Faculty of Engineering and Information Technology
University of Technology Sydney

Data Mining In Epigenetic Modification and Gene Expression

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Zhixun Zhao

October 2020

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I, Zhixun Zhao declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

[Zhixun Zhao]

DATE: 15th October, 2020

PLACE: Sydney, Australia

Acknowledgments

First and foremost, I want to thank my supervisor, Prof. Jinyan Li, for his extensive instruction and patient guidance during the last three and a half years, regardless of my research or daily life. His guidance helped me improve my research skills, such as scientific writing, academic communication, and presentation. I appreciate all his contributions of time, ideas, and funding to make all of this thesis possible.

I want to thank my advisor, Prof. Liang Fang, who advised me at my home university in China. His strong support and help gave me the chance to win the CSC (China scholarship council) scholarship and study overseas. I thank Prof. Shaoqing Li and Prof. Jihua Chen, who supervised me when I was pursuing my master's degree, for their initial guidance on my research. I thank my co-supervisor, Prof. Fang Chen, for her help in my research and providing me the opportunity to participate in industry projects.

My sincere thanks also go to my research team members: Dr. Hui peng, Dr. Yi Zheng, Dr. Yuansheng Liu, Dr. Chaowang Lan, Xiaocai Zhang, Xuan Zhang, Tao Tang, and Tian Lan, for their help in both my research and life. It's my honor to be one member of my research team, which is more like a family. Thank you all for bringing me the feeling of home and the unforgettable memories. I have special thanks to Dr. Huijun Wu, who is a truly loyal friend to me. We live together for two years and many thanks for your company.

I am also grateful to acknowledge the funding sources, including the tuition fee and living expenses provided by the China Scholarship Council

and Graduate Research School, travel funds provided by the Faculty of Engineering and Information Technologies, and vice-chancellor funding. Thanks to all staff of the Advanced Analytics Institute and School of Computer Science who provide services and conveniences to my study and research in UTS.

Lastly, I owe special thanks to my wife Zhujun Xue, my parents, and my parents-in-law. It's your love and encouragement that support me in completing my Ph.D. study. I dedicate this work to you all. Especially, I want to thank my wife for her understanding and sacrifice all these years. Love you forever.

Zhixun Zhao
October 2020 @ UTS

Contents

List of Figures

List of Tables

List of Publications

The journal and conference papers published during my PhD study are listed as follows:

Related to the Thesis :

1. **Z. Zhao**, H. Peng, J. Li et al. Imbalance learning for the prediction of N 6-Methylation sites in mRNAs [J]. BMC genomics, 2018, 19(1), 574.

2. **Z. Zhao**, H. Peng, J. Li et al. Identification of lung cancer gene markers through kernel maximum mean discrepancy and information entropy[J]. BMC Medical Genomics, 2019, 12(8): 1-10.

3. **Z. Zhao**, X. Zhang, J. Li et al. Accurate prediction of DNA N^4-methylcytosine sites via boost-learning various types of sequence features[J]. BMC genomics, 2020, 21(1), 1-11.

Others :

4. X. Zhang, **Z. Zhao**, Y. Zheng, and J Li. Prediction of Taxi Destinations Using a Novel Data Embedding Method and Ensemble Learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2019.

5. H. Peng, Y. Zheng, **Z. Zhao**, J. Li et al. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mis-match distributions

[J]. Bioinformatics, 2018, 34 (17), i757-i765.

6. X. Zhang, Y. Liu, Y. Zheng, **Z. Zhao**, J Li et al. Distinction Between Ships and Icebergs in SAR Images Using Ensemble Loss Trained Convolutional Neural Networks [C]. Australasian Joint Conference on Artificial Intelligence. Springer, 2018: 216-223.

7. Y. Zheng, H. Peng, X. Zhang, **Z. Zhao**, J. Li, et al. Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases [J], BMC Bioinformatics, 2018, 19(S19).

8. Y. Zheng, H. Peng, X. Zhang, **Z. Zhao**, J. Li, et al. Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces [J], BMC bioinformatics, 2019, 20(23): 605.

9. Y. Zheng, H. Peng, X. Zhang, **Z. Zhao**, J. Li, et al. DDI-PULearn: a novel positive-unlabeled learning method for large-scale prediction of drug-drug interactions [J], BMC bioinformatics, 2019, 20(19): 1-12.

Abstract

This thesis employs data mining techniques to discover domain knowledge in epigenetic modification and gene expression profile. Computational methods are developed for three research questions, namely, how to accurately predict DNA N^4-methylcytosine site, how to precisely identify mRNA N^6-methyladenosine sites, and how to identify lung cancer gene expression profile markers. The motivations of the proposed methods are improving the performance of computational methods via constructing efficient feature space, optimizing machine learning schemes, solving the data imbalance issue, and employing novel statistical analysis approach to provide researchers efficient computational tools.

DNA N^4-methylcytosine (4mC) is a critical epigenetic modification and plays various roles in the restriction-modification system. The computational methods have been explored to identify 4mC in the DNA sequence in recent years due to the high cost of experimental laboratory detection. However, the state-of-the-art methods have limited performance because of the lack of effective sequence features and the ad hoc choice of learning algorithms. Chapter 3 proposes a new method with novel sequence feature space and machine learning scheme. In sequence encoding, five essential sequence features are integrated into a 292-dimension feature space, representing both global and local sequence characteristics. Then a feature selection scheme is built, where the feature importance score produced from the training process of XGBoost machine is taken as the criterion of feature selection. At last, an SVM-based prediction model is trained with the selected features

and optimized by 10-fold cross-validations. In the result part, the impact of feature selection on model performance is evaluated by an independent test. The proposed method outperforms three state-of-art predictors in both independent test and 10-fold cross-validation. Furthermore, two case studies prove the effectiveness of our method in practical situations.

N^6-methyladenosine (m^6A) widely involves in mRNA metabolism and embryogenesis. Multiple computational human mRNA m^6A site predictors have been developed. However, there are two main drawbacks of the existing methods: first, inadequate learning of the imbalanced training data; second, the sequence text features are not outstanding in representing m^6A sequence characteristics. Chapter 4 proposes to use the cost-sensitive learning idea to solve the imbalance data issues in the problem. This cost-sensitive approach learns from the entire imbalanced dataset without a random selection of negative samples. In sequence representation, site location, entropy features and specific single nucleotide polymorphism (SNP) positions are taken as new features, which improve the performs significantly. In the comparison with existing predictors, our method achieves better correctness and robustness in both independent tests and case studies. The results suggest that imbalance learning is promising to improve the performance of m^6A prediction.

The early diagnosis of lung cancer has been a challenging problem in clinical practice for a long time. The identification of differentially expressed genes as a disease marker is a promising solution. Chapter 5 presents a novel approach to identify marker genes and define the boundary of gene expression profile for human lung cancer. By calculating the kernel maximum mean discrepancy, the proposed method evaluates the expression difference between normal, normal adjacent to tumor (NAT) and tumor samples. The expression level boundaries among different groups are defined with the information entropy theory for marker genes. Compared with two conventional methods t-test and fold change, the genes selected by MMD values have better performance under all metrics in 10-fold cross-validation. Furthermore, the GO and KEGG enrichment analysis validate the discovered

marker gene in function pathways. At last, we choose ten most meaningful genes as lung cancer markers and calculate the expression profile boundaries. The proposed method is more accurate than conventional DEA methods in marker gene identification and provides a reliable method for defining the gene expression level boundaries.

Chapter 1

Introduction

This chapter presents the research background, questions, contributions, and the structure of the thesis. In Section 1.1, the epigenetic modifications, including DNA and RNA base methylations, gene expression biomarker, and the application of data mining techniques in bioinformatics are introduced. Then the research questions and contributions of the thesis are discussed in Section 1.2 and Section 1.3, respectively. At last, the thesis structure is described in Section 1.4.

1.1 Background

This thesis presents my research topic: data mining applications in epigenetic modifications and gene expression profile biomarker. This section introduces the background knowledge of epigenetic modifications, gene expression biomarker, and data mining techniques in bioinformatics.

1.1.1 Epigenetic modification

Before the DNA was discovered as the molecule of genetic information, scientists had noticed that the genes were partially active in different organisms, despite that they shared the same genetic information (Kanwal & Gupta 2012). Introduced by Conrad Waddington in the early 1940s,

epigenetics was described as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being." (Waddington et al. 1939). In the original definition, epigenetics refers to molecular mechanisms that modulate the expression of genetic information into observable phenotype (Dupont, Armant & Brenner 2009). In many cases, the epigenetic gene expressions change the heritable phenotype without alterations in the primary DNA sequence (Esteller 2008). With the development of genetics, the definition of genetics was redefined and generally accepted as "the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in the DNA sequence." (Morris et al. 2001)

Epigenetics today refers to changes that impact the gene activity and expression, resulting from normal cell development or external environmental factors(Portela & Esteller 2010). As a critical inherited regulation method, epigenetics is required to be heritable in the progeny (Berger, Kouzarides, Shiekhattar & Shilatifard 2009). Epigenetic modifications are generally occurring in the DNA base, amino acids of histone proteins, nucleosome, and various types of RNAs. On a molecular level, the most widely studied epigenetic modifications can be categorized into the following groups: DNA methylation, histone modifications, nucleosome positioning, non-coding RNA, and message RNA methylation. In this thesis, the DNA methylation and message RNA methylation site detection are discussed.

DNA methylation

The covalent DNA modifications can modulate the gene expression that has been discovered by scientists since 1969 (Griffith & Mahler 1969). The domain modifications in mammalian DNA are methylation of cytosine, adenine, and guanine methylation (Hotchkiss 1948, Ratel, Ravanat, Berger & Wion 2006, Das & Singal 2004). Although the cytosine methylation mostly occurs in the context of CpG islands (Straussman, Nejman, Roberts, Steinfeld, Blum, Benvenisty, Simon, Yakhini & Cedar 2009), recent researches

find that cytosines in non-CpG sequences can also be methylated (Woodcock, Crowther & Diver 1987, Ramsahoye, Biniszkiewicz, Lyko, Clark, Bird & Jaenisch 2000). DNA methylation is catalyzed by a group of enzymes, named DNA methyltransferases (DNMTs) (Rodríguez-Paredes & Esteller 2011), known as DNMT1, DNMT1b, DNMT1o, NNMT1p, DNMT2, DNMT3A, DNMT3b with isoforms and DNMT3L (Bestor, Laudano, Mattaliano & Ingram 1988, Leonhardt, Page, Weier & Bestor 1992, Robertson 2002). In the dynamic methylation process, the enzymes that demethylate DNA base includes 5-methylcytosine glycosylase (Frémont, Siegmann, Gaulis, Matthies, Hess & Jost 1997) and MBD2b (Bhattacharya, Ramchandani, Cervoni & Szyf 1999), which removes the methylated group from cytosine. According to the location where methylated group occurs in the DNA sequence, there are many kinds of DNA base methylation, in which 5-Methylcytosine (5mC), N^6-methyladenine (6mA) and N^4-methylcytosine (4mC) are the most common types (Davis, Chao & Waldor 2013, Korlach & Turner 2012, Roberts, Vincze, Posfai & Macelis 2015).

RNA methylation

In the central dogma, the genetic information flows from DNA to RNA and then to histone protein. The reversible epigenetic modifications in DNA (Suzuki & Bird 2008, Kohli & Zhang 2013, Jones 2012, Bhutani, Burns & Blau 2011) and histone (Shi 2007, Klose, Kallin & Zhang 2006, Bird 2001) has been studied for decades. Although RNA has more than 100 chemical modifications, the function of most modifications are still uncertain (Grosjean, Benne et al. 1998, Grosjean & Grosjean 2005, Machnicka, Milanowska, Osman Oglou, Purta, Kurkowska, Olchowik, Januszewski, Kalinowski, Dunin-Horkawicz, Rother et al. 2012, Motorin & Helm 2011). The RNA modification was thought to be fixed and unalterable, and not considered as an epigenetic modification for a long time. RNA N^6-methyladenosine (m^6A) is the first discovered reversible RNA modifications with functional proteins 'writer', 'eraser' and 'reader' in the past several years

(Fu, Dominissini, Rechavi & He 2014, Jia, Fu, Zhao, Dai, Zheng, Yang, Yi, Lindahl, Pan, Yang et al. 2011, Zheng, Dahl, Niu, Fedorcsak, Huang, Li, Vgb, Shi, Wang, Song et al. 2013). m^6A has been widely detected in message RNA (mRNA) and long non-coding RNAs in higher eukaryotes (Krug, Morgan & Shatkin 1976, Schibler, Kelley & Perry 1977). Another highly-regarded RNA modification is 5-methylcytidine (m^5C), which exists extensively in tRNA and rRNA (Helm 2006, Squires & Preiss 2010). The recent evidence shows that the m^5C is dynamically modulated in the cellular response to stress (Chan, Dyavaiah, DeMott, Taghizadeh, Dedon & Begley 2010, Schaefer, Pollex, Hanna, Tuorto, Meusburger, Helm & Lyko 2010), and regulating the translation rates (Chow, Lamichhane & Mahto 2007).

1.1.2 Gene expression biomarker

Biomarker

Biomarkers are features that can be objectively detected and evaluated as a measurable indicator for biological, pathological, or therapeutic intervention pharmacological responses (Group, Atkinson Jr, Colburn, DeGruttola, DeMets, Downing, Hoth, Oates, Peck, Schooley et al. 2001). Generally, biomarkers are biomolecules found in the blood, other body fluids, or tissues of the human body. They are signs of normal or abnormal processes, or signs of illness and disease (Capelozzi 2009, Marshall, Bowman, Yang, Fong & Berg 2013, Vazquez, Koizumi, Henschke & Yankelevitz 2007). There are many types of biomarkers, including proteins, such as enzymes or receptors, nucleic acids, mRNA or other non-coding RNA, antibodies and peptide sequences, etc. Other landmark changes, such as gene expression, proteome and metabolome can also be used as biomarker (Rabinowits, Gerçel-Taylor, Day, Taylor & Kloecker 2009, Montani, Marzi, Dezi, Dama, Carletti, Bonizzi, Bertolotti, Bellomi, Rampinelli, Maisonneuve et al. 2015, Nagrath, Sequist, Maheswaran, Bell, Irimia, Ulkus, Smith, Kwak, Digumarthy, Muzikansky et al. 2007, Sozzi, Boeri, Rossi, Verri, Suatoni, Bravi, Roz, Conte, Grassi,

Sverzellati et al. 2014, Valenti, Huber, Filipazzi, Pilla, Sovena, Villa, Corbelli, Fais, Parmiani & Rivoltini 2006).

