amplified to produce clusters of identical sequences, employing either bridge amplification (in Illumina sequencing) or emulsion PCR (in Roche 454 sequencing).

1.1.1.3 Sequencing

Each cluster undergoes a sequencing-by-synthesis process, where fluorescently labeled nucleotides are added sequentially. The emitted fluorescence is recorded and employed to determine the order of nucleotides in the DNA fragment.

1.1.1.4 Data Analysis

The raw sequencing data is processed to remove errors, align sequences to a reference genome, and identify genetic variations, along with other analyses.

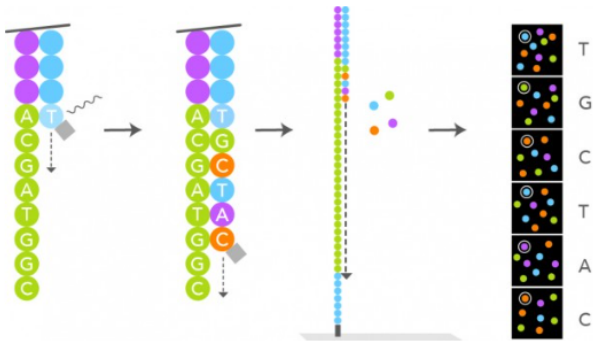

Figure 1.2 Addition of fluorescently labelled nucleotide and identifying the fluorophore Lexogen (2023)

1.1.2 Phage Display library

An essential aspect of this research involves the utilization of the phage display technique, a powerful tool for the study of protein-protein, protein-peptide, and protein-DNA interactions. Specifically, it provides a unique perspective for examining mimotope profiles within the framework of the immune system's response to cancer. This dissertation unveils the application of graph-based motif discovery methods to understand these interactions, contributing to a deeper understanding of immune system dynamics in cancer.

Phage display is a powerful laboratory method for exploring diverse molecular interactions. It involves displaying peptides, proteins, or antibodies on the surface of bacteriophage viruses. By displaying these molecules on the phage's exterior, researchers can isolate and amplify specific phage clones, thereby identifying molecules that bind to a desired target.

In phage display, mimotopes are peptide sequences displayed on phage surfaces that mimic the binding patterns of target proteins. NGS is employed to sequence these mimotopes, generating vast data sets of sequences. This data is crucial for understanding the

preferences and interactions of these peptides with specific targets.

Phage display library technique involves the following steps:

1. Phage Display Library Generation:

This step involves creating a diverse library of phages each displaying a different protein or peptide on their surface. Scientists insert variable DNA sequences, which code for the proteins of interest, into plasmids. These plasmids also carry a gene for a phage coat protein, an antibiotic resistance gene, and a packaging signal. The plasmids are then introduced into a bacterial vector (commonly E. coli) through a process called transformation. The bacteria are then infected with a helper phage, which provides the additional viral proteins necessary for the assembly of new phage particles. Each resulting phage displays a version of the protein of interest on its coat protein and carries the corresponding genetic sequence in its genome.

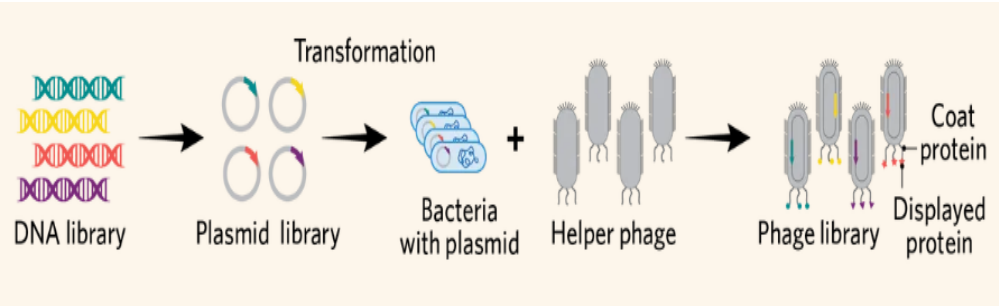

Figure 1.3 Phage library generation Doe (2023)

2. Affinity Purification:

In this step, researchers isolate phages that express surface proteins with a specific affinity for a target, typically using a surface ligand binding assay. The target (or ligand) is immobilized on a solid surface, and the phage library is introduced. Phages with surface

proteins that bind to the target are retained, while non-binders are washed away. The bound phages are then eluted, which may contain a mix of different surface protein epitopes that showed affinity for the target.

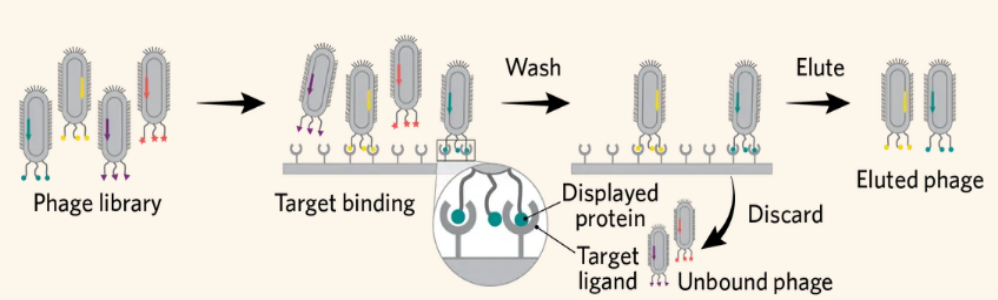

Figure 1.4 Affinity purification Doe (2023)

3. Phage Amplification:

The eluted phages from the affinity purification step are propagated in a fresh bacterial culture. This step allows for the multiplication of phages that successfully bound to the target. Researchers may repeat the affinity purification and amplification steps multiple times to enrich the population of phages that have the highest affinity for the target.

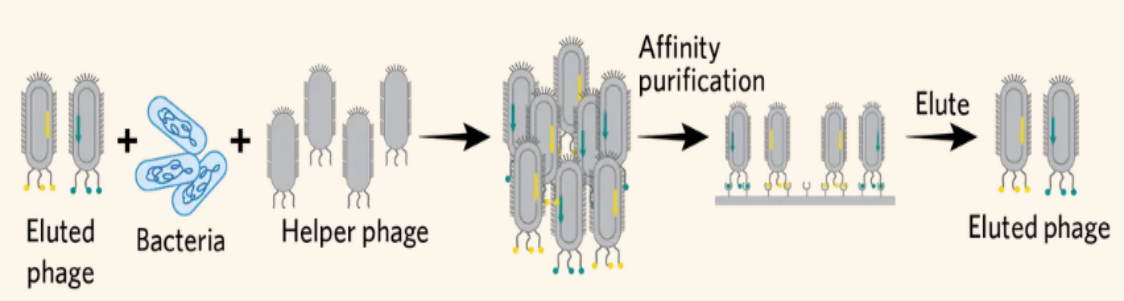

Figure 1.5 Phage amplification Doe (2023)

4. Clone Isolation:

In this final step, bacteria are infected with the selected phages from the amplification step and cultured on plates containing antibiotics. Only bacteria that have been successfully infected by the phages (and thus carry the antibiotic resistance gene from the phage plasmids) will grow on these plates. Researchers then pick antibiotic-resistant colonies and isolate the plasmids from these bacteria. The isolated plasmids, which contain the DNA sequences coding for the proteins of interest, are then sequenced. This sequencing step identifies the specific proteins or peptides that showed affinity for the target. These sequences can then be cloned into protein production vectors for various applications, such as therapeutic development, research into protein interactions, or other biotechnological uses.

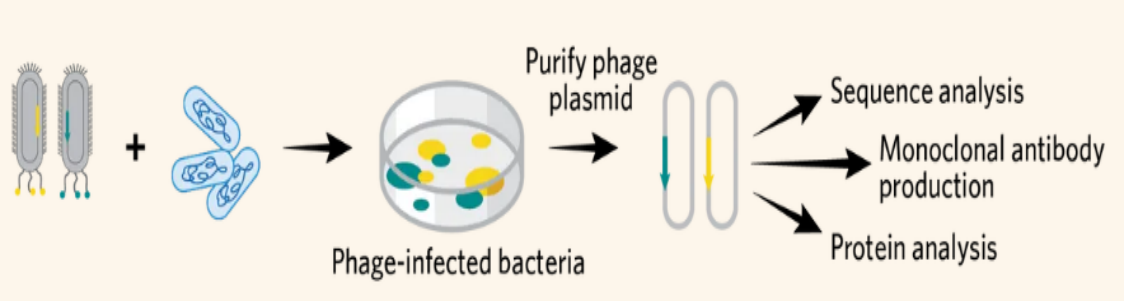

Figure 1.6 Clone isolation Doe (2023)

1.1.2.1 Mimotope Motifs and Finding Mimotope Motifs

The concept of regulatory motifs is central to this work. Motifs, in the context of bio-informatics, are conserved patterns or sequences within biological data, often representing functional elements. A motif, in this research, is represented as a position weight matrix that summarizes the amino acid preferences for each column of the alignment. The quality of a motif is quantified through measures such as information content (IC) or Kullback-

Leibler divergence (KLD). Information content quantifies how much information is contained in a motif, with higher IC values indicating more specificity. Kullback-Leibler divergence measures the dissimilarity between two probability distributions, allowing us to compare the motif's distribution to a background distribution, further assessing its uniqueness and significance. This dissertation introduces a novel method for discovering regulatory motifs, grounded in graph theory, and demonstrates its effectiveness through statistical testing on simulated and real data.

1.1.3 Single Cell Sequencing

Single-cell sequencing has emerged as a pivotal technology in cancer genomics, offering unprecedented insights into the heterogeneity and evolution of tumor cell populations. Unlike traditional bulk sequencing, which provides an average representation of a cell population, single-cell sequencing allows us to investigate the genomic, transcriptomic, and epigenomic profiles of individual cells.

The process involves isolating single cells, extracting and amplifying their genetic material, and subsequently sequencing it. This technology offers insights into cellular heterogeneity, revealing sub-populations, rare cell types, and dynamic changes within tissues, including those relevant to cancer. This work underscores the significance of single-cell recordings and presents a computational framework that harnesses graph theory to infer allele and haplotype-specific copy numbers. The resulting data not only unveils copy number aberrations but also contributes to the construction of phylogeny trees, shedding light on the evolutionary trajectories of cancer mutations.