There are many potential applications of biomarkers, and the widest application at present is to assist the early diagnosis and prognosis of diseases. Such markers include not only related biochemical indicators but also genetic indicators, which can reflect potential conditions and the development process of the disease. Furthermore, biomarkers can be used to assess the severity of the disease, evaluate the efficacy of treatment and detect side effects of drugs (Peters, Walters & Moldowan 2005). In the field of new drug development, biomarkers can also reflect the effectiveness of drugs, evaluating the pharmacological effects of the reaction of targeted drugs with receptors or enzymes, and providing useful guidance in the clinical use of treatment (Vargas & Harris 2016).

Genetic marker

Modern medical research has found that most diseases occur due to the joint effect of genetic and environmental factors. Gene variation plays a direct or indirect role in the occurrence of many diseases. At present, it is noted that the presence of diseases, such as primary hypertension, diabetes, mental illness, and tumors involves the mutation of multiple genes (Cowley 2006, Flannick & Florez 2016, Smyth, Plagnol, Walker, Cooper, Downes, Yang, Howson, Stevens, McManus, Wijmenga et al. 2008, Tarailo-Graovac, Shyr, Ross, Horvath, Salvarinova, Ye, Zhang, Bhavsar, Lee, Drögemöller et al. 2016, Jonsson, Stefansson, Steinberg, Jonsdottir, Jonsson, Snaedal, Bjornsson, Huttenlocher, Levey, Lah et al. 2013). By detecting the genetic markers related to these diseases, we can assess the risk of the patient's morbidity and intervene in the early stage, and delay or avoid the occurrence of the disease. The presence of most diseases is accompanied or directly caused by changes at the gene expression profiles. The changes often start from the beginning of the disease, when the body may have no apparent symptoms. It is difficult to diagnose the disease via conventional examination

methods. Therefore, early diagnosis and treatment of diseases can be achieved by detecting changes of genetic markers. The cure rate of high-risk diseases and the quality of patients' life can be improved significantly.

In recent decades, the pharmacogenomics studies have discovered many genetic variations related to drug transport, distribution, metabolism, or drug targets (McCarthy & Hilfiker 2000, Isla, Sarries, Rosell, Alonso, Domine, Taron, Lopez-Vivanco, Camps, Botia, Nunez et al. 2004). These genetic variations may lead to changes in the activities of corresponding drug metabolic enzymes, transporters, or drug-receptor proteins, causing individual differences in drug response. Therefore, the identification of related gene markers before the patient takes the drug can help explore a reasonable method to optimize the dosing regimen according to the patient's genetic characteristics, maximize the drug's efficacy, and minimize the adverse reactions (Crews, Hicks, Pui, Relling & Evans 2012, Daly 2010, Ma & Lu 2011). Besides, most patients with tumors face the risk of tumor metastasis or recurrence after surgery and medication. For different patients, the prognosis is related to the treatment plan and the patient's genetic background. Therefore, it would be possible to predict the prognosis of patients and guide clinical advancement by detecting relevant genetic markers.

1.1.3 Data mining in bioinformatics

In this thesis, two data mining techniques, including machine learning and statistic analysis, are applied to discover knowledge in epigenetic modification and gene expression profile. As powerful tools, data mining techniques have been extensively employed in computational biology. This section is a brief introduction of data mining technique and its applications in bioinformatics.

Data Mining

With the rapid increase of large, complex datasets, knowledge discovery in information-rich data has been a big challenge in science, engineering, and

business (Chakrabarti, Ester, Fayyad, Gehrke, Han, Morishita, Piatetsky-Shapiro & Wang 2006). Data mining (also named Knowledge Discovery in Database, KDD) (Fayyad, Piatetsky-Shapiro & Smyth 1996) is the process of discovering potential significant information and knowledge from the big quantitative, incomplete, noisy, fuzzy and random practical application data, with information technologies such as database, artificial intelligence, statistics, visualization, and parallel computing (Clifton 2010, Han, Pei & Kamber 2011).

Generally, the data mining process has the following steps: (1)Problem statement and hypothesis formulation; (2) Data collection and integration; (3) Data prepossessing including outlier detection, scaling, encoding and feature selection; (4) Knowledge mining based on appropriate data mining model; (5) Model interpretation and knowledge representation (Kantardzic 2011, Han et al. 2011). The primary task of data mining includes: (1) Classification and regression, developing the predictive learning model to class data items into predefined labels, or map data items to a prediction variable. (2) Clustering, categorizing data into a finite set of groups. (3) Summarization, describing data with a compact representation. (4) Dependency modeling, associate rules, and dependency patterns discovery in data. (5) Deviation detection, analyzing a few extreme cases of analysis to reveal the fundamental reason for the change.

Applications in bioinformatics

Bioinformatics is a research field that builds methods and software tools for understanding biological data, especially large and complex data (Baxevanis, Bader & Wishart 2020, Raza 2012). Some of the grand research area of bioinformatics contains the image and signal processing in molecular biology, sequencing and annotating genome in genetics, text mining in biological literature, gene and protein expression and regulation, and biological system modeling.

With the great progress in information technology, data mining has been

applied in following key researches: (1) Semantic integration of heterogeneous and distributed genetic databases (Sujansky 2001); (2) Protein structure prediction (McGuffin, Bryson & Jones 2000, Baker & Sali 2001); (3) DNA sequence similarity search and alignment (Madden 2013, Mount 2007, Noé & Kucherov 2005); (4) Multiple sequence alignment (Edgar 2004, Chenna, Sugawara, Koike, Lopez, Gibson, Higgins & Thompson 2003, Katoh & Standley 2013); (5) Association analysis and pathway analysis (Luo, Peng, Zhu, Dong, Amos & Xiong 2010, Torkamani, Topol & Schork 2008); (6) Biological data visualization (Gómez, García, Salazar, Villaveces, Gore, García, Martín, Launay, Alcántara, Del-Toro et al. 2013, Chen, Chen, He & Xia 2018); (7) Biomedical text mining (Cohen & Hersh 2005, Zweigenbaum, Demner-Fushman, Yu & Cohen 2007); (8) Gene microarray data analysis (Lock, Hermans, Pedotti, Brendolan, Schadt, Garren, Langer-Gould, Strober, Cannella, Allard et al. 2002, Quackenbush 2006); (9) Biomedical data mining based on privacy protection (Holzinger & Jurisica 2014, Jiang, Zhao, Wang, Malin, Wang, Ohno-Machado & Tang 2014).

1.2 Research questions

This thesis mainly focuses on two research topics: computational epigenetic modification detection and disease genetic biomarker discovery. Under these two topics, three research questions are raised to be solved, and the detailed formulations are illustrated below in Q1 to Q3.

Q1: DNA N^4-methylcytosine prediction

From the previous section, the N^4-methylcytosine has a dynamic methylation process regulated by enzymes, and the methylated group will change the physical and chemical properties of the local sequence. Thus, nucleotides in the flanking window around the 4mC site have specific sequence patterns, which is the basis of computational detection. In order to identify whether a cytosine site in the DNA sequence is N^4-methylcytosine or not, a local

sequence includes the flanking windows of cytosine is taken as the input of algorithm. The DNA fragment sequence has fixed length and target cytosine in the middle position, where nucleotides types include 'A', 'G', 'T', and 'C'.

For such DNA sequence S and target C, the DNA 4mC prediction problem can be expressed as the following formula:

$$f(S,C) = \begin{cases} 1 & \text{if the } C \text{ is } \mathrm{N}^4\text{-methylcytosine site} \\ 0 & \text{else} \end{cases} \tag{1.1}$$

where $f(S,C)$ is the trained binary classifier to label the target cytosine as positive (1, if the target cytosine is methylated) or negative (0, otherwise). The binary classifier is trained with labeled sequences, where the inputs of the algorithm are feature vectors of the sequence patterns, including local and global characteristics, and labels are experimentally validated status of the corresponding cytosine. From the above descriptions, the following factors are essential in the problem: (1)Sequence feature extraction. How to define the sequence patterns and convert the sequence into feature vectors (Chen, Yang, Feng, Ding & Lin 2017, He, Jia & Zou 2019). (2)Feature dimension reduction. Select the significant feature dimensions and build an effective feature space (Wei, Luan, Nagai, Su & Zou 2019). (3)Classification algorithm with high efficiency. The performance of the classifier is related to the data characteristics, and a proper classifier is vital to achieving better performance(Manavalan, Basith, Shin, Lee, Wei, Lee et al. 2019, Hasan, Manavalan, Shoombuatong, Khatun & Kurata 2020).

Q2: mRNA N^6-methyladenosine prediction

Like the DNA N^4-methyladenosine prediction, the mRNA N^6-methyladenosine prediction is based on sequence patterns around $\mathrm{m}^6\mathrm{A}$ site. The mRNA fragment sequence centered with adenosine is taken as an algorithm sample. The nucleotides in mRNA including 'A', 'G', 'C', and 'U'. Since the motif is discovered for $\mathrm{m}^6\mathrm{A}$ as [G/A/C] [G/A] A^* C [U/A/C] (where A^* stands for the $\mathrm{m}^6\mathrm{A}$ site) (Dominissini, Moshitch-Moshkovitz, Schwartz,

Salmon-Divon, Ungar, Osenberg, Cesarkas, Jacob-Hirsch, Amariglio, Kupiec et al. 2012, Meyer, Saletore, Zumbo, Elemento, Mason & Jaffrey 2012), the adjacent nucleotides near target site in input mRNA sequence should be conserved to the motif.

For a such mRNA sequence S and target A, the m^6A prediction problem can be expressed as the following formula:

$$f(S, A) = \begin{cases} 1 & \text{if the } A \text{ is } N^6\text{-methyladenosine site} \\ 0 & \text{else} \end{cases} \tag{1.2}$$

Where $f(S, A)$ is learned binary classifier to determine whether the target adenosine is methylated or not, the inputs of the classifier are feature vectors extracted from the mRNA sequence, and the labels are still divided into positive(1) and negative(0). The main problems in m^6A prediction question are: (1)Valid sequence features (Chen, Feng, Ding, Lin & Chou 2015, Chen, Tran, Liang, Lin & Zhang 2015). Besides the text features of the mRNA sequence, some biological characteristics should be considered in the feature space, such as adenosine location, single nucleotide polymorphism(SNP) variants, and entropy information . (2)Efficient site identification in flanking window. For biological features such as SNP feature, not all sites near m^6A make sense in the classification, and how to identify the efficient positions is an important question. (3)Imbalance learning (Zhou, Zeng, Li, Zhang & Cui 2016). As the number of non-m^6A sites (adenosine site that conserves to motif but is not methylated) is much larger than m^6A sites in the training data, the machine learning algorithm should cope with the imbalance problem.

Q3: Identification of lung cancer gene markers

The gene markers identification is a knowledge discovery task based on gene expression profiles produced by next-generation sequencing (NGS) technology. In this question, there are two main steps: (1) Identify the differential expressed gene. (2) Provide an expression level boundary for

each marker gene. For a gene expression profiles matrix $G = (g_1, g_2, ..., g_n)$, these two steps can be expressed as the following formula:

$$F(E) = \left\{ g_1^{'}, g_2^{'}, ..., g_m^{'} \right\} \tag{1.3}$$

where $F(E)$ is an analysis algorithm to select the gene subset, containing differential expressed genes g'_1, g'_2, ..., g'_m. Then the gene expression boundary of lung cancer is identified as:

$$\begin{cases} lower\ boundary_1 \leq g_1^{'} \leqslant upper\ boundary_1 \\ lower\ boundary_2 \leq g_2^{'} \leqslant upper\ boundary_2 \\ \qquad\qquad ... \\ lower\ boundary_m \leq g_m^{'} \leqslant upper\ boundary_m \end{cases} \tag{1.4}$$

Where *lowerboundary* and *upperboundary* are the expression profiles boundary of the corresponding gene in lung cancer samples, the challenges for these questions are (1) Gene differential expression analysis (DEA) algorithm. Currently, the DEA methods are mainly based on the statistical model under predefined distribution hypothesis, but it's hard to determine the statistical distribution of gene expression profiles in all situations (Soneson & Delorenzi 2013, Seyednasrollah, Laiho & Elo 2013, Rapaport, Khanin, Liang, Pirun, Krek, Zumbo, Mason, Socci & Betel 2013). (2) In cancer studies, the control group is normally collected from the normal tissue adjacent to tumors (NAT). However, recent research suggests that NAT tissue is not strictly equal to healthy samples, indicating that true healthy tissue should be taken into considerations (Aran, Camarda, Odegaard, Paik, Oskotsky, Krings, Goga, Sirota & Butte 2017). (3) There should be a valid method for the expression boundary estimation when the expression edge between two tissue types is not clear.

1.3 Research contributions

To solve the above three research questions, we have proposed novel methods described in Chapter 3 to Chapter 5. In each chapter, computational methods are built and optimized based on the main problems of the research question. The contributions of this thesis are summarized as follows (C1 to C3).

C1: Accurate DNA N^4-methylcytosine prediction

In Chapter 3, we propose a pre-computed method for accurate DNA N^4-methylcytosine prediction with novel sequence feature space and machine learning algorithm with a feature selection scheme. The main contributions are: (1) In this study, the sequence logos of training samples are firstly analyzed, and the particular patterns of adjacent nucleotides are observed. Along with the global features such as one-hit binary encode and sequential nucleotide frequency, the sequence features focused on local characteristics. Three corresponding features, such as k-nucleotide frequency, k-spectrum nucleotide frequency, and PseDNC, are extracted into feature space. (2) Unlike the existing predictors (He et al. 2019, Wei, Luan, Nagai, Su & Zou 2019), which only applied an F-score based feature selection, a novel embedding feature selection scheme is proposed. An XGBoost machine is trained to calculate the feature dimension importance with information entropy theory, more meaningful than F-score. Cross validations are conducted to evaluate the selected feature subset. (3)According to state-of-art researches, the SVM machine is efficient in predicting 4mC sites. The trained model is assessed via not only independent test and 10-fold cross-validation but also case studies from practical situations to avoid the model over-fitting.