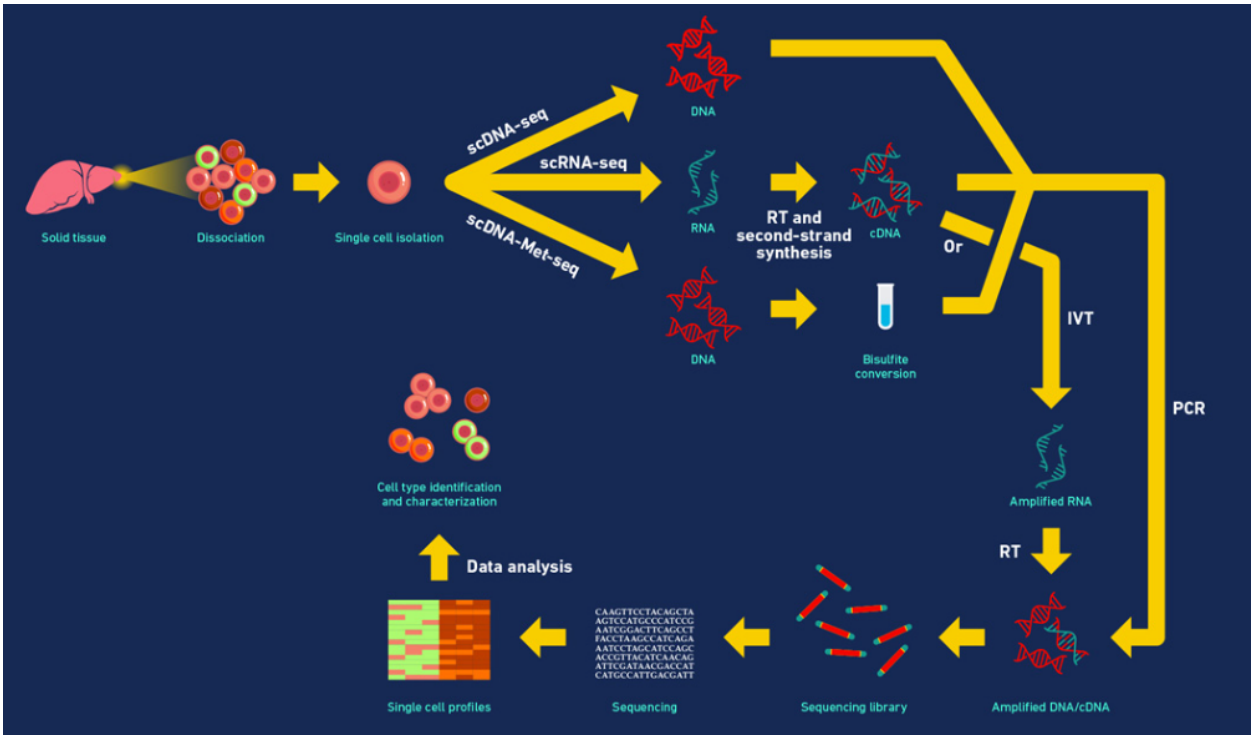

Figure 1.7 Single Cell Sequencing Workflow 10x Genomics (2019)

1.1.3.1 Copy Number Aberration

Copy number aberrations (CNAs) are changes in the number of copies of specific DNA segments. They are crucial in cancer biology because they can disrupt the balance of gene expression, influencing the activity of genes involved in cancer development and progression. CNAs can involve either an increase (amplification) or decrease (deletion) in the number of copies of DNA segments. The consequences of CNAs are significant because they can lead to the overproduction or underproduction of essential genes, contributing to tumor formation and diversity. Understanding the pattern of copy number alterations (CNAs) in cancer genomes is crucial for understanding the molecular processes that drive tumor development.

CNAs play a pivotal role in cancer biology, distinguishing themselves from single nucleotide variants (SNVs) in terms of scale and impact. This dissertation explores the biology of CNAs, elucidating their significance in the context of cancer progression. The integration of read depth ratio and B-allele frequency signals, underpinned by a Bayesian approach, forms the backbone of a novel methodology for CNA detection and characterization.

1.1.3.2 Phylogenetic Tree

Phylogenetic trees are graphical representations of the evolutionary relationships among a set of entities, such as species or, in your case, cancer mutations. These trees are invaluable tools for visualizing the history and relatedness of these entities.

Parsimony in phylogeny tree construction aims to find the simplest tree that explains the observed data. The principle of maximum parsimony aims to reduce the number of evolutionary changes, such as mutations or copy number alterations, necessary to explain the observed data.It provides a framework for reconstructing the evolutionary history of cancer mutations, revealing how these mutations have evolved over time. The concept of phylogeny trees and the principle of phylogeny tree parsimony are integral to this research. By leveraging phylogeny tree construction, this dissertation unveils a novel approach to understanding the evolution of mutations within cancer populations.

1.2 Bio-informatics Challenges

1.2.1 Challenges in NGS and Phage Display

The utilization of NGS in phage display experiments introduces specific bio-informatics challenges, which are addressed in this section. These challenges include the complexity of data analysis and the need for motif discovery within vast sequence data sets.

1.2.1.1 Data Analysis Complexity

NGS generates extensive data sets of peptide sequences from phage display experiments. Analyzing this data can be computationally intensive and requires sophisticated algorithms to extract meaningful insights.

1.2.1.2 Motif Discovery

Identifying binding motifs within the sequences is a critical aspect of phage display analysis. Existing tools face challenges such as scalability for large data sets, the requirement for prior knowledge about motif numbers, and issues with accuracy.

1.2.2 Challenges in Single Cell Sequencing

Single cell sequencing, while powerful, introduces its own set of bio-informatics challenges. These challenges include data sparsity, low coverage signal per cell, and the need to phase allele and haplotype-specific copy numbers.

1.3 Problem Formulations

In the first problem formulation, we aim to identify regulatory motifs from NGS data. We start with 7-mer peptides as input and create a graph where nodes represent 4-mers of the peptides. Edges are established based on a certain threshold of support from 7-mer peptides. By identifying all two-hop paths in this graph and aligning supporting peptides, we derive position probability matrices, uncovering critical motifs within the data.

In the second problem formulation, we tackle the challenge of inferring allele and haplotype-specific copy numbers from single-cell sequencing data sets. Our approach involves constructing a multi-partite graph, with nodes representing possible copy numbers in each genomic region (bins of size $5k$). Edge weights and node weights are determined based on read depth ratios and B-allele frequency signals. By utilizing Dijkstra's algorithm to find the shortest path from source to sink, we characterize allele and haplotype-specific copy numbers, enabling the construction of phylogenetic trees to elucidate cancer mutation evolution.

1.4 Contribution

The core findings of this dissertation are summarized in several significant contributions.:

- A novel method for the identification of regulatory motifs in mimotope profiles.
- Rigorous statistical analysis to assess the accuracy and recall of the motif discovery method on both simulated and real data.
- Creation of simulated data libraries using position probability matrices to support

research in cancer genomics.

- Development of a groundbreaking approach for characterizing allele and haplotype-specific copy number aberrations.
- Pioneering the calculation of B-allele frequency signals in single-cell data sequencing.
- Generation of simulated data for the inference of migration histories in metastatic cancers.

1.5 Road Map

The remainder of this dissertation is organized as follows. Chapter 2 embarks on the identification of regulatory binding sites through phage display library and NGS sequencing, offering a comprehensive review of the existing literature, the introduction of our novel method, and a thorough examination of results. Chapter 3 explores the realm of single-cell barcoding technologies and presents a method for inferring allele- and haplotype-specific copy numbers. This chapter also delves into the reconstruction of tumor evolution. The subsequent chapters provide further depth, analysis, and conclusions regarding the cutting-edge research within the realm of cancer genomics.

With these components in place, this dissertation aims to contribute significantly to the field of bioinformatics and cancer genomics, offering novel insights and methodologies that may pave the way for breakthroughs in cancer research.

1.6 Publications and Presentations

- Presented my research at the ISBRA 2023 conference in October 2023 in Poland.
- Published a paper titled "Graph-based motif discovery in mimotope profiles of serum antibody repertoire" in the proceedings of the International Symposium on Bioinformatics Research and Applications (ISBRA).
- Upcoming presentation at the ICCABS conference in December 2023 in Oklahoma,

CHAPTER 2

GRAPH THEORY INSIGHTS: UNCOVERING REGULATORY MOTIFS and EPITOPES via MAXIMUM CLIQUE ANALYSIS

Antibodies play a pivotal role in the human immune system's response to antigens, making them invaluable biomarkers for the detection of diseases such as cancer or viral infections [190]. The molecular interactions between an antigen's epitope and antibodies are often mediated by short linear motifs within the amino acid sequence of the antigen itself [39, 91]. Detecting and characterizing these interactions is essential for diagnostic and therapeutic purposes.

Experimental methods have been developed to identify peptides recognized by antibodies present in human serum. One such method involves the use of random peptide phage display libraries in combination with Next-Generation Sequencing (NGS) [15, 60]. This approach generates vast data sets consisting of hundreds of thousands of peptide sequences. The central computational challenge lies in identifying the true binding motifs corresponding to disease-related epitopes within these data sets.

In this chapter, we explore the computational aspects of discovering binding motifs and epitopes, addressing the major challenges associated with existing methods. These challenges include scalability issues, the requirement of prior knowledge about the number of motifs, and issues related to accuracy [60, 91].

Our approach is based on theory of intersection graphs We propose a novel approach to tackle the binding motif discovery problem efficiently and accurately.

2.1 Problem formulation

Consider the set $\mathcal{P}$ of peptides of length L. For an integer k, a $k-$part of a peptide $p \in \mathcal{P}$ corresponding to a subset $J = \{j_1, ..., j_k\} \subseteq [L]$ is a pair (s, r), where s is sub-string of p formed by amino acids at positions J, and r is the vector of numbers of intermediate positions between consecutive elements of J. For example, for $p = KKEGLHD$, $k = 4$ and $J = \{2,4,6,7\}$, the corresponding $k-$part is $(KGHD, (1, 1, 0))$. The value of k is user-defined parameter. Since the peptides in real data are short($L \leq 10$), it is possible to run the proposed algorithm with different values of k, with $k = 4$ or 5 usually being sufficient. We choose value of $L = 7$ and value of $k = 4$ as the condition of the experiments of serum samples allow us for such an assumption. In addition, in some similar researches like in [Ionov & Rogovskyy (2020)], the same assumptions are employed.

Instead of Analyzing the set $\mathcal{P}$, we propose to consider the set $\mathcal{H}_k$ of all unique $k-$parts of all peptides (together with their counts). Given the set $\mathcal{H}_k$, we construct an intersection graph $\mathcal{G} = \mathcal{G}(\mathcal{H}_k)$ with the vertex set $\mathcal{H}_k$, and two $k-$parts being adjacent, if they agree in all but two amino acids and all intervals (see Fig. 1). for each $k-$part v, we assign the set of $(k-1)-$ parts $H(v)$ obtained by removing each of its amino acids (Fig 1.b). Then the two $k-$parts u and v are adjacent in $\mathcal{G}$, if they share a $(k-1)$ part, i.e. $H(u) \cap H(v) \neq \emptyset$. For example $k-$parts $v1 =$(KGHD, $(1, 1, 0)$) and $v2 =$ (KMHD, $(2, 0, 0)$) share the $(k-1)-$part KHD , $(3, 0)$) and thus they are adjacent in $\mathcal{G}$.