C2: Accurate mRNA N^6-methyladenosine prediction

Chapter 4 presents a computational method for accurate mRNA N^6-methyladenosine prediction with novel sequence and biological features and

an imbalance machine learning algorithm. The contribution of this work is three folds: (1) Instead of extracting sequence text features only, we introduce three novel m^6A biological features, including location information, sequence entropy information, and single nucleotide polymorphism (SNP) variants. Together with the sequence features such as one-hit binary encode, k-mer adjacent nucleotides frequency, and chemical property with density, the three novel features make up an efficient feature space. (2) For the SNP feature, we adopt the Fisher's exact test and max-Relevance Min-Redundancy algorithms to identify the significant SNP positions in m^6A flanking windows. The identification of particular positions helps to reduce the feature dimensions and improve the feature efficiency for classification. (3) We learn a high-performance machine learning model based on a weighted XGBoost classifier to cope with the highly imbalanced training data. Other than the model trained with selecting negative samples and balanced training data, our model is more meaningful for the practical situations of m^6A prediction.

C3: Gene marker identification for lung cancer

In Chapter 3, we proposed a method for lung cancer gene marker identification through kernel maximum mean discrepancy and information entropy. The contributions of this part work include: (1) Inspired by transfer learning, we propose to apply kernel maximum mean discrepancy(kernel MMD) to conduct the gene differential expression analysis. The kernel MMD method calculates an MMD score to evaluate the degree of differential expression, and the genes with high degrees are regarded as the marker gene. (2) We introduce the true normal tissue in lung cancer marker identification for the first time. The experiment groups are divided into true normal, normal adjacent to the tumor, and tumor tissues. The differential expressed genes are identified between the three groups. (3) We present an information theory-based method for gene expression boundary detection. The information theory method can give a certain threshold for diagnosis when there is no

distinct gene expression gap between normal and tumor tissues.

1.4 Thesis structure

This thesis is composed of six chapters. The structure of this thesis is shown in Figure 1.1 and introduced briefly below:

Chapter 1 introduces the background knowledge of the research problems. The research questions and contributions are also described. The thesis structure is illustrated at the end of the chapter. **Chapter 2** reviews the related work about the studied research problems. The current research progress of the mentioned research problems is included. **Chapter 3** to **Chapter 5** describes the proposed three novel methods for solving the research questions, including accurate prediction of DNA 4mC sites, imbalance learning for mRNA m^6A prediction, and lung cancer gene marker identification. The details of method construction, evaluation, and experiment are presented in these chapters. **Chapter 6** provides a conclusion of this thesis, and future work is also discussed.

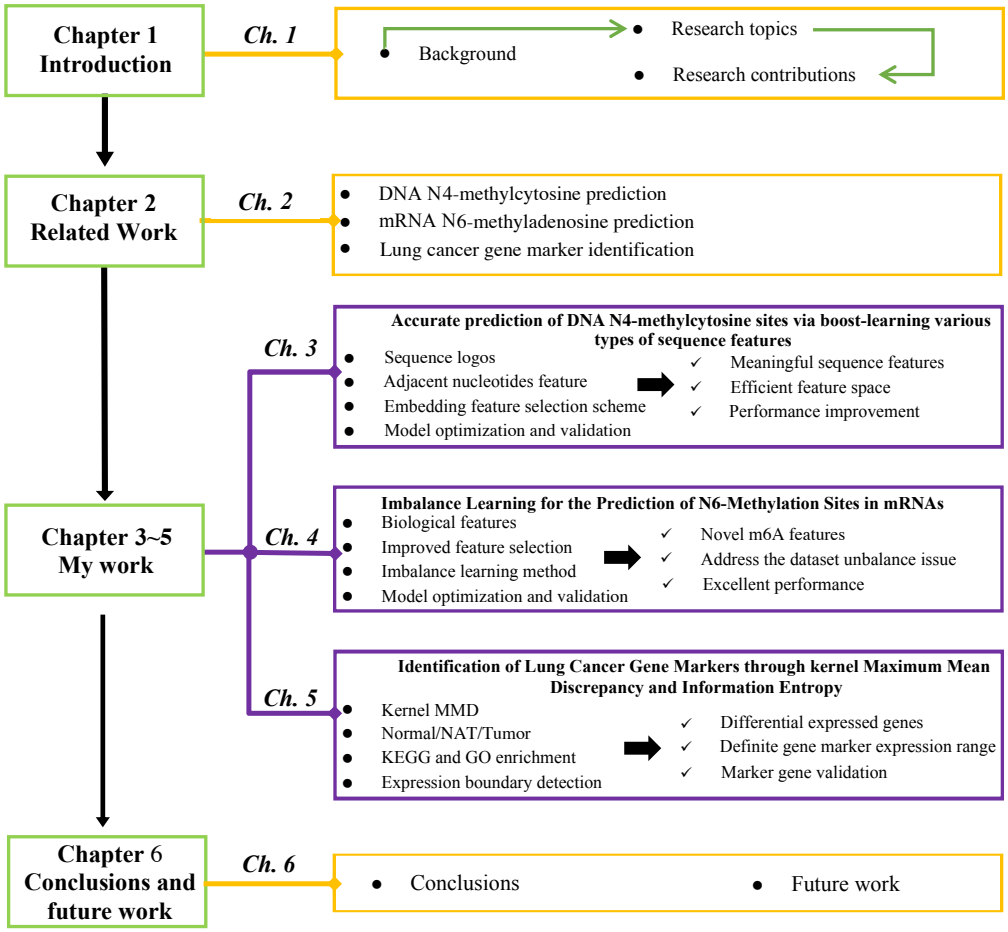

Figure 1.1: **Thesis structure.** It consists of the following four parts: introduction, related work, my work, and conclusions and future work. Short introduction of each part is shown in the right side.

Chapter 2

Related work and literature review

This chapter reviews the related work and literature of the research topics in this thesis. In Section 2.1, the development process of DNA N^4-methylcytosine is described. Then, the wet-lab and computational methods of the N^6-methylcytosine identification are introduced. In Section 2.3, the published lung cancer biomarkers and gene differential expression analysis methods are presented. Lastly, a brief summary is discussed in Section 2.4.

2.1 DNA N^4-methylcytosine prediction

As an essential epigenetic modification, DNA base methylation expands the DNA content and plays crucial roles in regulating various cellular processes (Rathi, Maurer & Summerer 2018, Stoiber, Quick, Egan, Lee, Celniker, Neely, Loman, Pennacchio & Brown 2016, Chen, Zhao & He 2016). According to the location where methylated group occurs in the DNA sequence, there are many kinds of DNA base methylation, in which N^5-Methylcytosine (5mC), N^6-methyladenine (6mA) and N^4-methylcytosine (4mC) are the most common types (Davis et al. 2013, Korlach & Turner 2012, Roberts et al. 2015). 5mC occurs at the C5-position of cytosine

and is the dominant methylation type in eukaryotic genomes, involving in differentiation, gene expression, genomic imprinting, preservation of chromosome stability, aging, suppression of repetitive element, and X chromosome inactivation (Robertson 2005, Jin, Li & Robertson 2011, Jones 2012, Tahiliani, Koh, Shen, Pastor, Bandukwala, Brudno, Agarwal, Iyer, Liu, Aravind et al. 2009). In prokaryotes, 6mA and 4mC constitute the majority of DNA base methylations (Heyn & Esteller 2015). 6mA occurs at the N6-position of adenine and is a marker in gene regulation, development, DNA replication, repair, and expression (Fu, Luo, Chen, Deng, Yu, Han, Hao, Liu, Lu, Doré et al. 2015, Greer, Blanco, Gu, Sendinc, Liu, Aristizábal-Corrales, Hsu, Aravind, He & Shi 2015, Zhang, Huang, Liu, Cheng, Liu, Zhang, Yin, Zhang, Zhang, Liu et al. 2015). 4mC exists at the N4-amino group of cytosine and participates in the restriction-modification system that provides a bacterial immune response against occupying DNA, DNA repair, expression, and replication (Cheng 1995, Modrich 1991, Messer & Noyer-Weidner 1988). Compared to 5mC and 6mA, the further biological function of 4mC is less studied for the lack of sufficient detection methods.

The precious location of DNA base methylation was a hard problem in the past for a long time. It is not affordable to locate the DNA 5mC on a large scale until the whole-genome bisulfite sequence, and the next generation sequence techniques were developed (Cokus, Feng, Zhang, Chen, Merriman, Haudenschild, Pradhan, Nelson, Pellegrini & Jacobsen 2008, Lister, O'Malley, Tonti-Filippini, Gregory, Berry, Millar & Ecker 2008). The detection of 6mA and 4mC in whole-genome became available after the Single-molecule real-time sequencing (SMRT) technology was introduced (Davis et al. 2013, Flusberg, Webster, Lee, Travers, Olivares, Clark, Korlach & Turner 2010). Then a next-generation sequence method called 4mC-Tet-assisted-bisulphite-sequencing and another method with engineered transcription-activator like effectors were developed for 4mC identification (Yu, Ji, Neumann, Chung, Groom, Westpheling, He & Schmitz 2015). However, the experimental methods were of high cost and

cannot identify 4mC on a large scale. Recently, the rapid development of machine learning provides a promising computational approach to address classification problems in bioinformatics, and researchers have explored using computational methods to identify 4mC sites in the DNA sequence.

Based on data collected from public SMRT sequence experiments, Ye built the first DNA 6mA and 4mC database named MethSMRT for 156 species (Ye, Luan, Chen, Liu, Xiao & Xie 2016). In (Chen, Yang, Feng, Ding & Lin 2017), Chen's team firstly constructed high-quality DNA 4mC benchmark datasets for six species and proposed an SVM based prediction model called iDNA4mC with the nucleotide chemical property and sequential nucleotide frequency features. Based on the benchmark datasets, 4mCPred and 4mcPred-SVM were built to improve the site prediction performance (He et al. 2019, Wei, Luan, Nagai, Su & Zou 2019). In 4mCPred, the authors used two new features PSTNP and EIIP with a simple feature selection. Wei's team built 4mcPred-SVM with four kinds of sequence features and a two-step feature optimization. These two predictors improved the 4mC site prediction by introducing new sequence features and feature selection. Recently, some other predictors have been developed to identified 4mC site in the DNA sequence for Mouse (Manavalan, Basith, Shin, Lee, Wei, Lee et al. 2019, Hasan et al. 2020), Escherichia coli (Lv, Wang, Ding, Zhong & Xu 2020), Rosaceae (Hasan, Manavalan, Khatun & Kurata 2019) and so on (Manavalan, Basith, Shin, Wei & Lee 2019, Wei, Su, Luan, Liao, Manavalan, Zou & Shi 2019).

Although these methods made some progress, the state-of-the-art methods have limited performance because of the lack of effective sequence features and the ad hoc choice of learning algorithms to cope with this problem.

2.2 mRNA N^6-methyladenosine prediction

Among more than 140 kinds of post-transcription modifications (PTMs) (Machnicka et al. 2012, Motorin & Helm 2011), N^6-methyladenosine (m^6A),

the methylation occurs at 6th nitrogen of adenosine, is one of the most abundant modifications (Wu, Jiang, Wang & Wang 2016, Fu et al. 2014). This methylation has been widely found in species such as Arabidopsis thaliana, Saccharomyces cerevisiae, bacteria, virus, human, and mouse (Wan, Tang, Zhang, Xie, Zhu, Wang & Lang 2015, Chen, Tran, Liang, Lin & Zhang 2015, Deng, Chen, Luo, Weng, Ji, Zhou & He 2015, Huang, Xiong, Yang, Liu, Yuan & Feng 2015). More exactly, these methylation events have occurred in the mRNAs at the 3' untranslated regions (UTRs) close to the stop codon, following a conserved sequence motif DRACH, [G/A/C] [G/A] A^* C [U/A/C], (where A^* stands for the m^6A site) (Dominissini et al. 2012, Meyer et al. 2012). The dynamic m^6A methylation involves many proteins such as METTL3, METTL14, WTAP, ALKBH5 and YTHDF2 (Wu et al. 2016, Wang, Li, Toth, Petroski, Zhang & Zhao 2014, Liu, Yue, Han, Wang, Fu, Zhang, Jia, Yu, Lu, Deng et al. 2014, Ping, Sun, Wang, Xiao, Yang, Wang, Adhikari, Shi, Lv, Chen et al. 2014).

The functions of m^6A in biological processes have been significantly redefined with the intensive investigation of this dynamic and reversible methylation in mRNAs recently. It is reported that m^6A disruption can affect translation efficiency (Wang, Zhao, Roundtree, Lu, Han, Ma, Weng, Chen, Shi & He 2015), cell viability (Bokar 2005) and cell development (Wang et al. 2014). The level changes of m^6A in mRNA can lead to abnormality of RNA export, protein translation, or RNA editing, causing cancer, obesity, and other human diseases (Shen, Huang, Huang, Xiong, Yang, Wu, Jia, Chen, Feng, Yuan et al. 2015, Yang, Huang, Huang, Shen, Xiong, Yuan, Qin, Zhang, Feng, Yuan et al. 2016, Choi, Ieong, Demirci, Chen, Petrov, Prabhakar, O'leary, Dominissini, Rechavi, Soltis et al. 2016, Tsai, Courtney & Cullen 2018). For example, strong relationships have been observed between m^6A and HIV-1 (Lichinchi, Gao, Saletore, Gonzalez, Bansal, Wang, Mason & Rana 2016, Riquelme Barrios, Pereira-Montecinos, Valiente-Echeverría & Soto-Rifo 2018), Zika virus infection (Lichinchi, Zhao, Wu, Lu, Qin, He & Rana 2016) and breast cancer stem cell

phenotype (Zhang, Samanta, Lu, Bullen, Zhang, Chen, He & Semenza 2016). The identification of m^6A sites is crucial for understanding the disease mechanisms and identifying novel medicine targets.