Cliques of $\mathcal{G}$ represent sets of $k-$parts, which mutually agree with each other, and the true motifs are represented by cliques. While in general generation of all candidate maximal

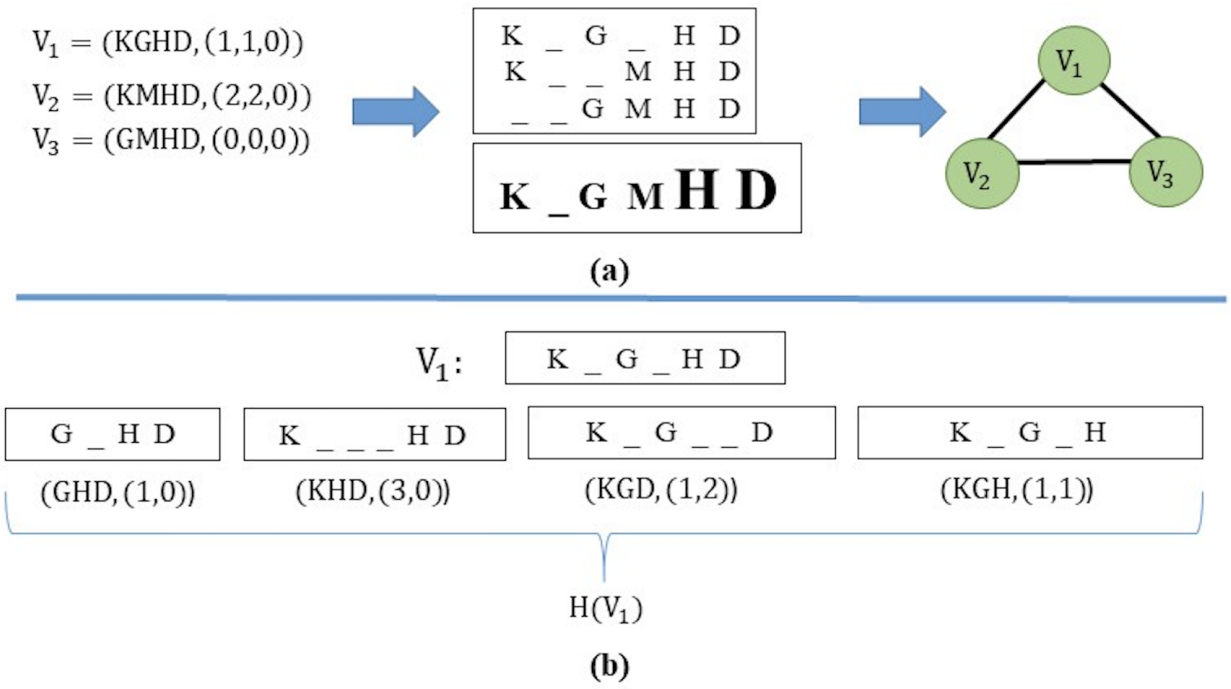

Figure 2.1 (a) Construction of the intersection graph for 3 k-parts v_1, v_2 and v_3. In the middle, an alignment of k-parts and the motif corresponding to that parts is shown (sizes of symbols in the motif represent their frequencies). (b) $(k-1)$-parts $H(v_1)$ induced by the k-part v_1

cliques requires exponential time, in our case we can efficiently solve this problem using combinatorial properties of the graph $\mathcal{G}$. By Construction, $\mathcal{G}$ is an intersection graph of a linear k-uniform hypergraph(i.e. a hypergraph with all hyperedges of size k and every pair of hyperedges having at most one common element). We can prove that such graphs have the following properties:

For a fixed k the graph $\mathcal{G}$ contains a polynomial number of maximal cliques.

Any two cliques with at least $k^2 - k + 2$ vertices have at most 1 common vertex.

Let C be a clique with at least $k^2 - k + 1$ vertices. Then C is a subset of a single maximal clique.

#Implanted Motifs	Sample Size	#Extracted Cliques	Largest Clique	Sensitivity	Specificity	Clique Count
1	1000	25654	2600	%100	%100	1
2	1000	14037	1300	%100	%100	2
3	1000	8968	871	%100	%100	3
4	1000	8464	650	%100	%100	4
5	1000	8247	520	%100	%100	5
6	1000	8505	442	%100	%100	6
7	1000	8498	377	%100	%100	7

Table 2.1 Results of graph based motif extraction on simulated data.

Theorem 1 implies that all maximal cliques of $\mathcal{G}$ could be efficiently generated. We used the modified Bron-Kerbosch algorithmBron & Kerbosch (1973); Tomita et al. (2006) which terminates each recursion, once a clique C of size $k^2 - k + 1$ is generated (in our case $k = 4$), whereupon C is expanded to a maximal clique. Using theorem 1, the expansion could be done in time linear by the number of edges having at least one end in C. Once the candidate cliques are generated, the corresponding motifs are assembled to obtain the set of candidate motifs $\mathcal{M}$. Next we select reliable motifs from the set $\mathcal{M}$. The simplest criteria is to select motifs corresponding to largest cliques. In order to eliminate sequencing chimeras, it is also important to ensure that k -sets in the clique have high counts. Thus we propose to select motifs M with the highest values $d(M) = n(M)c(M)$, where $n(M)$ be the size of corresponding clique and $c(M)$ be the total count of k-parts forming that clique. Finally, motifs based on their information content will be selected.

2.2 Results

Table 2.1 represents the performance of graph based algorithm on the simulated data. Sensitivity and specificity of our method is shown in this table.

We also compared running time of our modified Bron-Kerbosch algorithm with original

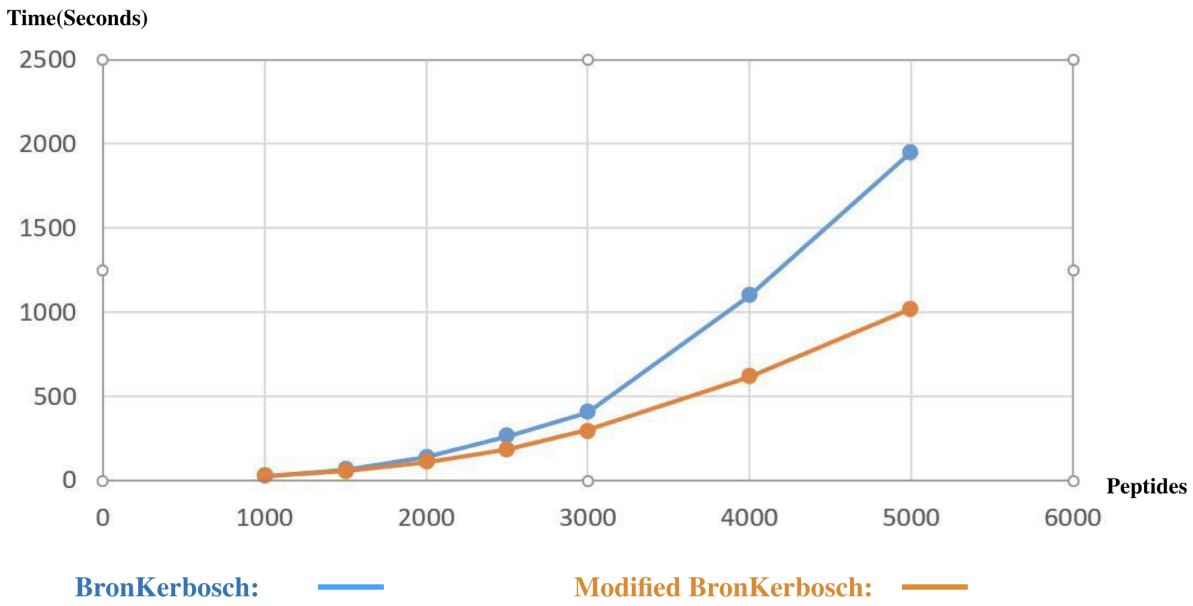

Figure 2.2 Running time of Modified Vs original Bron-Kerbosch algorithm

one. In Figure 3 we illustrate running time of both algorithms. As the number of peptides increases, our method runs outstandingly faster on the original Bron-Kerbosch algorithm.

CHAPTER 3

GRAPH-BASED MOTIF DISCOVERY IN MIMOTOPE PROFILES OF SERUM ANTIBODY REPERTOIRE

3.1 Introduction

Phage display (or biopanning) is a technique for studying different protein interactions including protein-DNA, protein-peptide and protein-protein interactions using bacteriophage (a type of virus that only infects bacteria)(Smith 1985). In this technique, antibody genes are combined on a strand of DNA. The DNA is then packaged in a protein coat made from bacteriophage. The antibody genes make the antibody hat (receptor), which is attached to the top surface of the virus coat. The virus is called phage and the combination is called phage antibody. Each phage antibody hat is unique and binds to a specific target molecule (for example an antigen, an epitope, or a peptide). Target refers to the substance that is used to scan phage library and template is considered its natural partner. Only the antibody phage hat that fits the shape of a disease target will bind to the target molecule. Changing the antibody genes will change the type of antibody hat and what it can bind to (antigen, or epitope). Many of these antibody phages have been made and the pool contains billions of unique antibody phages. Together all these antibody phages are called phage display library. This pool of antibody phages contains unique receptors for specific target binding and thus, can be screened to reveal specific disease targets. For example, cancer patients' serum can be incubated with phage library to reveal cancer specific epitopes. Once antibody phages bind with specific targets, they can be pulled out and further replicated using a host

bacteria(Murphy & Weaver 2016).

Geysen et. al. (Geysen et al. 1986) first identified peptide binding to target and mimicking the binding site on the template, which was called mimotope. Mimotope is useful in many applications such as epitope mapping(Smith & Petrenko 1997), vaccines(Knittelfelder et al. 2009), therapeutics(Macdougall et al. 2009), defining drug target(Rodi et al. 1999), protein network detection(Tong et al. 2002)(Thom et al. 2006). As a result of exposure to antigenic proteins, patient's immune system produces antibodies, which could be used as biomarkers for cancer or viral infection detection(Zhong et al. 2006). Molecular interactions between antigen's epitope and antibody are often mediated by short linear motifs of the amino acid sequence of the antigen(Dinkel et al. 2014; Krejci et al. 2016). Such interactions could be experimentally detected using random peptide phage display libraries.

Traditional phage display is laborious and prone to finding false positive hits. In recent years, many studies have been devoted to taking advantage of Next Generation Sequencing (NGS) technique in the analysis of phage display screens (Rentero Rebollo et al. 2014; Christiansen et al. 2015). NGS enables phage display screening to produce huge number of outputs (short peptides). Another contribution of NGS to the analysis of phage display screening is that it accelerates and improves selection process and therefore, avoids repetitive selections and restricts the number of false positive hits, in contrast with traditional phage display (Wang & Yu 2004). In effect, a library of all possible peptides of fixed length is generated, and peptides recognized by antibodies contained in the human serum are selected, amplified in bacteria and sequenced using NGS (Bratkovič 2010; Gerasimov et al. 2017). Such

methods produce data sets consisting of hundreds of thousands of peptide sequences. The computational problem consists in discovery of true binding motifs corresponding to epitopes related to the diagnosed disease.

The problem of detection of epitope-specific binding motifs from NGS data is computationally challenging for several reasons. The generated data is large and usually noisy, as a result of biopanning; mimotopes which are considered as desired signals are mixed with target unrelated peptides (TUPs) that are undesired signals. Thus, a significant portion of sequenced peptides is not related to the repertoires of antibody specificities, but produced by nonspecific binding and preferential amplification in bacteria(Gerasimov et al. 2017). High heterogeneity of antigen and antibody populations, as well as antibody-antigen recognition poly-specificity manifests itself in presence of multiple binding motifs of various lengths within the same data set(Krejci et al. 2016; Van Regenmortel 2019).