Experimental approaches including two-dimensional thin-layer chromatography (Keith 1995), high performance liquid chromatography (Zheng et al. 2013), and high-throughput methods (e.g., m^6A-seq (Dominissini et al. 2012) and MeRIP-Seq (Meyer et al. 2012)) have been applied to identify m^6A sites in mRNAs. However, they can only detect m^6A-containing transcript fragments instead of identifying the exact methylated adenines (Liu, Flores, Meng, Zhang, Zhao, Rao, Chen & Huang 2014). Based on the single-nucleotide resolution m^6A maps in mRNAs, researchers have explored computational methods with sequence features and machine learning algorithms to make m^6A sites prediction. For instance, iRNA-Methyl (Chen, Feng, Ding, Lin & Chou 2015), m^6Apred (Chen, Tran, Liang, Lin & Zhang 2015) and RAM-ESVM (Chen, Xing & Zou 2017) are predictors aiming at yeast m^6A site prediction (Schwartz, Agarwala, Mumbach, Jovanovic, Mertins, Shishkin, Tabach, Mikkelsen, Satija, Ruvkun et al. 2013); methods SRAMP (Zhou et al. 2016), Methy-RNA (Chen, Tang & Lin 2017) and RAM-NPPS (Xing, Su, Guo & Wei 2017) are built on human and mouse m^6A maps (Ke, Alemu, Mertens, Gantman, Fak, Mele, Haripal, Zucker-Scharff, Moore, Park et al. 2015, Linder, Grozhik, Olarerin-George, Meydan, Mason & Jaffrey 2015). There are also some predictors developed for Arabidopsis thaliana (Xiang, Yan, Liu, Zhang & Sun 2016, Chen, Feng, Ding & Lin 2016, Wang & Yan 2018). Recently, several new methods haven been developed for RNA m^6A site prediction with machine learning (Liu, Lei, Meng & Wei 2020, Chen, Wei, Zhang, Wu, Rong, Lu, Su, de Magalhaes, Rigden & Meng 2019, Qiang, Chen, Ye, Su & Wei 2018), deep learning (Huang, He, Chen, Chen & Li 2018, Zhang & Hamada 2018) and ensemble learning algorithm (Liu, Lei, Fang, Tang, Meng & Wei 2020). However, there are two main drawbacks to these methods. The first is the inadequate learning of the imbalanced m^6A samples, which are much less than the non-

m^6A samples by their balanced learning approaches. Second, the features used by these methods are not outstanding in representing m^6A sequence characteristics.

2.3 Lung cancer gene markers identification

Small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC) are two main types of lung cancer, comprising the majority of clinic cases (Schnabel & Junker 2015). As the most common cancer, lung cancer is the leading cause of cancer-related deaths all over the world (Parkin 2001, Minna, Roth & Gazdar 2002). However, most lung cancer cases were diagnosed in a very late stage when symptoms like coughing, coughing up blood, shortness of breath, and chest pains appeared. Many early-diagnosed lung cancer cases were detected by accident (Minna et al. 2002, Jemal, Siegel, Ward, Murray, Xu, Smigal & Thun 2006). In the clinic practice, the most widely used examinations for lung cancer are chest radiography and computed tomography (CT), but these two methods require visible and irreversible histological variants in human lung, resulting in rather low sensitivity in the early stage (Fontana, Sanderson, Taylor, Woolner, Miller, Muhm & Uhlenhopp 1984, Frost, Ball Jr, Levin, Tockman, Baker, Carter, Eggleston, Erozan, Gupta, Khouri et al. 1984, Hussain, Khatri, Casali, Batchelor & West 2014). Therefore, it is a crucial issue to find more timely and accurate approaches for lung cancer early-stage diagnosis.

Due to the progress in molecular biology, some molecules which play vital roles in lung cancer development are possible to diagnose cancer and distinguish the specific cancer sub-types (Capelozzi 2009, Marshall et al. 2013, Vazquez et al. 2007). Researchers have explored to identify efficient biomarkers from these molecules as the indicator of the pathogenic process to improve the diagnosis sensitivity (Jantus-Lewintre, Usó, Sanmartín & Camps 2012). These explorations are mainly focused on genetic mutations, DNA methylation profile, miRNA synthesis profile, and especially blood proteins

(Rabinowits et al. 2009, Mitas, Hoover, Silvestri, Reed, Green, Turrisi, Sherman, Mikhitarian, Cole, Block et al. 2003, Andre, Schartz, Movassagh, Flament, Pautier, Morice, Pomel, Lhomme, Escudier, Le Chevalier et al. 2002, Montani et al. 2015, Nagrath et al. 2007, Sozzi et al. 2014, Sozzi, Conte, Leon, Cirincione, Roz, Ratcliffe, Roz, Cirenei, Bellomi, Pelosi et al. 2003, Valenti et al. 2006). Till now, panels of protein markers have been identified and intensively used in clinic applications. For example, the combinations of CEACAM, CYFRA 21-1, ProGRP, CA125, NSE (neuron-specific enolase) and NY-ESO (cancer-testis antigen) are popular lung cancer diagnosis markers (Doseeva, Colpitts, Gao, Woodcock & Knezevic 2015, Goetsch 2011, Mizuguchi, Nishiyama, Iwata, Nishida, Izumi, Tsukioka, Inoue, Uenishi, Wakasa & Suehiro 2007, Pujol, Grenier, Daurès, Daver, Pujol & Michel 1993, Okada, Nishio, Sakamoto, Uchino, Yuki, Nakagawa & Tsubota 2004). Recently, researchers also discovered that β-chain of human haptoglobin (Kang, Sung, Ahn, Park, Lee, Park & Cho 2011), SAA (serum amyloid A) (Sung & Cho 2008), APOA1 (apolipoprotein A-1) (Maciel, Junqueira, Paschoal, Kawamura, Duarte, Carvalho & Domont 2005) and some other proteins (Indovina, Marcelli, Maranta & Tarro 2011) may be potential biomarkers. Despite the advances in protein marker discovery, some disadvantages of protein markers are still existing, like genetic heterogeneity of tumors, poor reproducibility of laboratory tests, and low concentration of the proteins (Zamay, Zamay, Kolovskaya, Zukov, Petrova, Gargaun, Berezovski & Kichkailo 2017, Sozzi et al. 2003). In recent years the next-generation sequence technologies have promoted the study of disease-related genomes. Projects like The Cancer Genome Atlas (TCGA) (Tomczak, Czerwińska & Wiznerowicz 2015) and the Genotype-Tissue Expression (GTEx) (Lonsdale, Thomas, Salvatore, Phillips, Lo, Shad, Hasz, Walters, Garcia, Young et al. 2013) have collected a large number of sequencing experiments and provided tissue-specific gene expression data in public. As some genes have distinct expression levels between normal and tumor tissues for the reason of disease development, they are promising to diagnose lung

cancer more timely and accurately.

During the past years, gene differential expression analysis (DEA) has been extensively applied in the pre-process of high-throughput profiling data collected from micro-arrays (Wang, Gerstein & Snyder 2009, Marioni, Mason, Mane, Stephens & Gilad 2008, Bullard, Purdom, Hansen & Dudoit 2010). Based on statistical models, researchers developed tools to identify genes that had distinct expression levels between different experimental groups. Compared with the micro-array data, the RNA-seq raw data comes with the unique feature of discrete reads, which should be analyzed under an appropriate statistical hypothesis (Soneson & Delorenzi 2013). According to the statistical hypothesis, the existing RNA-seq analysis models can be categorized into the Poisson model (Kvam, Liu & Si 2012, Bullard et al. 2010), negative binomial model, beta-binomial model (Robinson, McCarthy & Smyth 2010, Anders & Huber 2010), and Bayesian model (Hardcastle & Kelly 2010, Seyednasrollah et al. 2013, Rapaport et al. 2013). These models can tell whether the gene expression levels are the same between experiment groups and calculate a confidence coefficient scores (also named p-value) suggesting the magnitude of expression difference. However, the most existing gene differential expression analysis (DEA) methods have two main drawbacks: First, these methods are based on fixed statistical hypotheses and not always effective; Second, these methods can not identify a certain expression level boundary when there is no obvious expression level gap between control and experiment groups.

2.4 Summary

This chapter reviews the state-of-art researches for the research problems, including DNA N^4-methylcytosine prediction, mRNA N^6-methylcytosine prediction, and lung cancer gene marker identification. The development of these problems and the drawbacks of the existing methods are introduced.

Generally, the computational methods have been explored to reduce

the wet-lab experiment cost, and the state-of-art methods cannot achieve satisfactory performance. There are still problems in the existing researches. For the epigenetic base modification detection problems, the most common computational technique is machine learning, where the problem is defined as binary classification problems for machine learning algorithms (Chen, Yang, Feng, Ding & Lin 2017, Chen, Feng, Ding, Lin & Chou 2015). The limitations of these methods mainly lie in the lack of meaningful feature representation and efficient machine learning schemes. Besides, the key problems of gene marker identification include the significant general statistic DEA theory and gene expression boundary identification approach.

Chapter 3

Accurate prediction of DNA N^4-methylcytosine sites via boost-learning various types of sequence features

3.1 Background

As mentioned in Section 2.1, the core idea of the previous DNA N^4-methylcytosine prediction method is to transform 4mC-contained DNA sequences into various features as the input of the machine learning algorithms. However, these features are not adequate to make the prediction methods to achieve excellent performance. Through the analysis of the sequence logos, we observe that the adjacent nucleotides' characteristics are potentially essential. We extract the contiguous nucleotides sequence characteristics like k-nucleotide frequency, k-spectrum nucleotide pair frequency, and PseDNC as features to describe the sequences. Besides, two global sequence features, one-hot binary, and sequential nucleotide frequency are also merged into our feature space. As global features have the complete information of DNA sequence and the local features can underline specific

sequence patterns, the combined feature space is highly expected to improve the prediction performance.

Since feature selection can reduce the feature space dimension and the modeling complexities (Wei & Billings 2006), two of the existing 4mC prediction methods 4mCPred and 4mCPred_SVM both employed a feature selection scheme based on the F-score and sequential forward search (SFS) strategy (He et al. 2019, Wei, Luan, Nagai, Su & Zou 2019). Although the F-score can evaluate the feature importance according to the relevance between the feature and label, the performance of the selected feature subset was not good enough. In this paper, we propose an embedded feature selection scheme, in which features are ranked with the feature importance scores derived by the XGBoost classifier (Chen & Guestrin 2016) training process. Supported by information entropy theory, the feature importance here is more meaningful than F-score. Then lower-ranked features are removed one by one, each round with a cross-validation assessment on the performance of the selected feature subset.

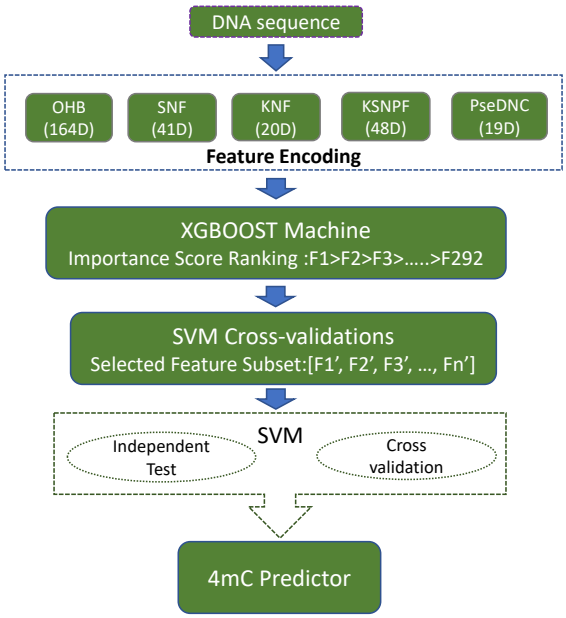

Figure 3.1: **Framework of proposed model construction**

The flowchart of our approach is shown in Figure 3.1, where the new sequence feature space and feature selection scheme is depicted for DNA 4mC site prediction. First, the DNA sequence is encoded into five kinds of features, a total of 292 dimensions. Second, an XGBoost machine is trained, and the feature importance scores from the training are used to rank all the features. Last, an SVM-based prediction model is built, and the parameters are optimized with 10-fold cross-validation.

In the results section, we firstly analyze the feature importance in our feature space and show that feature selection improves the model performance significantly in the independent test. Besides, we compare the proposed method with the three state-of-art methods, iDNA4mC, 4mCPred, and 4mCPred_SVM in independent test and 10-fold cross-validation on benchmark datasets, and the proposed method can achieve much better performance. Two detailed case studies for 4mC site prediction on the dlk-1 and DSCAM genes partly prove the effectiveness of our approach in practical situations.

3.2 Materials and methods

3.2.1 Benchmark datasets

From the DNA 4mC database MethSMRT (Ye et al. 2016), Chen and his team constructed the benchmark databases containing *Caenorhabditis elegans (C.elegans), Droso-phila melanogaster (D.melanogaster), Arabidopsis thaliana (A.thaliana), Escherichia coli (E.coli), Geoalkalibacter subterraneus (G.subterraneus) and Geobacter pickeringii (G.pickeringii)* (Chen, Yang, Feng, Ding & Lin 2017). In the benchmark datasets, the 41-bit 4mC-centred DNA sequences were obtained from MethSMRT with a Modification QV threshold of 30. The CD-HIT software (Fu, Niu, Zhu, Wu & Li 2012) was used to remove the redundant positive samples. The same number of negative samples were selected randomly to construct a balanced dataset. The negative samples were also 41-bit cytosine-centered DNA sequences and

were not detected by SMRT. To compare with the existing predictors, we use the same division of the datasets for independent tests. The summary of benchmark datasets is listed in Table 3.1.

Table 3.1: **Summary of six benchmark datasets**

Species	Positive Sample	Negative Sample	Total
C.elegans	1554	1554	3108
D.melanogaster	1769	1769	3538
A.thaliana	1978	1978	3956
E.coli	388	388	776
G.subterraneus	906	906	1812
G.pickeringii	569	569	1138

3.2.2 Feature space construction

The sequence logos of all the six species are plotted using the web tool 'two sample logos' (Crooks, Hon, Chandonia & Brenner 2004) to visualize the difference between the positive and negative sequences. See Figure 3.2. The sequence characteristics are distinct among the six species, especially positions near the 4mC sites that exhibit different patterns in the positive and negative samples. Also, nucleotides not in the near flanking window around 4mC show the difference in different labels. Thus an expanded feature space combining global and local patterns is good to construct accurate models for all the species.