The development of regulatory motif finding algorithms began in the late 1980s and early 1990s, when researchers began using computational methods to identify and analyze patterns in DNA sequences. Early motif finding algorithms used sequence patterns to identify conserved motifs. One of the first algorithms was Gibbs Sampling, developed by Lawrence et al. in 1993 (Lawrence et al. 1993). This algorithm used a probabilistic approach to identify common patterns in a set of DNA sequences. In the mid-1990s, researchers developed a new way to represent motifs called position weight matrices (PWMs). PWMs show the probability of each nucleotide at each position in a motif. This information can be used to develop algorithms like MEME (Bailey et al. 1994) and AlignACE (Roth et al. 1998), which

can find motifs in DNA sequences. Phylogenetic footprinting is another way to find regulatory motifs in DNA sequences by comparing them to sequences from other species. This method looks for regions that are conserved across species, because these regions are more likely to be functional. Notable examples of phylogenetic footprinting algorithms include FootPrinter (Blanchette & Tompa 2003) and PhyloGibbs. (Siddharthan et al. 2005). In addition, machine learning techniques have been used to develop new motif finding algorithms that are based on classification and regression models. These models, such as support vector machines (SVMs), hidden Markov models (HMMs), and neural networks, have been used to discover motifs in DNA sequences. Some examples of machine learning-based algorithms include SVMotif (Kon et al. 2007), MDscan-Motif (Liu et al. 2002), and DeepBind (Alipanahi et al. 2015).

The tools that have been developed to address the limitations of existing motif finding methods use a variety of algorithmic techniques, such as clustering (Gerasimov et al. 2017; Krejci et al. 2016), Gibbs sampling (Andreatta et al. 2013; Nielsen et al. 2004), artificial neural networks (Nielsen & Lund 2009), and mixture model optimization (Kim et al. 2012). However, these methods face serious challenges, when dealing with NGS data: many of them are not scalable for large data sets, require prior knowledge about the number of motifs to identify, and have low accuracy(Gerasimov et al. 2017; Krejci et al. 2016).

We introduce a method for finding regulatory motifs, which relies on results from graph theory. In order to evaluate our method, we generated samples of sequences with different lengths. We planted predefined motifs into the samples and applied our algorithm to simu-

lated data. We also applied our method to mouse microbiome data including two groups of 5 mice. Results indicate that the graph-theoretic approach successfully identifies motifs in replicates samples.

In the continue, we introduce our method and then we validate our approach on both simulated and real data and analyze the results.

3.2 Method

Our graph-based method uses the concept of graphs in mathematics to formulate the problem of motif discovery. In this method, we construct a directed graph in which nodes are $k-$mers. We assume that $k = 4$ and thus, we deal with $4-$mers in this research. The reason is that all peptides are of length 7 and according to Ionov & Rogovskyy (2020) choosing $k = 4$ results in more statistically significant motifs.

3.2.1 Problem Formulation

For given set of peptides, we find all $4-$mers of that set. By $4-$mer we mean a string of amino acids of length 4 in which the first and the last positions occupied by amino acid and the second and the third positions either amino acid or gap (dash). In other words, at most, one deletion is allowed in the 2nd or 3rd positions. We build a graph in which vertices are $4-$mers. Two vertices are connected by an edge, if there are enough M number of peptides supporting simultaneously both $4-$mers.

Let u, x be two $4-$mers belonging to the same peptide P, and the first position of u strictly precedes the first position of x in P. Then there is a directed edge $u \to x$. Together

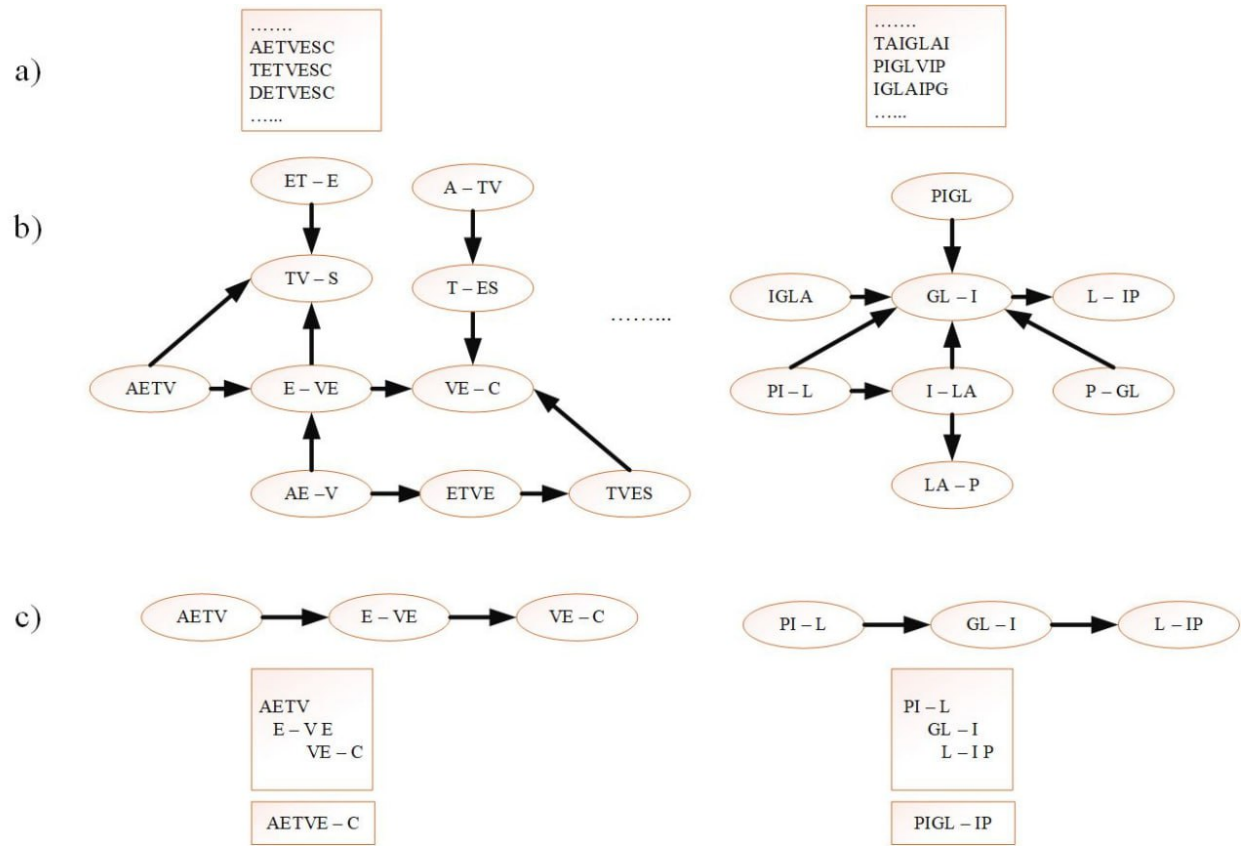

Figure 3.1 Schematic of the Graph-based method. a) Peptides that contribute to particular part of the graph. b) A sample directed graph made from 4−mers as vertices. Two vertices are connected if they both belong to a peptide. Direction of the edge is from the 4−mer that fills lower indexes to the 4−mer that occupies higher indexes when align to the peptide. c) Examples of paths of length 2 (2− hops) in the graph. 4−mers are aligned and as a result, $k-$ subsets corresponding to paths are created.

u and x can make a 5−, 6− or 7−subset $k-ux$ inside P. This $k-$ subset forms an edge $(k-ux)$, if the first amino acid belongs to u and the last belongs to x. The support of an edge $(k-ux)$ is equal to the number of peptides containing $k-$ subset $k-ux$. We consider only edges with the support greater than M.

After building the graph, we extract all paths of the graph with length equal to 2, i.e. we find all 2-hop paths of the graph (see Fig. 3.1−c). Each path uxv contains two directed

edges with $(k1 - ux)$ and $(k2 - xv)$ subsets with enough number of M supports. Together $(k1-ux)$ and $(k2-xv)$ make a $(k-uxv)$ subset of the whole path. We assume that support of a path uxv is equal to minimum of {support (ux), support (xv)}.In the next step, we align all peptides that support the path and align them to the achieved $k-$subset $(k-uxv)$. As a result, for each path of length 2, we obtain a set of aligned peptides which accounts for a motif and we can represent this motif with a position probability matrix. The following formula is used to calculate probability $P(Aa)$ of each amino acid at an individual position of the motif

$$P(Aa) = \frac{N_A a + \frac{p}{n}}{\sum N + p} \tag{3.1}$$

where $N_A a$ is the number of amino acid at each position, N is the total number of peptides contributing in motif, p is the pseudo count which is added to the nominator and denominator of the fraction, and n is the total number of amino acids (n=20). The reason for applying pseudo counts to the formula above is that in some positions, counts of one particular amino acid would be zero (in contrast with other positions), which results in probability of some significant motifs to be zero.

In studying motifs, some positions might be more important and subsequently, contain more information. To explore this, information content matrix is calculated based on Shannon entropy equation (Ash 2012). Alternatively, in some researches, information content is represented as relative entropy, Kullback-Leibler divergence (KL divergence). In contrast with Shannon entropy, KL divergence takes into account the non-uniform background frequencies. The relative entropy is calculated using the following equation:

$$IC(Aa) = P(Aa) \times \log_2 \frac{P(Aa)}{B_A a} \tag{3.2}$$

in which $B_A a$ is the background frequency of amino acid a. In this study, we accounted for the non-uniform background frequencies of amino acids (i.e., the probability of each amino acid is not equal). To do this, we calculated the frequency of each amino acid at each position of $7-$mers in each sample. In equation 2, IC can take negative values. In order to avoid negative values, we simply replace them with zero.

3.2.2 Motif Validation

Methods of quantifying similarity between motifs include (but not limited to) Pearson Correlation Coefficient (PCC) Pietrokovski (1996), Average Log-Likelihood Ration (ALLR) Wang & Stormo (2003), Fisher-Irwin exact test (FIET) Schones et al. (2005), Kullback-Leibler divergence (KLD) Roepcke et al. (2005), Euclidean distance Choi et al. (2004) and Tomtom (E value) Gupta et al. (2007). We used Pearson Correlation Coefficient to measure if two motifs are identical.The Pearson correlation coefficient (PCC) is a measure of the linear relationship between two motifs. It is calculated by comparing the occurrence profiles of the two motifs across a set of sequences. A high positive PCC indicates that the two motifs are similar, while a low or negative PCC indicates that they are dissimilar. To this end, we selected a threshold correlation number $correlation \geq 0.75$. Any two motifs with PCC greater than 0.75 are assumed identical motifs. Accordingly, we calculated the number of retrieved motifs. The reason we used the Pearson correlation coefficient over other alternative approaches

		Actual	
		Positive	Negative
Predicted	Positive	A, B, E	C, D
	Negative	3	$NotApplicable$

Table 3.1 Confusion matrix built after matching related to Fig. 3.2

in this study is that we represent motifs as position probability matrices. Employing the Pearson correlation coefficient for comparing two matrices aligns most effectively with our graph-based methodology.

In order to verify the discovered motifs obtained through our graph-based method, we conducted tests using simulated data. The simulated data was generated by creating several position probability matrices, which were used to generate sets of 7-mers. These 7-mers served as the intentionally placed motifs within our simulations. If any gaps were present in the motifs, we filled them with random amino acids. Subsequently, we applied a graph-based approach to analyze the simulated set of peptides. We utilized simulated data to evaluate the performance of our graph-based method in detecting the target motifs. By creating position probability matrices and generating corresponding peptides, we introduced noise and assessed the algorithm's ability to identify the desired motifs in the presence of such noise. In Fig. 3.2, the U set represents the planted motifs, while the V set represents the collection of all retrieved motifs. To determine the correlation between the members of set U and set V, we employed the Pearson correlation coefficient. The matching between the two sets was accomplished using the Gale-Shapley algorithm.