Among the existing methods, iDNA4mC only use nucleotide chemical property and frequency feature, which cannot extract the local adjacent nucleotide patterns; in 4mCPred and 4mCPred_SVM, the features mainly focus on the trinucleotide or dinucleotide nucleotide patterns, ignoring the spectrum nucleotide patterns. In this study, the feature space covers five types of features, one-hot 4-bit binary feature (OHB), sequential nucleotide frequency (SNF), k-nucleotide frequency (KNF), k-spectrum nucleotide pair

frequency (KSNPF) and PseDNC. The OHB and SNF feature possess the information of the whole sequence and represent the global sequential properties, while KNF, KSNPF, and PseDNC features capture the local sequence patterns.

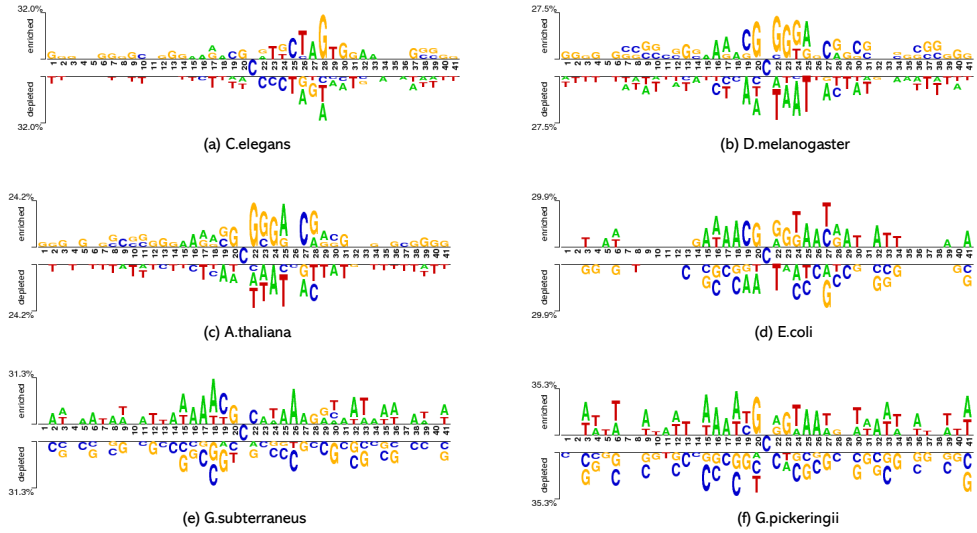

Figure 3.2: **Sequence logos for DNA samples in the benchmark datasets**

One-hot Binary Feature

The one-hot binary feature is the most widely used sequence representation feature. It converts each of the nucleotides in the DNA sequence into a 4-bit vector, which contains only one '1'. The length of the OHB feature is related to the number of nucleotide types and length of the sequence. Since the DNA sample sequence here is 41-bit and has four types of nucleotide, the one-hot binary feature is 164 bits. The encoding rules in this study are as follows: 'A'- (1,0,0,0), 'G'- (0,1,0,0), 'T'- (0,0,1,0), 'C'- (0,0,0,1). From the rule, it is obvious that the OHB feature is sparser than 2-bit or 3-bit binary

features. The one-hot binary feature makes it more reasonable to calculate the importance sore for each dimension in feature space and to discover local motifs.

Sequential Nucleotide Frequency

The sequential nucleotide frequency, also known as nucleotide density, is the frequency that the corresponding nucleotide occurs before the current position. SNF is commonly used together with the binary encoding feature as a global density feature. For an n-bit long sequence, SNF calculates n values for each position in the sequence and produces an n-dimensional feature that starts with '1'. The SNF feature d_i is defined as:

$$d_i = \frac{1}{|S_i|} \sum_{j=1}^{i} f(s_j), f(s_j) = \begin{cases} 1 & s_j = s_i \\ 0 & s_j \neq s_i \end{cases} \tag{3.1}$$

where S_i denotes the length of sequence before the current position i and s_i is the nucleotide at position i. For example, a sequence like *'AACGTACT'* can be converted into the SNF feature vector (1, 0.5, 0.33, 0.25, 0.2, 0.5, 0.28, 0.25).

k-Nucleotide Frequency

The k-nucleotide (k-mer) frequency is a classic concept in DNA sequence encoding. KNF feature is the frequency that adjacent k nucleotides occur in the whole sequence. The length of the KNF feature vector is 4^{k}, determined by the parameter k. The calculation of KNF is as below:

$$F(n_1 n_2 ... n_k) = \frac{C(n_1 n_2 ... n_k)}{S - k + 1} \tag{3.2}$$

where $n_1 n_2 ... n_k$ donates the adjacent k nucleotides and $n_i \in$ *(A, C, G, T)*. F and C is the feature value and total count of the adjacent nucleotides, while S is the length of sequence. When k = 1, the KNF is a vector like (F_A, F_C, F_G, F_T); when k =2, the KNF of a sequence is like (F_{AA}, F_{AC}, F_{AG}, F_{AT},

F_{CA}, F_{CC}, F_{CG}, F_{CT}, F_{GA}, F_{GC}, F_{GG}, F_{GT}, F_{TA}, F_{TC}, F_{TG}, F_{TT}) with a dimension of $4^2 = 16$.

k-Spectrum Nucleotide Pair Frequency

The KSNPF feature depicts the sequence context by calculating the frequency of k-spaced nucleotide pairs (e.g., AXXT is a two-spaced nucleotide pair, and CXXXG is a three-spaced nucleotide pair). Like the adjacent nucleotides pair above, the feature dimension of the KSNPF is 16 for each k. The calculation of this feature is as follows:

$$F(n_1X...Xn_2) = \frac{C(n_1X...Xn_2)}{S-k-1} \tag{3.3}$$

where $n_1X...Xn_2$ donates the k-spaced nucleotides pair and $n_i \in$ *(A, C, G, T).*

PseDNC

As an essential sequence feature, PseDNC combines global and local structural properties and has been widely used in sequence site prediction problems (Chen, Feng, Lin & Chou 2013). For a DNA sequence, the PseDNC feature is a vector:

$$F_{PseDNC} = [d_1, d_1, ...d_{16}d_{16}...d_{16+\lambda}]^T \tag{3.4}$$

where,

$$d_k = \begin{cases} \frac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le k \le 16) \\ \frac{w\theta_\mu - 16}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16 < k \le 16+\lambda) \end{cases} \tag{3.5}$$

where f_k denotes the normalized frequency of two adjacent nucleotide pairs; w is the weight factor, and θ is the correlation factor of j-tier, representing the correlation of all j-tier from the sequence. The definition of θ is:

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Theta_{i,i+j}(j = 1, 2, ..., \lambda; \lambda < L) \tag{3.6}$$

where Θ is the correlation function and given by:

$$\Theta_{i,i+j} = \frac{1}{\mu}\sum_{u=1}^{\mu}\left[P_u(R_iR_{i+1}) - P_u(R_jR_{j+1})\right]^2 \tag{3.7}$$

where μ is the length of sequence; $P_u(R_iR_{i+1})$ is the numerical value of the u-th DNA local property for the adjacent nucleotide pair R_iR_{i+1} at position i. In this study, PseDNC feature is computed by a python package 'repDNA' (Liu, Liu, Fang, Wang & Chou 2015) and the λ value is default to 3. The names of 38 DNA local properties utilized in the definitions here are detailed in the supplementary Table S1 of Additional File 1.

3.2.3 Feature selection scheme

Feature selection can reduce the dimension of feature space and speed up the model training. A lot of feature selection strategies have been employed in machine learning (Li, Cheng, Wang, Morstatter, Trevino, Tang & Liu 2017). In particular, a filter feature selection scheme has been used to improve the prediction performance. The filter feature selection scheme has two steps: first, F-score is calculated for each dimension in feature space according to the relevance between feature and label; second, a selection strategy called SFS is adopted to ascertain the feature subset. In this study, we proposed an embedded feature selection method also with two steps. However, we rank features with importance scores produced from the XGBoost training process (Chen & Guestrin 2016) and select the top features with cross-validations.

In our method, XGBoost is the predefined classifier to analyze the feature importance. XGBoost has been proven to be an efficient tool in data science. In the training process, the XGBoost classifier calculates the feature importance score for each dimension based on the dimension location and the split efficiency in the boosting tree. In this study, XGBoost is implemented with a python package $'xgboost'$ of vision 0.90. The feature importance scores are obtained through the function $'get\_score'$. According to the calculation method, the feature importance score has 5 types: $'weight'$, $'gain'$, $'cover'$,

$'total\_weight'$, $'total\_gain'$ and here we use the default $'weight'$ importance score.

With the importance scores derived by the XGBoost classifier, feature dimensions are ranked from the highest to the lowest. Then the lower-ranked features are removed from the feature space one by one, and the feature subset performance is evaluated by 10-fold cross-validation with a support vector machine. The feature subset with the best performance is taken as the final feature space for 4mC prediction.

3.2.4 Support vector machine

Support vector machine (SVM) is a popular machine learning classifier and has been proven to be more efficient than the other algorithms for DNA 4mC prediction in the state-of-the-art researches (Wei, Luan, Nagai, Su & Zou 2019). In this study, SVM machine is implemented with the python package $'scikit-learn(vision 0.22)'$ (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg et al. 2011). The kernel function of the SVM prediction model is set as a radial basis kernel function (RBF). The hyperparameter C and γ are optimized by a grid search with cross-validations and the search ranges are listed below:

$$\begin{cases} 2^{-5} \leq C \leq 2^{10} & step = 2 \\ 2^{-15} \leq \gamma \leq 2^{2} & step = 2^{-1} \end{cases} \tag{3.8}$$

With the output of the probability scores, the ROV curve can be plotted. The threshold of probability score is set as 0.5 to obtain the predicted label.

3.2.5 Performance evaluation metrics

To compare with the existing predictors, the evaluation metrics in this study are consistent with the state-of-the-art methods, including Sensitivity (Sn), Specificity (Sp), Accuracy (ACC) and Matthews correlation coefficient

(MCC). The definitions of these four metrics are as follows:

$$\begin{cases} Sn = \frac{TP}{TP+FN} \times 100\% \\ Sp = \frac{TN}{TN+FP} \times 100 \\ ACC = \frac{TP+TN}{TP+FN+TN+FP} \times 100 \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \end{cases} \tag{3.9}$$

Sn shows the model capability of identifying positive samples, while Sp tells the capacity of classifying negative samples; ACC is the prediction accuracy of all samples; MCC evaluates the overall performance of a predictor. In this study, the receiver operating characteristic (ROC) curve is also used to analyze model performance. The ROC curve is plotted in a coordinate graph where the x-axis is the false positive rate (1-Sp) and the y-axis is the true positive rate (Sn). The area under the curve(AUC) evaluates the classification performance, and larger AUC means better performance.

3.3 Results

This section reports the feature importance scores obtained from the XGBoost machine and analyzes the influence of the feature selection on prediction performance. Then three state-of-the-art predictors are compared with the proposed method in the independent test and 10-fold cross-validation on benchmark datasets. At last, we present results from two case studies which were conducted to identify the 4mC sites in the C.elegans and D.melanogaster genes.

3.3.1 Feature importance analysis

As stated, five types of sequence features are created to constitute a 292-dimensional feature space. Among the 292 dimensions, OHB is from D1 to

D164; SNF is from D165 to D205; KNF is from D206 to 225; KSNPF is from D226 to 273 and PseDNC is from D274 to D292. The feature importance scores are obtained from the training process of the XGBoost machine. The importance score distributions for all the datasets are illustrated in Figure 3.3. Top 30 feature dimensions are reported in Table S2 of Additional File 1, and feature importance scores of all the feature dimensions are in Additional File 2.

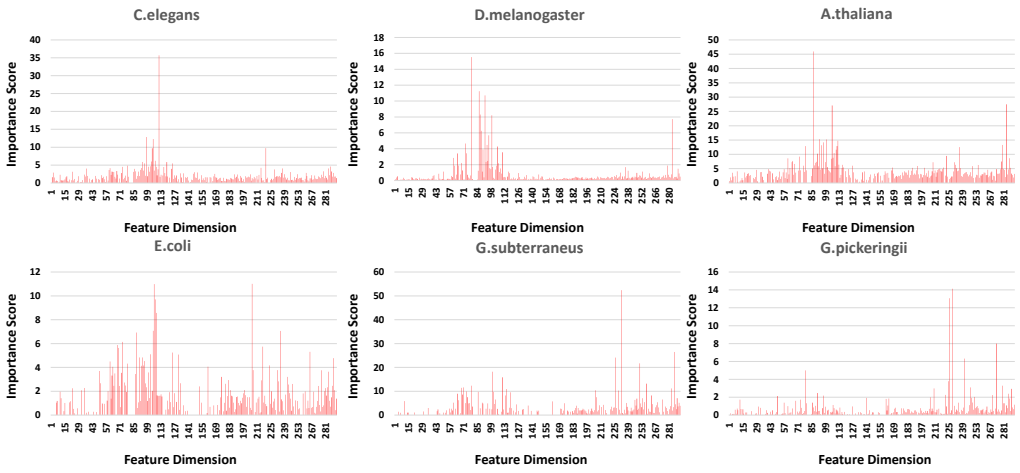

Figure 3.3: **Sequence feature importance distribution**

It is understood that each feature dimension has distinct importance scores in different species. OHB and PseDNC features have relatively high average scores in all species. In particular, OHB features have the highest average score in C.elegans, D.melanogaster, and A.thaliana. KSNPF feature not only gets a high importance score in A.thaliana, E.coli, and G.subterraneus like KNF features but also has the highest average score in G.pickeringii. SNF feature just stands out in E.coli. The features' importance score ranges from 0 to 50, and some feature dimensions' scores are such low that they are less important in the classification and may have noise effects on model performance. Thus, the feature selection before the training is potentially useful to improve model accuracy.

3.3.2 Impact of feature selection on classification

We first evaluate the model performance via the independent test without feature selection before model training. Then the independent test is carried out with feature selection, where the benchmark datasets divisions and SVM parameters are kept the same. Table 3.2 and Figure 3.4 show the independent test performance before and after feature selection.