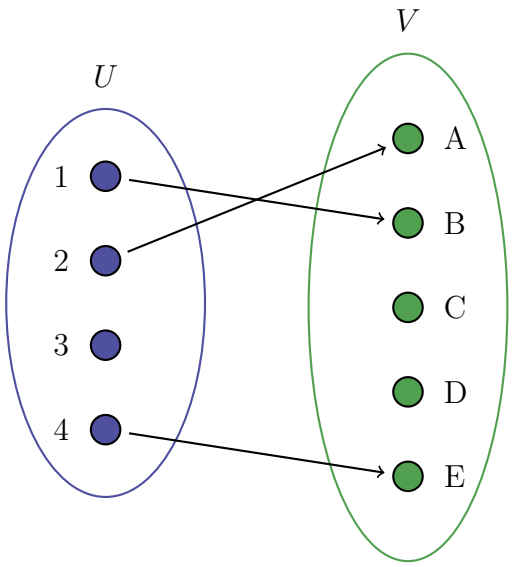

Figure 3.2 set U represents set of true planted motifs, set V represents set of retrieved motifs. When motifs are retrieved (in set V), Gale-Shapley algorithm is used to do matching between set U and set V. Each motif in set U is matched with only one motif in set V with whom it has highest Pearson Correlation Coefficient

3.3 Results and Discussion

3.3.1 Data set

Our mouse microbiome data is the mixture of the 3 libraries: $M2$, $L2$ and $M1$. M means that the IgM antibodies were used for analysis and the L means that antibody were isolated from the same serum samples by using protein L, which do not discriminate between classes of antibodies. With IgM antibodies the experiment was repeated with the same serum samples two times , this is why there are the $M2$ and the $M1$ libraries. L library include all antibodies (IgG, IgA, IgE) except the IgM. We expect that IgM repertoire is the most sensitive to the environment changes and we should find real differences in the IgM repertoire that is in the $M2$ and the $M1$ samples. So, totally we have 3 libraries $M2$, $L2$ and $M1$.

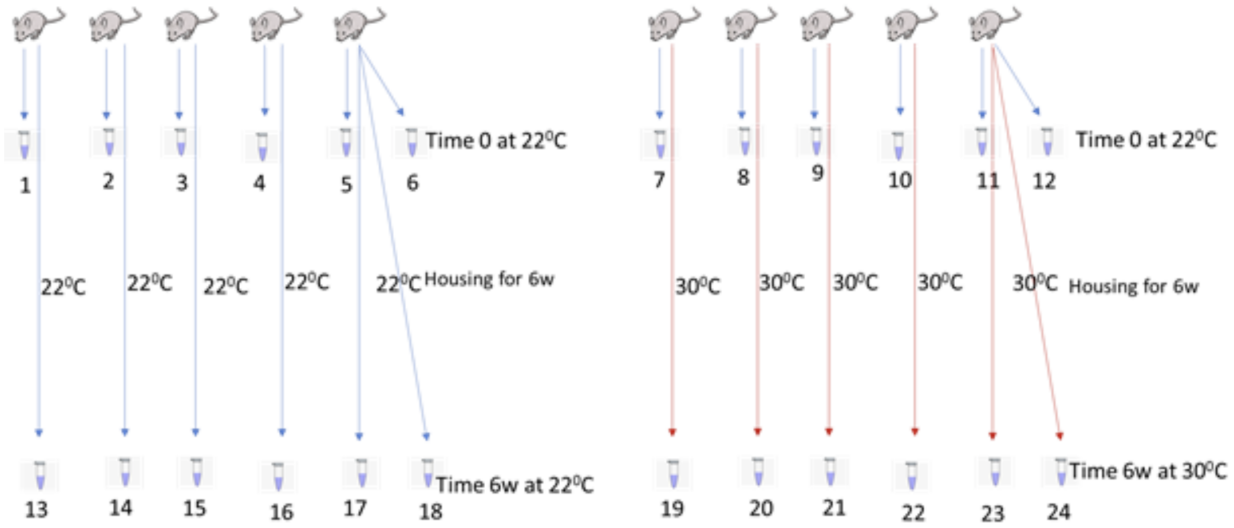

Figure 3.3 Mouse microbiome data

Each library consist of 24 serum samples obtained from two groups of mice. Prior to time $t = 0$, all mice from both groups were maintained at 22 °C. At t = 0, serum samples were collected from mice in both groups, resulting in samples $S1 - S6$ and $S7 - S12$. Each group initially contained 5 mice, but the serum samples from mouse 5 in the first group and mouse 10 in the second group were duplicated as technical replicas. Thus, there were 6 profiles in each group.

After the initial bleeding, the first group of mice was kept at 22 °C, and the second bleeding occurred after 6 weeks ($t = 6w$) at the same temperature, producing samples $S13 - S18$. Again, serum samples from mouse 5 were duplicated. Meanwhile, the second group of mice was kept at 30 °C, and after 6 weeks at $t = 6w$, samples $S19 - S24$ were collected. Serum from mouse 10 was duplicated in this case. The phage DNA used for insert sequencing was derived from antibody-bound phages immediately after the initial incubation

of the phage library with serum antibodies. This process was carried out without amplifying the isolated antibody-binding phages in bacteria. Consequently, the number of sequencing reads cannot serve as a quantitative measure of the antibody titer. This limitation arises because, in many cases, the number of corresponding antibodies far exceeds the number of their specific targets.

To create quantitative profiles, we propose quantifying the number of distinct peptide sequence variants or determining the size of the peptide family related to each motif. This approach allows us to calculate how frequently each motif appears in each profile, enabling the generation of a motif signature that can differentiate between housing at 22°C and 30°C. In summary, to discern the impact of temperature conditions, we focus on measuring motif occurrences and constructing signature profiles rather than relying on sequencing reads or antibody titers.

3.3.2 Results

3.3.2.1 Simulated Data:

In order to confirm the motifs identified through our graph-based method, we conducted tests using simulated data. The simulated data was generated by using multiple position probability matrices, which were then utilized to generate sets of 7-mers. Essentially, we created a pool of diverse position probability matrices, each representing an identical motif with varying levels of information content associated with it. We did this by collecting all of the position probability matrices that were generated when we applied our method to

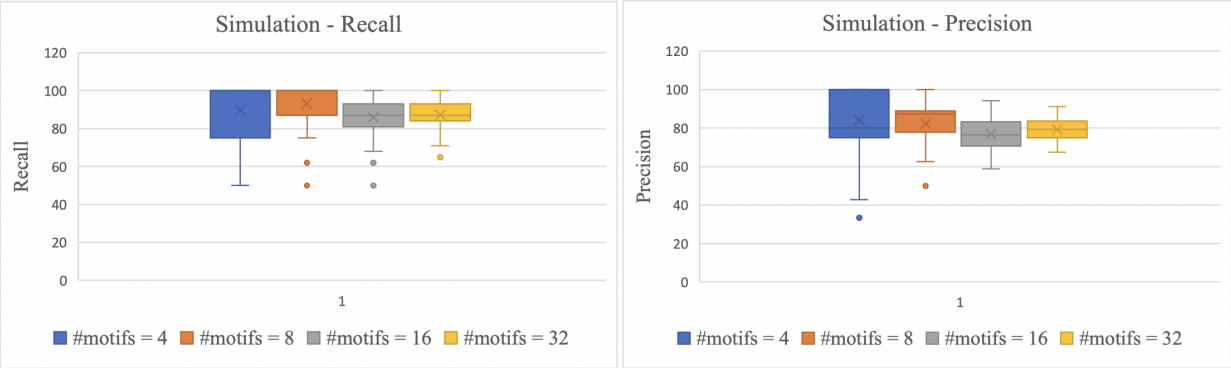

Figure 3.4 Whisker bar plot of the simulated data. Number of planted motifs is on the x axis while the y axis shows recall and precision for each group.

real data. Subsequently, we randomly selected a specific number of matrices from this pool and generated a corresponding number of 7-mer peptides based on each selected matrix. To address any gaps within the motif, we inserted random amino acids, thereby introducing noise to the simulated data. Next, we applied a graph-based approach to analyze the set of generated peptides. The objective was to determine whether the graph-based algorithm was able to detect the intended target motifs or not. We conducted a series of experiments involving the insertion of sets of 4, 8, 16, and 32 motifs. Each set was tested 100 times, and the recall (sensitivity) and precision were calculated for each experiment. Figure 3.4 visualizes the results of all simulations through two whisker bar plots.

3.3.2.2 Real Data:

The primary objective of our research was to investigate any qualitative or quantitative differences in the profiles associated with the second bleeding of both the first and second groups of mice, which were respectively kept at 22°C and 30°C. The focus was to determine

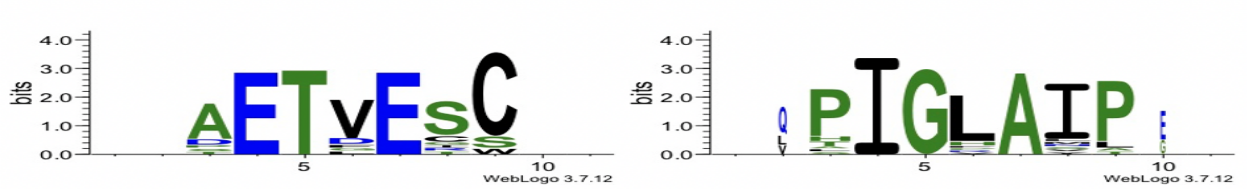

Figure 3.5 Two technologically inserted motifs were successfully identified in all 24 samples with graph-based method

if variations in both time and temperature would lead to significant differences or patterns indicating an increase or decrease in motifs within the samples. However, our analysis did not reveal any significant differences or discernible patterns that would suggest an increase or decrease in motifs across the samples. Despite the variations in time and temperature, the motifs in the samples did not exhibit notable changes in their quantity or quality. To further explore the data, we employed 4−mer sequences instead of motifs, as most motifs comprised consensus sequences of 4 amino acids. There were a total of $160,000$ distinct 4-mers, and for each profile, we calculated the number of occurrences of each tetra peptide. This involved determining how many different 7−mer sequences contained a particular tetra peptide in each profile, generating a signature of tetra peptides related to housing at 30 °C. The tetra peptides could exhibit either increased or decreased numbers with the temperature shift. However, we did not find a meaningful relationship between the samples in terms of increasing or decreasing tetra peptides. It's important to note that although no significant differences were observed in this study, negative results are valuable as they contribute to the understanding of the regulatory motifs' behavior under specific conditions. These findings indicate that the specific motifs we investigated were not significantly affected by