Table 3.2: **The independent test performanc before and after feature selection**(Sn, Sp and ACC:%)

Datasets	Selection	Sn	Sp	ACC	MCC
C.elegans	before	82.69	75.00	78.85	0.58
	after	94.23	78.85	86.53	0.74
D.melanogaster	before	74.57	77.12	75.85	0.52
	after	84.74	86.44	85.59	0.71
A.thaliana	before	82.57	76.51	79.54	0.59
	after	80.30	83.33	81.81	0.64
E.coli	before	92.30	69.23	80.76	0.63
	after	88.46	88.46	88.46	0.77
G.subterraneus	before	83.33	75.00	79.17	0.59
	after	91.67	81.67	86.67	0.74
G.pickeringii	before	81.57	78.94	80.26	0.61
	after	86.84	89.47	88.15	0.76

The results of independent test after feature selection are improved significantly in all the species. In C.elegans, feature selection improved Sn, Sp, ACC and MCC by 7.54%, 3.85%, 7.74% and 0.16. In D.melanogaster, the model performance has the most considerable improvement by 10.17%, 9.32%, 9.74% and 0.19 for Sn, Sp, ACC and MCC, respectively. For A.thaliana, Sp increased by 6.82% while ACC and MCC slightly increased by 2.27% and 0.05. Besides, Sp, ACC and MCC improved by 9.23%, 7.7% and 0.14 in E.coli dataset. In G.subterraneus, the metrics improvement is

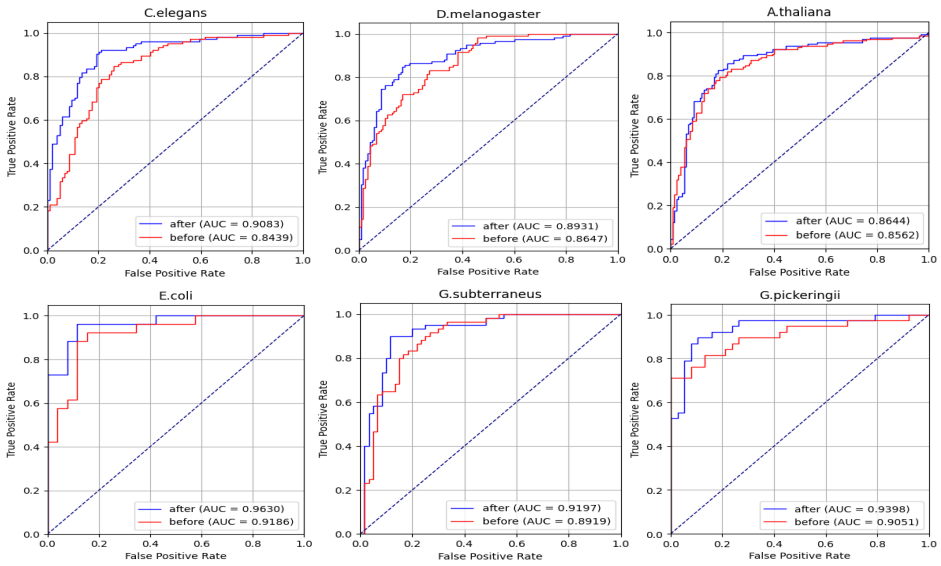

Figure 3.4: **The independent test performanc before and after feature selection**

by 8.34% for Sn, 6.67% for Sp, 7.5% for ACC and 0.15 for MCC. As for G.pickeringii, the performance is improved by 5.17%, 10.73%, 7.89% and 0.15 in terms of Sn, Sp, ACC and MCC with feature selection. From Figure 3.4, it's obvious that the AUCs after feature selection become better in all the species. The most massive AUC growth exists in C.elegans by 0.06 and the least growth is by 0.01 in A.thaliana. The results imply that the proposed feature selection scheme enhances the performance of the SVM model by selecting effective features from the original feature space.

3.3.3 Comparison with state-of-art predictors

Three state-of-the-art DNA 4mC prediction methods, iDNA4mC, 4mCPred, and 4mCPred_SVM are compared with the proposed method. The comparison was conducted using the independent test and cross-validation test on the benchmark datasets.

The independent test results by iDNA4mC and 4mCPred were reported

Table 3.3: **Independent test results on benchmark datasets**(Sn, Sp and ACC:%)

Methods	Datasets	Sn	Sp	ACC	MCC
iDNA4mC	C.elegans	80.77	73.08	76.92	0.54
	D.melanogaster	74.58	77.97	76.27	0.53
	A.thaliana	80.3	77.27	78.79	0.58
	E.coli	96.15	69.23	82.69	0.68
	G.subterraneus	85.00	76.67	80.83	0.62
	G.pickeringii	81.58	78.95	80.26	0.61
4mCPred	C.elegans	85.58	78.85	82.21	0.65
	D.melanogaster	83.90	81.36	82.63	0.65
	A.thaliana	76.52	76.52	76.52	0.53
	E.coli	84.62	80.77	82.69	0.65
	G.subterraneus	91.67	75.00	83.33	0.68
	G.pickeringii	86.84	68.42	77.63	0.56
this study	C.elegans	94.23	78.85	86.53	0.74
	D.melanogaster	84.74	86.44	85.59	0.71
	A.thaliana	80.30	83.33	81.81	0.64
	E.coli	88.46	88.46	88.46	0.77
	G.subterraneus	91.67	81.67	86.67	0.74
	G.pickeringii	86.84	89.47	88.15	0.76

in (He et al. 2019), and we cannot find the independent test results of 4mCPred_SVM method. Since 4mCPred_SVM only provides the final prediction model, it's not available to rebuild the independent test. Thus, here we compare our method with iDNA4mC and 4mCPred in the independent test under the same division of training and testing data. The results of the independent test are presented in Table 3.3. Our method outperforms the other methods in all species. Generally, the proposed method improves ACC from 3.02% to 7.89% and increases MCC from 0.06 to 0.15. Especially, a significant improvement of our approach can be observed in G.pickeringii

(improving Sn by 5.26%, Spby 10.52%, ACC by 7.89%, and MCC by 0.15).

Table 3.4: **Cross-validation results on benchmark datasets**(Sn, Sp and ACC:%; TP: true positive, FN: false negative, FP: false positive, TN: true negative)

Datasets	Methods	Sn	Sp	ACC	MCC	TP	FN	FP	TN
C.elegans	iDNA4mC	79.7	77.5	78.6	0.572	1328	316	349	1205
	4mCPred	82.5	82.6	82.6	0.652	1282	272	270	1284
	4mCPred_SVM	82.4	80.7	81.5	0.631	1280	274	300	1254
	this study	84.9	80.4	82.6	0.653	1319	235	305	1249
D.melanogaster	iDNA4mC	83.3	79.1	81.2	0.625	1474	295	369	1400
	4mCPred	82.4	82.1	82.2	0.646	1458	311	317	1452
	4mCPred_SVM	83.8	82.2	83.0	0.661	1483	286	314	1455
	this study	85.4	83.2	84.3	0.686	1510	259	297	1472
A.thaliana	iDNA4mC	75.7	76.2	76.0	0.519	1498	480	471	1507
	4mCPred	75.5	78.0	76.8	0.536	1494	484	435	1543
	4mCPred_SVM	77.8	79.6	78.7	0.573	1538	440	404	1574
	this study	78.3	80.5	79.4	0.589	1549	429	385	1593
E.coli	iDNA4mC	82.0	77.8	79.9	0.598	318	70	86	302
	4mCPred	81.9	83.2	82.6	0.655	318	70	65	302
	4mCPred_SVM	85.8	80.7	83.3	0.666	333	51	67	321
	this study	86.1	82.5	84.3	0.686	334	54	68	320
G.subterraneus	iDNA4mC	82.2	80.8	81.5	0.630	745	161	174	732
	4mCPred	81.8	83.7	82.8	0.662	742	164	148	758
	4mCPred_SVM	84.0	83.4	83.7	0.674	760	145	150	755
	this study	83.6	85.7	84.7	0.694	757	148	129	776
G.pickeringii	iDNA4mC	82.4	83.8	83.1	0.663	469	100	92	477
	4mCPred	85.0	81.0	83.0	0.668	484	85	108	461
	4mCPred_SVM	86.3	85.8	86.0	0.721	491	78	81	488
	this study	86.3	89.1	87.7	0.754	491	78	62	507

We performed 10-fold cross-validation with the same process as the existing methods. The cross-validation results of the three state-of-the-art predictors were reported in the publication of 4mCPred_SVM (Wei, Luan, Nagai, Su & Zou 2019), where the reported performance of 4mCPred has

been modified by solving the over-estimated problem. The summary of cross-validations is illustrated in Table 3.4. Except for the four evaluation metrics, we also list the sample count of TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative). As shown in the table, in D.melanogaster, A.thaliana and Gpickeringii, our method has the most TP and TN counts, increasing ACC by 0.7% to 1.7% and MCC by 0.015 to 0.033. In G.subterraneus, our method has the highest TN, improving more ACC and MCC by 1% and 0.02% than 4mC_SVM, which has the second-best performance. Additionally, the TP and TN of our method are not the highest in C.elegans and E.coli, but our method slightly improves the ACC and MCC by 1% and 0.02 in E.coli and has a comparative performance with 4mCPred, better than other two methods in C.elegans.

It's clear that our method achieves better overall performance than the existing predictors in independent and cross-validation tests. The improvement of ACC indicates that our method accurately identifies more 4mC sites, and the increase of MCC means that our method has a more balanced performance for classifying positive and negative samples. Therefore, our method is more effective in identifying DNA 4mC sites than the existing predictors.

3.3.4 Case study

Two detailed case studies are conducted to confirm the effectiveness of our method to solve practical problems. C.elegans and D.melanogaster are model organisms widely applied in human disease-related research works, like Parkinson and human aging research investigations (Feany & Bender 2000, Auluck, Chan, Trojanowski, Lee & Bonini 2002, Van Ham, Thijssen, Breitling, Hofstra, Plasterk & Nollen 2008, Feng, Li, Ward, Piggott, Larkspur, Sternberg & Xu 2006). As 4mC plays critical roles in DNA expression and replication in these models, we describe how our method can help identify 4mC sites more accurately in the related genes. We focus on the dlk-1 gene which can promote mRNA stability and local translation

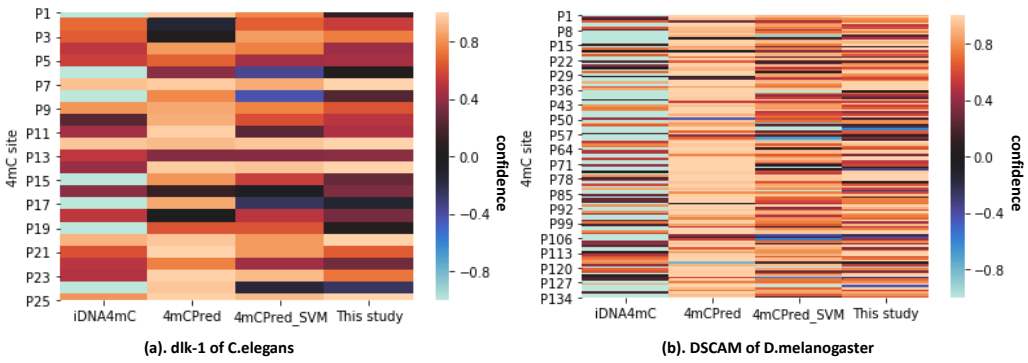

Figure 3.5: **The confidence of predicted label in case studies**

in C.elegans (Yan, Wu, Chisholm & Jin 2009), and on DSCAM gene, which can contribute to the specificity of neuronal connectivity in D.melanogaster (Schmucker, Clemens, Shu, Worby, Xiao, Muda, Dixon & Zipursky 2000).

Table 3.5: **4mC site identificaiton in case studies**(TP: True Postive,; FN: False Negative)

Case	Methods	Total	TP	FN
dlk-1	iDNA4mC	26	19	7
	4mCPred	26	25	1
	4mCPred_SVM	26	20	6
	This study	26	24	2
DSCAM	iDNA4mC	137	70	67
	4mCPred	137	121	16
	4mCPred_SVM	137	122	15
	This study	137	126	11

The 26 and 137 validated 4mC sites in dlk-1 and DSCAM gene are collected from the MethSMRT database. The collected 4mC-contained DNA sequences are all 41-bit, which can be directly submitted into the web tools of three state-of-the-art methods. The prediction result is depicted in Figure 3.5

and Table 3.5. Figure 3.5 shows the label confidence predicted by these four predictors, where the positive confidence refers that the corresponding site is predicted to be 4mC site, and the negative confidence means the site is predicted to be a non-4mC site. As shown in the figure, iDNA4mC achieves the worst performance in both cases, and half of the predictions are incorrect in the DSCAM gene. 4mCPred, 4mCPred_SVM, and the proposed method have similar performance in the DSCAM gene case, while the results made by 4mCPred and our proposed method on the dlk-1 gene are better than 4mCPred_SVM.

More details of the prediction are presented in Table 3.5. Since the testing data in the case study only contains positive samples, there are only TP and FN counts in the results. For the dlk-1 case, 4mCPred has only one wrong prediction, and the proposed method has made two false predictions out of 26 samples, while iDNA4mC and 4mCPred_SVM have 7 and 6 incorrect predictions respectively. For the DSCAM case, there are 137 4mC sites tested, and our proposed method has made 126 correct predictions (i.e., only 11 incorrect predictions). 4mCPred and 4mCPred_SVM have 16 and 15 false predictions, while iDNA4mC has made 67 false predictions. More detailed results can be found at the supplementary Additional file 3.

3.4 Discussion and summary

The 4mC site prediction is a typical sequence site classification problem. The state-of-the-art research work has made some explorations, but their performance still needs improvement. In this chapter, we designed a novel computational method for accurate 4mC site prediction, solving the research question **Q1** (see in Section **1.2**) This method constructs a more effective feature space, integrating five types of sequence features, and use a novel learning algorithm with XGBoost based feature selection scheme. The results show that the feature selection improves the performance, and the prediction model outperforms the other three existing predictors in the

independent tests and the cross-validations. In the future, we will continue to optimize our feature space with novel sequence features of important biological characteristics. Furthermore, we will expand the size of the benchmark datasets to enhance the model's accuracy and generalization ability. Also, since the number of 4mC is much smaller than non-4mC sites in practical situations, the data imbalance will be considered in the next research. At last, we will apply our method to solve other sequence site prediction problems.

Chapter 4

Imbalance learning for the prediction of N^6-methylation sites in mRNAs

4.1 Background

As mentioned in Section 2.2, one critical issue of N^6-Methylation prediction problem is that non-m^6A sites are much more than m^6A sites in the training data. The existing computational methods have overlooked this imbalance issue. They trained the model with balanced datasets containing roughly equal sizes of m^6A samples and randomly selected non-m^6A samples. Such sampling of non-m^6A samples may lead to inadequate learning, and the prediction models would change when the selected non-m^6A samples are different.