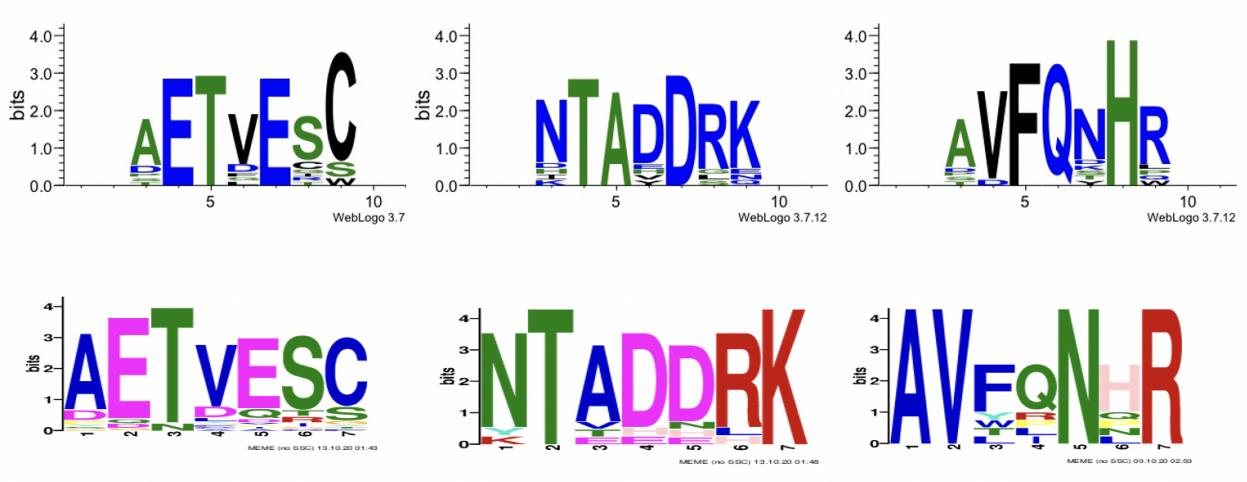

Figure 3.6 The first row shows some sample motifs discovered by our graph-based method. The second row is assigned to motifs discovered by MEME. The logos of the first row was created using a version of WebLogo Crooks et al. (2004) modified to display aligned pairs of Logos

the time and temperature variations in the experiment. Additionally, it is worth highlighting that our method successfully identified two technologically inserted motifs ("IGLAEIP" and "AETVESC") in all of the samples (see Fig. 3.5). This discovery suggests that our method is valid and capable of detecting important motifs. Comparatively, we also applied a well-known motif discovery method called MEME Bailey et al. (1994) to analyze all 24 samples. However, MEME failed to identify these two motifs in all of the 24 samples. In Fig. 3.6, we have summarized a number of discovered motif logos using our graph-based and MEME.

3.3.3 Discussion

To evaluate the results, we obtained a set of results for different values of three critical parameters: the number of peptides that support a two-hop path in the graph (which we call the "support path"), the information content of the planted motifs, and the correlation

number. These parameters are all assumed to have a significant impact on the measurements.

We first hypothesized that increasing the path support would result in stronger motifs that were easier to retrieve. We started with a path support of 4 and increased it from there. We saw that the number of extracted motifs increased until we reached a certain point. However, the results remained unchanged when we chose higher values for the path support. This is because target motifs are mainly clustered in groups of a certain number of peptides. For example, if we start with a path support of 4, meaning that there are at least 4 peptides supporting the path, as we increase the path support, more target motifs will be detected by the graph-based method. However, from a certain number of path support onward, we get plenty of non-targeted motifs and the number of successfully detected target motifs will not increase.

We planted a variety of motifs with different information contents in our simulated data. As we expected, motifs with higher information contents were more likely to be retrieved than those with lower information contents.

Our graph-based motif discovery model has several advantages over existing methods. First, it is scalable to large data sets, making it possible to find motifs in hundreds of thousands of peptides in a reasonable amount of time. Second, it does not require prior knowledge of the number of motifs, unlike the MEME method. Third, it has been shown to be more accurate on simulated data than the MEME method.

In summary, our research did not yield significant differences or patterns in the quantity or quality of motifs or tetra peptides when comparing the serum samples collected after

6 weeks and between samples kept at 22 °C and 30 °C. These results suggest that the temperature shift did not have a noticeable impact on the identified motifs or tetra peptides in the samples.

CHAPTER 4

IDENTIFYING COPY NUMBER ABERRATIONS IN SINGLE CELL SEQUENCES

4.1 Introduction

Single-cell analysis has emerged as a powerful tool in genomics research, allowing for the characterization of genomic heterogeneity at the resolution of individual cells. Traditional bulk sequencing methods provide an average view of the genome, masking important events such as allele-specific copy number aberrations (CNAs) and copy-neutral loss-of-heterozygosity (LOH) (Zaccaria & Raphael 2021). However, recent advancements in single-cell barcoding technologies have enabled whole-genome sequencing of thousands of individual cells in parallel, providing an opportunity to study genomic heterogeneity at unprecedented resolution (Zaccaria & Raphael 2021).

The 10x Genomics Chromium platform (see Fig.2.1) has made significant contributions to single-cell genomics by providing a scalable and cost-effective way to analyze gene expression patterns at the single-cell level. It has been used for a wide range of applications, such as understanding cellular heterogeneity, characterizing rare cell populations, and studying developmental processes and disease mechanisms at the individual cell level.

The single-cell sequencing methodology introduced by 10x Genomics uses the Chromium platform to perform high-throughput single-cell analysis. The first step is to obtain a biological sample containing the cells of interest. The sample is then dissociated into single cells.

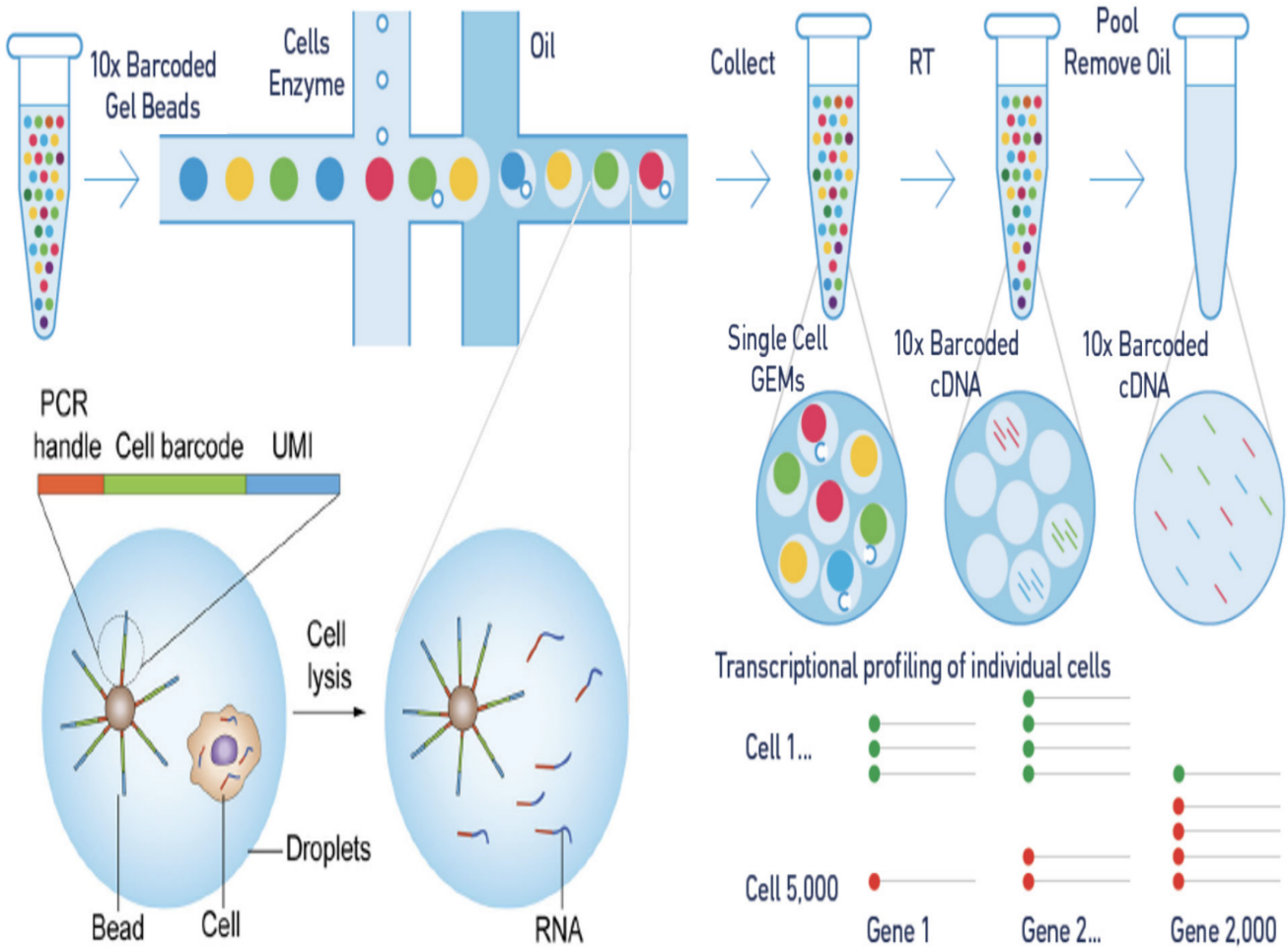

Figure 4.1 The Chromium Single Cell CNV Solution workflow.10x Genomics (2019)

The next step is to encapsulate each single cell in a droplet along with a uniquely barcoded gel bead. This is done using a microfluidic device called the Chromium controller. Within each droplet, the captured cell undergoes lysis, releasing its RNA content. The released RNA molecules are then captured by the barcoded gel bead, which preserves the cellular identity associated with each RNA molecule. The captured RNA is then reverse transcribed into complementary DNA (cDNA). This cDNA is then amplified and used to create a library of DNA sequences that represent the RNA transcripts from the original single cell.

The cDNA libraries are then sequenced using high-throughput next-generation sequencing (NGS) platforms. The sequencing reads obtained from each cDNA molecule contain information about the expressed genes in the original single cell.

The raw sequencing data is then processed and analyzed using specialized bioinformatics tools. The barcodes associated with each cDNA molecule enable the identification of the cell of origin for each transcript, allowing for downstream analysis at the single-cell level. The data analysis includes steps such as read alignment, transcript quantification, identification of differentially expressed genes, clustering of cells into subpopulations, and visualization of the results.

Studying genetic changes in single cell sequencing can provide valuable insights into the molecular basis of diseases and can help us to develop new treatments. These changes include Copy Number Aberrations (CNAs), Loss of Heterozygosity (LoH), and Whole Genome Duplications (WGDs).

Copy number aberrations are changes in the number of copies of a particular gene or genomic region. They can be caused by a variety of factors, including DNA damage, chromosomal rearrangements, and epigenetic changes. Single cell sequencing allows the detection and characterization of CNAs at the cellular level, enabling the identification of genetic variations associated with diseases and understanding their clonal evolution within heterogeneous cell populations. CNAs can have a significant impact on gene expression and can contribute to the development of diseases such as cancer.

Loss of heterozygosity is the loss of one of the two copies of a gene. This can occur

when a cell inherits a mutation in one copy of a gene, or when a gene is deleted. LOH can also be caused by chromosomal rearrangements or somatic mutations. LOH can affect gene expression and can increase the risk of developing cancer. Single cell sequencing enables the detection of LOH events in individual cells, shedding light on clonal expansion, tumor progression, and identifying candidate genes implicated in disease development.

Whole genome duplication refers to the complete replication of the entire genome within an organism. It occurs when the entire set of chromosomes is duplicated, resulting in multiple copies of each chromosome in the genome. Whole genome duplication can have significant impacts on the evolution, development, and genetic diversity of organisms.

One area of interest in single-cell analysis is the characterization of allele and haplotype-specific copy number aberrations using single-cell sequencing. Copy number variations (CNVs) are genomic alterations that involve the duplication or deletion of large segments of DNA. These CNVs can have significant implications for gene dosage and expression, and are known to contribute to human genetic variation and disease (Handsaker et al. 2015). However, traditional methods for detecting CNVs are limited in their ability to distinguish between the two homologous chromosomes in humans, providing a limited view of tumor heterogeneity and evolution (Zaccaria & Raphael 2021).