Another issue of computational m^6A prediction is the lack of valid features. The state-of-the-art features are usually derived from window sequences with m^6A at the centre position. These features include binary encoding sequence features (Xiang et al. 2016, Zhou et al. 2016), k-mers (Xiang et al. 2016), physical-chemical properties (Liu, Xiao, Yu, Jia, Qiu & Chou 2016, Zhang, Sun, Liu, Ren, Shen & Yu 2016), position-specific

nucleotide propensities (Li, Liu, Shen & Yu 2016), pseudo nucleotide compositions (Chen, Xing & Zou 2017, Wan, Duan & Zou 2017, Zou, Wan, Ju, Tang & Zeng 2016), nucleotide pair spectrums (Zhou et al. 2016) and multi-internal nucleotide pair positions (Xing et al. 2017).

Here we use a cost-sensitive XGboost classifier to address the imbalance issue. Similar to previous works, m^6A samples and non-m^6A samples are labeled as positive and negative, respectively. The classifier is then trained with all the samples without selecting a subset of negative samples and prevents over-fitting by defining different costs for the incorrect classified positive and negative samples. The model minimizes the cost function in the learning stage and improves the precision of classifying positive samples. Besides, ROC rather than accuracy is set as the training cost function. Owing to training on the whole dataset without sampling noise, our method HMpre exhibits higher performance and better robustness.

To improve the effectiveness of feature space, we present three types of novel m^6A features. First, we extract novel features to capture specific single nucleotide polymorphism (SNP) variants in the window sequences through the MRMR method and Fisher's exact test (Peng, Long & Ding 2005). These features are relevant because single nucleotide variants can affect m^6A dynamics (Zheng, Nie, Peng, He, Liu, Xie, Miao, Zuo & Ren 2017). Moreover, m^6A occurs richly in some particular regions of transcripts. Thus we calculate the absolute and relative locations of m^6A sites as new features. To further exploit the distribution properties of nucleotides, entropy information is also considered as new features. Together with these newly proposed features, conventional features including 4-bit binary, overlapping chemical property with density, and k-mers are integrated into our feature space to describe comprehensive characteristics of methylation.

In the performance evaluation of our method HMpre, we first report specific SNP positions as new features. Then we report a detailed comparison result with three existing balance learning predictors on an independent test dataset. HMpre achieves a much better performance of precision 0.3035, F1

0.3961, and MCC 0.3329. Since the ratio of positive sites over negative sites in a test mRNA is unknown, HMpre and existing predictors are also evaluated on nine datasets containing different ratios of positive sites over negative sites. Results show that HMpre works better and has stronger robustness on the ratio change. In practical use, the inputs to a predictor are always individual transcripts. Therefore the four methods are then applied to make predictions on single transcripts. Again, HMpre achieves the best overall performance. Furthermore, we evaluate the effectiveness of the features with 10-fold cross-validation and feature importance scores from the XGBoost classifier. The new features are all meaningful, and the proposed feature space improves performances notably. In the case studies, the c-Jun gene's transcript is taken as an example to demonstrate the prediction details. Then we evaluate our method on the transcript of the CBFB gene relating to HIV-1 infection, and our method also achieves better results than the other predictors.

4.2 Materials and methods

Datasets

Currently validated human mRNA m^6A sites were all obtained by Ke and Linda from single nucleotide resolution maps (Ke et al. 2015, Linder et al. 2015). To guarantee the reliability of negative samples, non-m^6A sites conforming to the conserved motif DRACH were all produced from these validated transcripts. Based on these datasets, Zhou has built a human mature mRNA m^6A dataset, which is the largest human m^6A dataset so far. The dataset used in our experiments is downloaded from Zhou's work (Zhou et al. 2016). After removing redundant and unaligned samples, we get 7506 mature human transcripts in total. We reserved 6280 transcripts for training and 1226 transcripts for independent testing. For each transcript, the number of non-m^6A conforming to the DRACH motif is much larger than m^6A sites. The training dataset contains 26512 positive samples and 271214 negative samples, while the independent test dataset contains 5644 positive

samples and 54744 negative samples. Each sample contains the transcript id, the location of the target adenine, and the flanking window sequence. All samples used in our dataset are listed in Additional File 4.

4.2.1 Feature space construction

Computational prediction methods usually build features from a flanking window sequence with m^6A at the center position. The size of the flanking window varies from 20 to 50 nts in previous works, and we choose the size of 25-nt, which is similar to other human predictors. Thus the features are extracted from the 51-nt long sequence. Based on the sequence characteristics of the m^6A site, we introduce three types of new features: site location-related features, features related to entropy information, and SNP features. Three types of conventional features are also used. There are a total of 509 dimensions in our feature space. The transcript sequences, length information (including coding region and UTRs), and SNP variants are obtained from the Ensembl online human gene database (GRCh38.p10). A diagram of the feature space construction is presented in Figure 4.1.

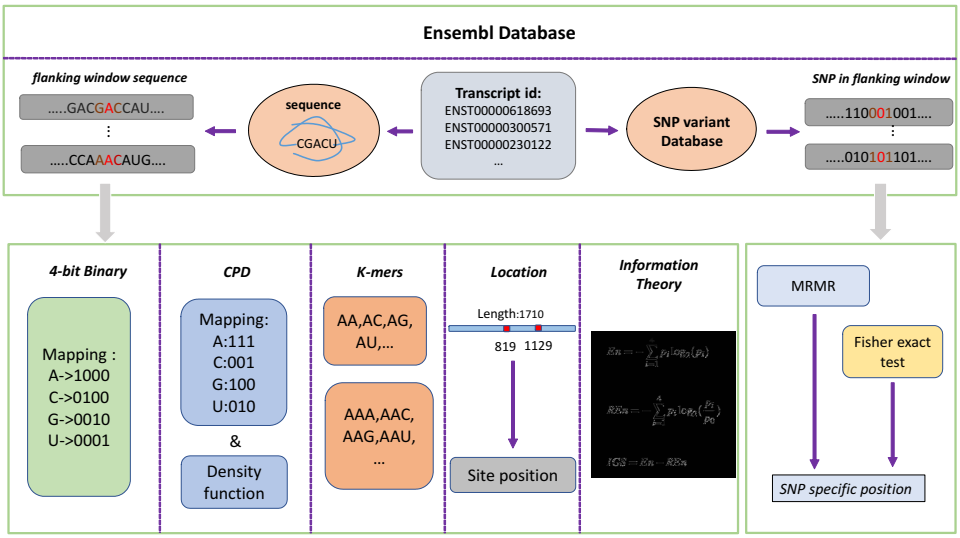

Figure 4.1: **Feature space construction**

Three New Types of Sequence Features

Site location Related Features In mature transcripts, m^6A sites are rich in some special regions, such as the 3' UTRs near the stop codon (Meyer et al. 2012). However, non-m^6A sites conforming to the DRACH motif are randomly distributed over the entire transcript. Thus the location of the target adenine site in the transcript can be taken as a new feature. Specifically, site location refers to the distance between the target site and the transcript start site. The relative location of the target site in the whole transcript is also taken as a new feature, which is the ratio of the site location over the transcript length.

Features Related to Entropy Information Because of motif conservation for regulating protein binding sites, the nucleotides around m^6A sites have some unique distributions. Shannon information theory can be used to evaluate these nucleotide distributions in the transcript fragment sequences. We calculate Shannon entropy (En), relative entropy (REn) and information gain score (IGS) of all samples as a new type of feature. The scores of these features are calculated as:

$$En(s) = -\sum_{i\in\{A,G,U,C\}} p_i^s \log_2(p_i^s) \tag{4.1}$$

$$REn(s) = -\sum_{i\in\{A,G,U,C\}} p_i^s \log_2(\frac{p_i^s}{p_0}) \tag{4.2}$$

$$IGS(s) = En(s) - REn(s) \tag{4.3}$$

where p_i^s is the frequency of A, G, U, C in sequence s, and p_0 is the uniform distribution of each nucleotide occurrence, namely $p_0 = 1/4$. The frequency of each nucleotide is then combined with the entropy features as a 7-dimension feature vector.

SNP Features Singe nucleotide polymorphism is a kind of variant at specific sites in the genome. For SNP sites, several possible nucleotide variations are alleles for this position. As a synonymous single nucleotide

variant, SNP changes the sequence of mRNA but does not alter the amino acid sequence of protein (Sauna & Kimchi-Sarfaty 2011). Also, m^6A is regulated by some proteins which also have fixed RNA binding sites, which means the flanking window sequence around m^6A site has specific base groups patterns. The SNP variant of mRNA sequence may disrupt the DRACH motif or protein binding regions, leading to failures of m^6A dynamic regulations (Zheng et al. 2017). Hence, we attempted to find positions with unique SNP states. From the Ensembl database, we map SNP variants in the transcript and convert sample sequence into a 51-bit 0/1 vector (i.e., 0 denotes a non-SNP variant position; 1 donates an SNP variant position). As there are various methods to select effective features (Zou, Zeng, Cao & Ji 2016, Saeys, Inza & Larrañaga 2007), in this paper Max-Relevance Min-Redundancy (MRMR) algorithm (Peng et al. 2005) and Fisher's exact test are adopted to recognize special SNP positions.

MRMR selects positions with a maximal statistical criterion based on mutual information. MRMR tries to find a position subset, which has maximum relevance (dependency) with class and minimum internal redundancy. MRMR adds positions into the subset one by one, and the order is determined by relevance to the target class and the redundancy with the other positions. Fisher's exact test is a statistical significance test. For an individual position, it investigates the SNP variant distribution difference between the positive and negative samples and derives a p-value from assessing the difference. A low p-value means the SNP variant at this position has a great difference between the negative and positive samples. Finally, we can rank positions with Fisher's exact test p-value and the MRMR selection order. By calculating the average ranking of MRMR and Fisher's exact test, positions with a significant SNP specificity can be identified. The SNP variant states of such specific positions are considered as SNP features. The detailed SNP specificity identification algorithm is presented in Algorithm S1 of Additional file 5.

Conventional Sequence Features

4-bit Binary Features Binary encoding is a common feature extraction method to characterize RNA sequences. As the mRNA sequence contains four nucleotides A, C, G, and U, this encoding method can map every single nucleotide into a 4-bit binary code. The mapping rules are: 'A'- (1,0,0,0), 'C'- (0,1,0,0), 'G'- (0,0,1,0), 'U'- (0,0,0,1). In this way, a 51-nt sequence can be transformed into a 204-dimension feature vector.

Chemical Property with Density (CPD) Based on differences in chemical property, four kinds of nucleotides can be categorized into different groups (Chen, Tang, Ye, Lin & Chou 2016). In terms of ring numbers in a single base group, C and U have only one ring while A and G have two. Besides, C and G have strong hydrogen bonds when forming secondary structures, whereas hydrogen bonds in A and U are both weak. When considering chemical functionality, amino group contains A and C while keto group includes G and U. Thus, we can divide the nucleotides by different chemical properties and use overlapping encoding rules: 'A': (1,1,1), 'C': (0,0,1), 'G': (1,0,0), 'U': (0,1,0). In literature work, the density of nucleotide is always used with chemical property features, which calculates the frequency of a nucleotide occurring before current position. Density feature d_i is defined as

$$d_i = \frac{1}{|S_i|} \sum_{j=1}^{i} f(s_j), f(s_j) = \begin{cases} 1 & s_j = s_i \\ 0 & s_j \neq s_i \end{cases} \tag{4.4}$$

K-mer Features In the mRNA sequence, adjacent nucleotide pairs have an influence on mRNA structures and functions. K-mer is the frequency of k-nt adjacent nucleotides. As a global feature, k-mer has been proven to be effective in many sequence-based site predictions. The length of k-mer feature is 4^k bits. In this paper, we adopt 2-mer and 3-mer. Each sample has an 80-dimension k-mer feature vector.

4.2.2 Imbalance learning

Imbalance learning has been explored for protein binding site prediction (Yu, Hu, Huang, Shen, Qi, Tang & Yang 2013, Hu, He, Yu, Yang, Yang & Shen 2014, Song, Li, Zeng, Wu, Guo & Zou 2014) and protein-protein interaction sites identification (Wei, Han, Yang, Shen & Yu 2016, Liu, Shen & Yu 2016). However, imbalance learning for m^6A prediction has not been explored. An intuitive way to address this problem is to integrate sampling and ensemble techniques, which trains basic classifiers with different sampling data and combines the results in an ensemble way to reduce the random sampling bias. But it requires effective sampling techniques to select meaningful negative subsets, and there are some researches focuses on dynamic and cluster ways (Lin, Chen, Qiu, Wu, Krishnan & Zou 2014). Another viable strategy is to introduce cost-sensitive learning models, like weighted support vector machine and cost-sensitive decision trees, using different matrices to describe the costs for classifying samples into the wrong class (He & Garcia 2009).

Here we use a cost-sensitive XGBoost classifier as a learning model. XGBoost (eXtreme Gradient Boosting) is a tree boosting algorithm developed by Chen (Chen & Guestrin 2016). It is an advanced implementation of the gradient boosting algorithm, which has been widely applied for classification problems. XGBoost has some advantages over other cost-sensitive classifiers. Firstly, the regularization can effectively prevent the training model from over-fitting. Secondly, embedded parallel processing allows a faster learning speed. Thirdly XGBoost is of high flexibility and allows users to define custom optimization objectives and evaluation criteria. Moreover, the XGBoost classifier can learn from imbalance training data by setting class weight and taking ROC as evaluation criteria. Here we implement the model with a python package named xgboost (vision 0.6a2). The parameters can be optimized by 10-fold cross-validation in the learning stage. The parameters in our model are: 'lambda': 700, 'max-depth': 6, 'eta': 0.1, 'silent': 1, 'objective': 'binary:logistic', 'booster': 'gbtree', 'scale-pos-weight': 6, 'eval-

metric': 'auc' and training boost round is 400, while other parameters are all default values.

In this paper, our method is compared with three recently published human m^6A prediction methods. These three literature methods are: SRAMP (Zhou et al. 2016), Methy-RNA (Chen, Tang & Lin 2017) and RAM-NPPS (Xing et al. 2017). They all have open access web predictors, and SRAMP also provides a tool package for local implementation. The prediction results of Methy-RNA and RAM-NPPS are obtained from the web predictors, while the results of SRAMP are derived from tool package in mature mode.