To address this limitation, several computational methods have been developed to infer allele and haplotype-specific copy numbers in single cells and subpopulations of cells. One such method is CHISEL (copy-number haplotype inference in single cells using evolutionary links), which aggregates sparse signals across thousands of individual cells to infer allele and

haplotype-specific copy numbers (Zaccaria & Raphael 2021). CHISEL has been successfully applied to single-cell sequencing datasets from breast cancer patients, revealing extensive allele-specific CNAs, including copy-neutral LOH, whole-genome duplications, and mirrored-subclonal CNAs (Zaccaria & Raphael 2021).

The study by Zaccaria & Raphael (2019) introduced CHISEL as the first method to infer allele and haplotype-specific copy numbers in single cells and subpopulations of cells (Zaccaria & Raphael 2021). They applied CHISEL to 10 single-cell sequencing datasets from breast cancer patients and identified extensive allele-specific CNAs, including copy-neutral LOH and whole-genome duplications (Zaccaria & Raphael 2021). This study demonstrated the ability of CHISEL to provide a more comprehensive view of tumor heterogeneity and evolution compared to traditional methods (Zaccaria & Raphael 2021).

Another study by Handsaker et al. (2015) analyzed whole-genome sequence data from the 1000 Genomes Project to identify large multi-allelic CNVs and investigate their impact on gene dosage and expression (Handsaker et al. 2015). They found that mCNVs contribute significantly to gene-dosage variation and generate abundant variation in gene expression (Handsaker et al. 2015). This study highlights the importance of considering allele and haplotype-specific copy numbers in understanding genetic variation and disease (Handsaker et al. 2015).

Zahng et al. (2021) reanalyzed single-cell sequencing data from breast cancer patients using the CHISEL algorithm and identified large-scale allele-specific CNAs that were uncharacterized by previous total copy-number analysis (Zhang & Sjöblom 2021). This study

further demonstrates the utility of CHISEL in uncovering allele-specific CNAs in cancer genomes (Zhang & Sjöblom 2021).

Wu et al. (2021) developed Alleloscope, a method for allele-specific copy number estimation that can be applied to single-cell DNA and ATAC sequencing data (Wu et al. 2021). They applied Alleloscope to gastric, colorectal, and breast cancer samples and found pervasive occurrence of highly complex, multi-allelic copy number aberrations (Wu et al. 2021). This study highlights the interplay between somatic copy number aberrations and chromatin remodeling in shaping the clonal diversity of cancers (Wu et al. 2021).

In summary, single-cell analysis and the characterization of allele and haplotype-specific copy number aberrations have provided valuable insights into genomic heterogeneity and tumor evolution. Computational methods such as CHISEL and Alleloscope have enabled the inference of allele and haplotype-specific copy numbers, revealing the complex landscape of CNAs in cancer genomes. These findings have important implications for understanding genetic variation, disease progression, and the development of targeted therapies.

4.2 Method

Input:

1. n cells with m allelic positions (or bins like in Zaccaria & Raphael (2021))

2. For each cell i, we have Read-Depth Rations (RDR)

$$x_i = (x_{i,1}, ..., x_{i,m}) \in \mathbb{R}^{>} \geq 0 \tag{4.1}$$

and B-allele frequencies (BAF)

$$y_i = (y_{i,1}, ..., y_{i,m}) \in \mathbb{R}^{\geqslant} \geq 0 \tag{4.2}$$

. These numbers can be estimated using an EM approach of (Zaccaria & Raphael 2021).

Find:

1. For each cell i, integer copy numbers

$$a_i = (a_{i,1}, ..., a_{i,m}) \in \mathbb{N}^{\geqslant} \tag{4.3}$$

and

$$b_i = (b_{i,1}, ..., b_{i,m}) \in \mathbb{N}^{\geqslant} \tag{4.4}$$

where $a_{i,t}$ and $b_{i,1}$ are the numbers of copies of the alleles on the first and second haplotype of the cell at the position of t respectively.

2. Character-based phylogeny T describing the evolutionary relations between copy number aberration events.

It is assumed that input and output parameters are related as follows:

$$\gamma_i x_{i,t} = a_{i,t} + b_{i,t} \tag{4.5}$$

$$y_{i,t} = \frac{mina_{i,t}, b_{i,t}}{a_{i,t} + b_{i,t}} \tag{4.6}$$

where $i = 1, ..., n; t = 1, ..., m$ and γ_i is the unknown cell-specific scale factor. These relations do not allow to straightforwardly estimate haplotype-specific copy numbers yet: first, because γ_i are unknown; second because $a_{i,t}$ and $b_{i,t}$ are symmetric in these formulas and therefore phasing is ambiguous; and third, because of the integrality of copy numbers the equations (1) - (2) may have no feasible solutions and the equations should be relaxed and treated as approximate. Therefore, we assume that instead of (2) $y_{i,t}$is drawn from distribution with the density function

$$G(y_{i,t}|\frac{min(a_{i,t}, b_{i,t})}{a_{i,t} + b_{i+t}}) \tag{4.7}$$

with the mean

$$\mu = \frac{mina_{i,t}, b_{i+t}}{a_{i,t} + b_{i+t}} \tag{4.8}$$

in (Zaccaria & Raphael 2021) normal distribution is used.

4.2.1 Generation of candidate haplotypes

We construct an $m + 2-$partite digraph $H = (H_1, H_2, ..., H_{m+2})$ as follows:

- The first part H_1 consists of sources $s_i : i = 1, ..., n$ corresponding to the cells.
- Each of the next m parts $H_2, ..., H_{m+1}$ corresponds to genomic positions. The part $t+1$ consists of vertices (a_t, b_t) representing haplotype-specific copy numbers, $1 < a_t, b_t < C_max$ where C_max is the maximal allowed copy number.
- The $m + 2$th part H_{m+2} consists of a sink r.
- There are all possible edges between parts t and $t + 1$

Any (s_i, r)-path in this graph correspond to a pair of vectors of haplotype-specific copy numbers of the cell i. In order to rank the paths, we introduce vertex and edge weights as follows:

- For each cell i and vertex (a_t, b_t), the corresponding vertex weight is defined as

$$g_i(a_t, b_t) = -log(G(y_{i,t}|\frac{mina_t, b_t}{a_t + b_t})) \tag{4.9}$$

- For each cell i and edge $((a_t, b_t), (a'_t, b'_t)) \in E(H)$, the corresponding edge weight is

$$f_i((a_t, b_t), (a'_t, b'_t)) = ||\frac{a_t + b_t}{x_{i,t}} - \frac{a'_{t+1} + b'_{t+1}}{x_{i,t+1}}||_p^p \tag{4.10}$$

The formula (2.10) comes from the fact that for every cell i according to (1) $\frac{a_{i,t}+b_{i,t}}{x_{i,t}} = \gamma_i$, where γ_i does not depend on t. Thus, we do not have to find γ_i like in (Zaccaria & Raphael 2021), but just need to ensure that in a most likely solution $\frac{a_{i,t}+b_{i,t}}{x_{i,t}}$ are approximately equal for all genomic positions t.

In addition to weight functions f_i and g_i, we can introduce the third vertex weight function h that assign lower weights to balanced copy number pairs $(1, 1)$, $(2, 2)$, $(3, 3)$, ... to reflect high prevalence of whole-genome duplications (WGDs).

Next, for each cell i we find k shortest paths or k bottleneck shortest paths $P_i = (P_{i,1}, ..., P_{i,k})$ between si and r with respect to weights f_i and g_i (or $g_i + h$). It can be easily done using a modified Dijkstra's algorithm. Each path correspond to a candidate pair of haplotype-specific copy number vectors.

For typical single cell sequence data, the graph H has the structure and a size of moderate neural network. Thus, all calculations can be efficiently implemented and parallelized, thus

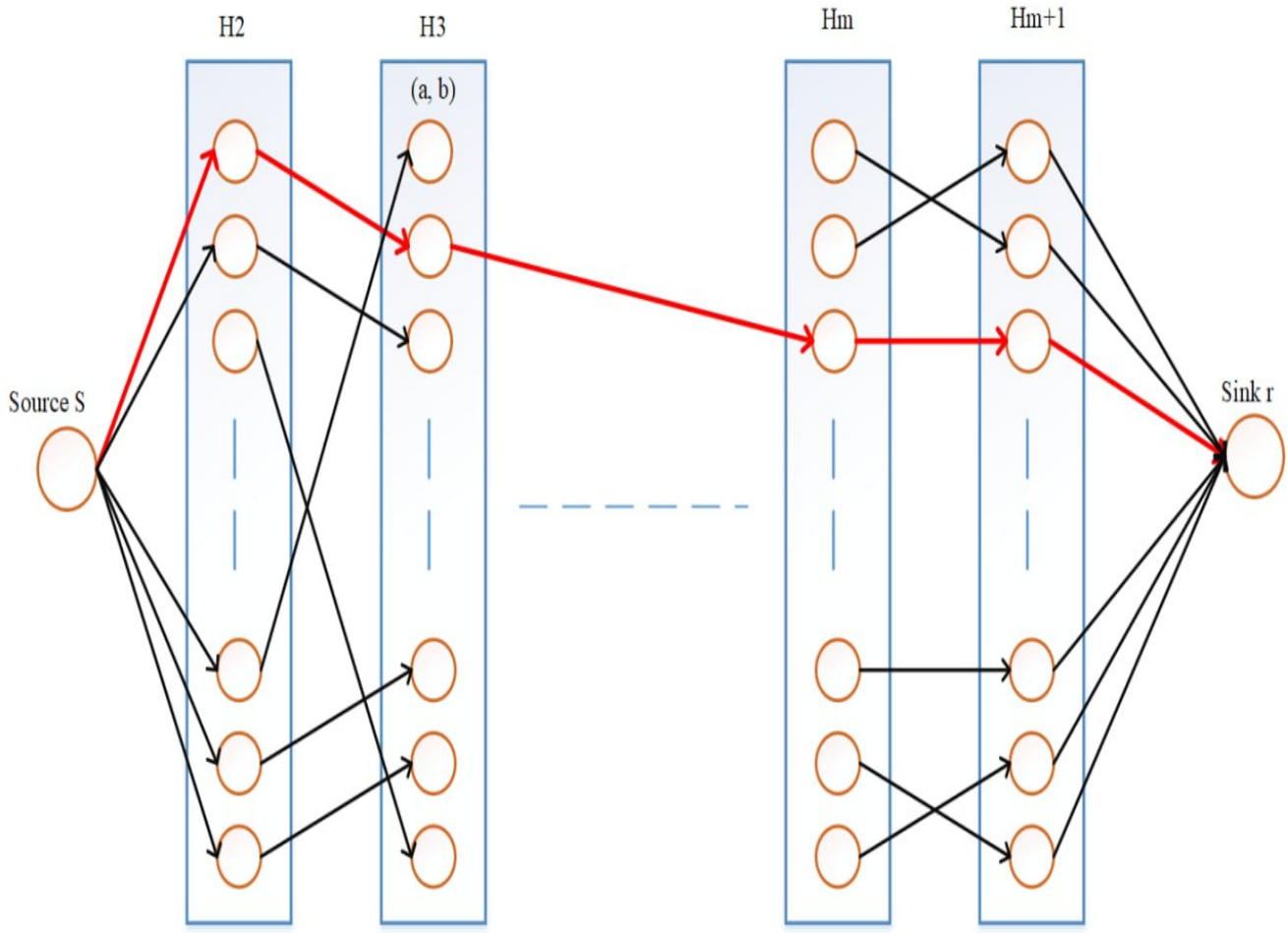

Figure 4.2 Schematic of the proposed phasing algorithm for finding allele and haplotype specific copy numbers

making this approach scalable.

4.2.2 Finding optimal copy number vectors and the corresponding phylogeny

The next goal is to select paths from the sets of candidate paths that define a phylogeny with required properties.

1. Perfect phylogeny. The first natural assumption is that all copy number aberration

events occur at each genomic position at most once. For a collection of paths $P^* = (P_{1,i1}, ..., P_{n,in})$ in order to guarantee this the following condition should be satisfied:

(a) for every genomic position $t = 1, ..., m$, the paths from P^* cover at most one vertex from $H_{t+1} \setminus \{1_t, 1_t\}$.

(b) for any two genomic positions $t1$, $t2$ and vertices (a_{t1}, b_{t1}), (a_{t2}, b_{t2}), $a_{t1} + b_{t1} \geq 3, a_{t2} + b_{t2} \geq 3$, paths from P^* do not cover all four $2-$subsets of vertices of the set $\{(1_{t1}, 1_{t1}), (1_{t2}, 1_{t2}), (a_{t1}, b_{t1}), (a_{t2}, b_{t2})\}$

Such subset of paths can be easily found using Integer Linear Programming (ILP). Specifically, we introduce a bipartite graph $Q = (Q_1, Q_2)$ with $Q = P_1 \cup ... \cup P_n$, $Q_2 = \{1, ..., m\}$ and vertices t and P_{i,j_i} being adjacent, whenever P_{i,j_i} covers a vertex (a_t, b_t) with $a_t + b_t \geq 3$. In order to satisfy b the sub-graph of Q induced by the paths of P^* and their neighbors should not contain a $W-$sub-graph. It is easy to enumerate such sub-graphs in Q; assume that the characteristic vectors of their Q_1-parts are $w_1, ..., w_L$. Introduce the binary variables $Z_{i,j}$ that is equal to 1 if and only if $P_{i,j} \in P^*$. Then the problem can be represented via ILP as follows:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} cost(P_{i,j}) z_{i,j} \to min \tag{4.11}$$

$$\sum_{j=1}^{k} z_{i,j} = 1; i = 1, \cdots, n \tag{4.12}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_{l,i,j} z_{i,j} \leq 2; l = 1, \cdots, L \tag{4.13}$$

4.3 Results

We introduced a novel approach based on graph theory for deducing copy numbers that are specific to alleles and haplotypes. In our method, we constructed a multi-partite directed graph where nodes represent potential copy numbers at various genomic regions (bins). Nodes originating from different genomic regions are linked, while nodes within the same genomic region are not connected. Both nodes and edges in the graph carry weights, resulting in two types of weights: node weights and edge weights. These weights are computed using data from read depth ratios and B-allele frequencies. Notably, we employed a Bayesian approach to estimate B-allele frequency signals.

The process begins with individual single-cell signals starting at a source node labeled as s and traversing the multi-partite graph, ultimately concluding at a sink node marked as r. Our approach primarily revolves around finding the shortest path from the source to the sink, accomplished using the Dijkstra algorithm. Once we have characterized the allele and haplotype specific copy numbers for each genomic region, we construct a phylogenetic tree based on this information. This tree aids us in gaining insights into the evolutionary history of mutations.

4.4 Future Work

While existing methods, including the one we proposed in our dissertation, offer valuable insights into allele and haplotype-specific copy numbers, there is room for further improvement, especially in terms of accuracy, scalability, and adaptability to diverse genomic regions and data types. Recent advancements in deep learning and attention mechanisms present an exciting avenue for enhancing the capabilities of copy number inference methods.

The future work aims to leverage state-of-the-art deep learning architectures and attention mechanisms to enhance the inference of allele and haplotype-specific copy numbers from single-cell data. Here are the key components of this research direction:

1. Transformer-Based Models:

Exploring the integration of Transformer-based models, originally designed for natural language processing tasks, into the context of single-cell data analysis. Transformers have shown remarkable success in capturing long-range dependencies in sequential data, which can be relevant for analyzing genomic data.

2. Graph Attention Networks (GATs):

Adapting Graph Attention Networks (GATs) to work with your multi-partite graph representation. GATs can learn node and edge weights dynamically, which can be advantageous for modeling complex relationships between genomic regions and single-cell signals.

3. End-to-End Learning:

Developing an end-to-end deep learning architecture that takes raw single-cell data as input and directly predicts allele and haplotype-specific copy numbers. This eliminates the

need for manual feature engineering and preprocessing, making the method more adaptable to different datasets.

Expected Outcomes: The expected outcomes of this future work include a powerful and adaptable deep learning-based method for allele and haplotype-specific copy number inference. This method should provide improved accuracy, scalability, and ease of use, enabling researchers to gain deeper insights into genomic variations in single cells.

Conclusion: By embracing the capabilities of deep learning and attention mechanisms, this future work aims to push the boundaries of copy number inference in single-cell data analysis, ultimately contributing to a better understanding of genetic diversity and evolution at the single-cell level.

Appendices

CHAPTER A

Software Developed

A Motif Discovery

This software is designed for the discovery of regulatory motifs in NGS sequences. It employs graph theory algorithms to identify patterns and repetitions that are biologically significant. The tool is specially useful in comparative genomics and functional genomic studies.

Github Repository: `https://github.com/HosseinSaghaian/Motif_Discovery`

B SCGraphCNA

SCGraphCNA is a tool developed for analyzing single-cell genomic data. It assists in the detection and analysis of copy number variations in single cells, providing insights into genomic instability and heterogeneity within cell populations.

Github Repository: `https://github.com/HosseinSaghaian/SC_Copy_Number`

REFERENCES

10x Genomics. 2019, Chromium Single Cell V(D)J Reagent Kits with Feature Barcoding technology for Cell Surface Protein, 10x Genomics, document Number CG000186 Rev A

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. 2015, Nature biotechnology, 33, 831

Andreatta, M., Lund, O., & Nielsen, M. 2013, Bioinformatics, 29, 8

Ash, R. B. 2012, Information theory (Courier Corporation)

Bailey, T. L., Elkan, C., et al. 1994

Blanchette, M., & Tompa, M. 2003, Nucleic acids research, 31, 3840

Bratkovič, T. 2010, Cellular and molecular life sciences, 67, 749

Bron, C., & Kerbosch, J. 1973, Communications of the ACM, 16, 575

CD Genomics. 2023, library preparation, `https://www.cd-genomics.com/blog/principle-and-workflow-of-illumina-next-generation-sequencing/`, accessed: 2023-12-03

Choi, I.-G., Kwon, J., & Kim, S.-H. 2004, Proceedings of the National Academy of Sciences, 101, 3797

Christiansen, A. et al. 2015, Scientific reports, 5, 1

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. 2004, Genome research, 14, 1188

Dinkel, H. et al. 2014, Nucleic acids research, 42, D259

Doe, J. 2023, Phage Display: Finding the One in a Million, https://www.the-scientist.com/methods/phage-display-finding-the-one-in-a-million-71509, accessed: 2023-12-04

Gerasimov, E., Zelikovsky, A., Măndoiu, I., & Ionov, Y. 2017, BMC bioinformatics, 18, 1

Geysen, H. M., Rodda, S. J., & Mason, T. J. 1986, Molecular immunology, 23, 709

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. 2007, Genome biology, 8, 1

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. 2015, Nature genetics, 47, 296

Ionov, Y., & Rogovskyy, A. S. 2020, Plos one, 15, e0226378

Kim, T., Tyndel, M. S., Huang, H., Sidhu, S. S., Bader, G. D., Gfeller, D., & Kim, P. M. 2012, Nucleic acids research, 40, e47

Knittelfelder, R., Riemer, A. B., & Jensen-Jarolim, E. 2009, Expert opinion on biological therapy, 9, 493

Kon, M. A., Fan, Y., Holloway, D., & DeLisi, C. 2007, in Sixth International Conference on Machine Learning and Applications (ICMLA 2007), IEEE, 573–580

Krejci, A., Hupp, T. R., Lexa, M., Vojtesek, B., & Muller, P. 2016, Bioinformatics, 32, 9

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. 1993, science, 262, 208

Lexogen. 2023, Addition of fluorescently labelled nucleotide and identifying the fluorophore, https://www.lexogen.com/rna-lexicon-next-generation-sequencing/, ac-

cessed: 2023-12-03

Liu, X. S., Brutlag, D. L., & Liu, J. S. 2002, Nature biotechnology, 20, 835

Macdougall, I. C., Rossert, J., Casadevall, N., Stead, R. B., Duliege, A.-M., Froissart, M., & Eckardt, K.-U. 2009, New England Journal of Medicine, 361, 1848

Murphy, K., & Weaver, C. 2016, Janeway's immunobiology (Garland Science)

Nielsen, M., & Lund, O. 2009, BMC bioinformatics, 10, 296

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., & Lund, O. 2004, Bioinformatics, 20, 1388

Pietrokovski, S. 1996, Nucleic acids research, 24, 3836

Rentero Rebollo, I., Sabisz, M., Baeriswyl, V., & Heinis, C. 2014, Nucleic acids research, 42, e169

Rodi, D. J., Janes, R. W., Sanganee, H. J., Holton, R. A., Wallace, B., & Makowski, L. 1999, Journal of molecular biology, 285, 197

Roepcke, S., Grossmann, S., Rahmann, S., & Vingron, M. 2005, Nucleic acids research, 33, W438

Roth, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. 1998, Nature biotechnology, 16, 939

Schones, D. E., Sumazin, P., & Zhang, M. Q. 2005, Bioinformatics, 21, 307

Siddharthan, R., Siggia, E. D., & Van Nimwegen, E. 2005, PLoS computational biology, 1, e67

Smith, G. P. 1985, Science, 228, 1315

Smith, G. P., & Petrenko, V. A. 1997, Chemical reviews, 97, 391

Thom, G. et al. 2006, Proceedings of the National Academy of Sciences, 103, 7619

Tomita, E., Tanaka, A., & Takahashi, H. 2006, Theoretical computer science, 363, 28

Tong, A. H. Y. et al. 2002, Science, 295, 321

Van Regenmortel, M. H. 2019, in HIV/AIDS: Immunochemistry, Reductionism and Vaccine Design (Springer), 39–56

Wang, L.-F., & Yu, M. 2004, Current drug targets, 5, 1

Wang, T., & Stormo, G. D. 2003, Bioinformatics, 19, 2369

Wu, C.-Y., Lau, B. T., Kim, H. S., Sathe, A., Grimes, S. M., Ji, H. P., & Zhang, N. R. 2021, Nature biotechnology, 39, 1259

Zaccaria, S., & Raphael, B. J. 2021, Nature biotechnology, 39, 207

Zhang, X., & Sjöblom, T. 2021, Pharmaceuticals, 14, 57

Zhong, L., Coe, S. P., Stromberg, A. J., Khattar, N. H., Jett, J. R., & Hirschowitz, E. A. 2006, Journal of Thoracic Oncology, 1, 513