4.2.3 Performance evaluation metrics

The proposed prediction method is evaluated by 10-fold cross-validations and independent test dataset with four frequently used metrics: precision, recall, F1-score, and Matthews correlation coefficient (MCC). As RAM-NPPS and Methy-RNA cannot return prediction probabilities, we do not use AUROC or AUPRC as evaluation metrics.

Precision and recall reflect the tendencies of classifier prediction. Recall (also called sensitivity in binary classification) illustrates how many positive samples are rightly classed, and precision shows the ratio of true positive sample ratio in all predicted positive-label samples. There is always a trade-off between precision and recall, so we introduce F1 and MCC to evaluate the overall performance of a predictor. F1-score combining precision and recall together can assess the performance on both balanced and unbalanced test datasets. MCC is also a frequently used metric in classifier evaluation, which returns a value between -1 to 1: 1 standing for perfect prediction and -1 for reversed prediction.

Table 4.1: **Ranking details of top 12 specific SNP positions** (FET: Fisher's exact test)

No.	Position	FET	MRMR	Average	Ranking
1	-2	1	1	1	1
2	-1	2	5	3.5	2
3	-24	6	7	6.5	3
4	-21	10	4	7	4
5	-19	7	12	9.5	5
6	2	3	23	13	6
7	-25	4	24	14	7
8	-11	19	9	14	7
9	-4	8	21	14.5	8
10	-15	21	11	16	9
11	-9	15	17	16	9
12	-23	9	25	17	10

4.3 Results

We report the specificity results of SNP identification as new features. In the performance comparison and evaluation, we tested our HMpre method and other existing predictors on the independent test dataset. To demonstrate the robustness of our method to deal with the unknown percentages of positive samples in real transcripts, we compared our method with three existing human m^6A predictors on datasets of different positive-and-negative sample ratios. To evaluate the performance for practical use, we tested all the predictors on single transcripts. Lastly, we report the feature effectiveness results of HMpre and XGBoost classifier feature importance scores.

4.3.1 Specific SNP status as new features

MRMR and Fisher's exact test is applied to analyze sequence SNP variant states in the training dataset and identify positions with specific SNP variant

states as a new feature. As presented in Figure 4.2, MRMR and Fisher's exact test give rankings to all the positions numbered from -25 to 25 in the window sequence.

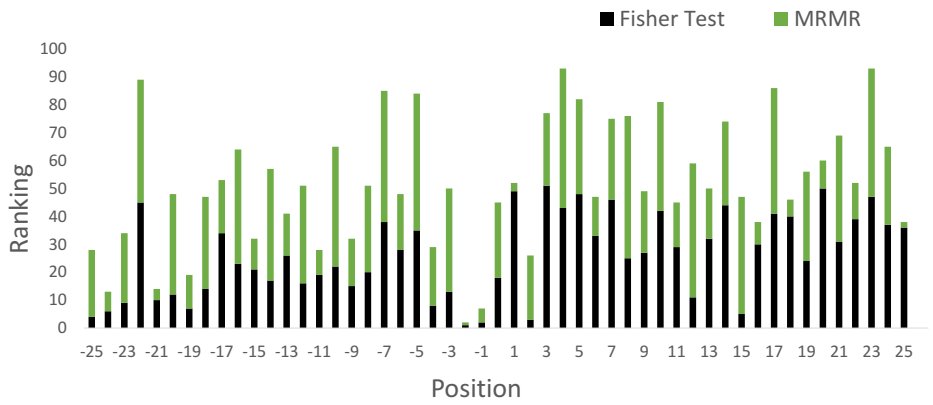

Figure 4.2: **SNP specificity ranking** The black blocks stand for the Fisher's exact test rankings and the green blocks stand for the MRMR rankings. X-axis is the window sequence sites from -25 to 25. Y-axis is the total ranking of each position. A low ranking means a high SNP specificity at this position.

In the process of selecting a position subset, MRMR defines mutual information to evaluate the subset for the inner redundancy and relevance with the target class, then it gives out the order of position selection, and we take the order as position importance ranking. The top 12 positions are -2, 25, 1, -21, -1, 18, -24, 16, -11, 20, -15 and -19. Fisher's exact test can statistically recognize the SNP variant distribution difference for these individual positions between the positive and negative samples, as described by a p-value. With Fisher's exact test p-values (details in Table S1 of Additional file 5), we can also rank all these positions. The top 12 positions are -2, -1, 2, -25, 15, -24, -19, -4, -23, -21, 12 and -20. Finally, we choose the top 12 positions with the highest average ranking as SNP features. These highly ranked positions are illustrated in Table 4.1. These positions have relatively higher ranking both in MRMR and Fisher's exact test. Detailed

results are listed in Table S2 of Additional file 5.

4.3.2 Performance on the independent dataset

Table 4.2: **Performance on the independent test dataset** (Methy: Methy-RNA; NPPS: RAM-NPPS)

Methods	Precision	Recall	F1	MCC
Methy	0.065	0.5184	0.1163	-0.1619
NPPS	0.1656	**0.6339**	0.2626	0.1833
SRAMP	0.2638	0.4812	0.3408	0.2653
HMpre	**0.3035**	0.5698	**0.3961**	**0.3329**

Our proposed HMpre is compared with three existing prediction methods on the independent test dataset. The prediction results are reported in Table 4.2. HMpre achieves the best performance under all metrics except recall; RAM-NPPS has a better recall of 0.6339 than HMpre. The precision of HMpre is 0.3035, 0.04 higher than SRAMP, which is the best in the existing predictors. Overall, HMpre achieves an F1 score of 0.3961, higher than the best F1 value of the other three predictors (0.3408 by RAM-NPPS). In terms of MCC, Methy-RNA has a value of -0.1619, and SRAMP is 0.2653, about 0.08 higher than RAM-NPPS, but still lower than HMpre's 0.3329.

4.3.3 Robust performance when tested on datasets with different imbalance ratios

In normal situations, the numbers of m^6A and non-m^6A sites are unknown before prediction. Therefore, a practical m^6A predictor should have strong robustness against the imbalance level change. To appraise the robustness of HMpre and other predictors, we test them on nine datasets whose negative samples to positive samples ratios range from 1:1 to 9:1. Here we adopt the overall metrics F1 and MCC as evaluation criteria. The results are reported

in Figure 4.3. The F1 and MCC values of all the methods have a trend of decreasing when the imbalance level increases. The F1 scores of RAM-NPPS and Methy-RNA decrease more rapidly than HMpre and SRAMP. For the MCC values, HMpre also has a relatively slow-changing rate while the other methods are comparable. Moreover, HMpre has a better performance on all of these datasets under F1 and MCC, proving that HMpre has stronger robustness.

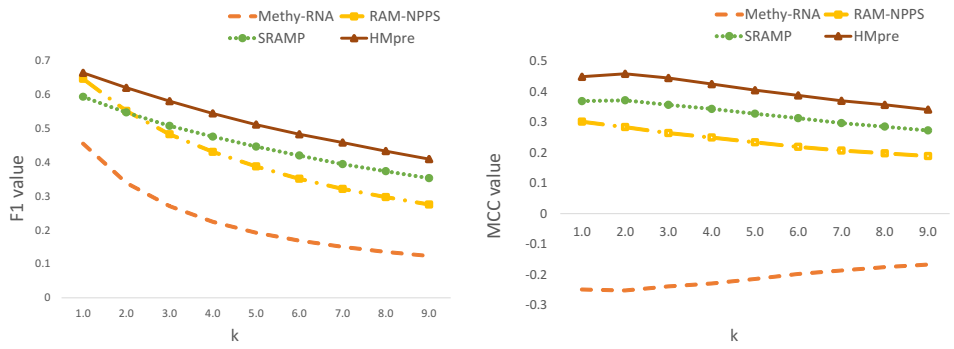

Figure 4.3: **Performance on datasets of different imbalance levels** The F1 and MCC values of four predictors are represented. X-axis k is the ratio of the negative samples to positive samples (imbalance level) in a test dataset; Y-axis is metric value.

4.3.4 Performance on 1226 individual transcripts

Since the testing objects are always single transcripts in real cases, the four predictors are evaluated on individual transcripts. There are 1226 transcripts in the independent dataset for the four methods to make predictions. The imbalance levels of the 1226 transcripts are different, and we calculate the average metric values of all the transcripts as the final results for each method. The results are reported in Table 4.3. Although RAM-NPPS has the highest recall of 0.6582, HMpre achieves the best performance under the remaining four metrics (precision 0.2972, recall 0.6062, F1 0.3658, and MCC

0.3239). Especially, the overall metrics F1 and MCC of HMpre are about 0.07 and 0.08 higher than SRAMP, the best existing predictor.

Table 4.3: **Average performance on individual 1226 transcripts** (Methy: Methy-RNA; NPPS: RAM-NPPS)

Methods	Precision	Recall	F1	MCC
Methy	0.0723	0.5075	0.1174	-0.1614
NPPS	0.1770	**0.6582**	0.2529	0.1907
SRAMP	0.2484	0.4759	0.2928	0.2387
HMpre	**0.2972**	0.6062	**0.3658**	**0.3239**

4.3.5 Feature importance analysis

Three types of new features are extracted to add to the existing feature space to improve the prediction performance. 10-fold cross-validations with different feature spaces are used to verify whether the new feature space actually improves the prediction performance. The performance of the three types of traditional features and their merged features are compared with the proposed feature space in Table 4.4. The three types of traditional features (four-bits binary coding, chemistry property with density, and k-mers) achieve distinct performance, and the 4-bit binary features are better than the other two types of features. By joining the three types of conventional features together, all metrics increase comparing with individual features. The proposed feature space, combining conventional and new features together, exhibits the best performance under all metrics.

We also attempted to understand more about the role of each feature in prediction. XGBoost can make an inner analysis of feature importance during the learning process and output scores for all the features. The importance scores can reveal how meaningful the features are when building a model and tell which features play leading roles in the feature space.

The feature importance scores boxplot is presented in Figure 4.4. There

Table 4.4: **Different feature space performance in cross validation** (CPD: Chemical Property with Density; Joint: joint of conventional features)

Feature	Precision	Recall	F1	MCC
K-mers	0.1392	0.3426	0.2461	0.1572
CPD	0.0.2460	0.4816	0.3256	0.2532
Binary	0.25	0.4906	0.3312	0.2601
Joint	0.2519	0.5035	0.3358	0.2661
Proposed	**0.2669**	**0.5248**	**0.3538**	**0.2877**

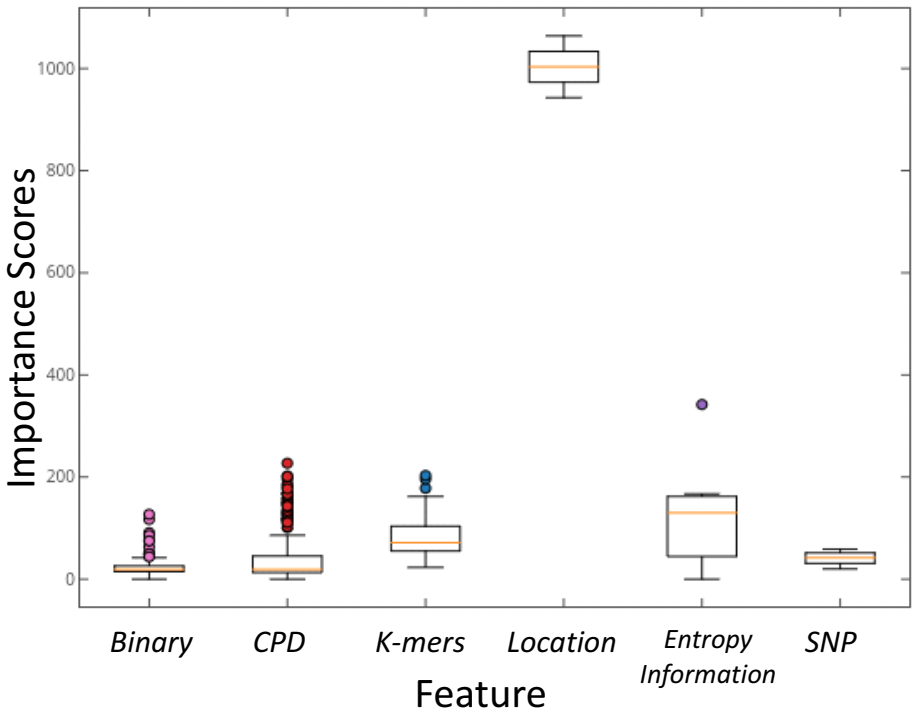

Figure 4.4: **Boxplot of feature importance scores**

are 509 features, and their distribution is presented in Table S3 of Additional file 5. The importance scores have a wide range from 0 to 1064. The features with a 0 score are from 4-bit binary and CPD features, corresponding to the motif adjacent sites, which are 'GAC' or 'AAC' in all the samples. The dimension with the highest score 1064 (f501) refers to the site distance

from the transcript start site, followed by features of relative location in the transcript (f500, scored 943) and sequence entropy (f506, scored 342). Besides, density features in CPD features have relatively high importance scores. Detailed importance scores are shown in Figure S1. For the average score, binary and CPD are much lower than other features, while site location and entropy information are obviously higher. K-mers and SNP have comparable average scores. From the results, the three types of new features are indeed significant in the feature space.

4.4 Case studies

In this section, we report two detailed case studies to understand the difference between the four predictors and evaluate their capacity in practical use. First, we describe the prediction results for the c-Jun transcript from the test dataset. The second case study is about the m^6A sites in the mRNAs of the CBFB gene, which can modulate HIV-1 replication and infection (Lichinchi, Gao, Saletore, Gonzalez, Bansal, Wang, Mason & Rana 2016).

4.4.1 m^6A site prediction for c-Jun transcript

Transcript ENST00000371222 of the c-Jun gene contains 25 verified m^6A sites and 47 non-m^6A sites conforming to the DRACH motif. HMpre predicted 21 m^6A sites, 18 of which are true positives, and three are false positives, while SRAMP predicted 12 true positive m^6A sites and three false positives. RAM-NPPS made 14 true positives and 12 false positives. Methy-RNA made the most 31 false positive predictions and identified only 19 true m^6A sites. Thus, Methy-RNA achieved the highest true positive rate, but it made the most number of false-positive predictions. See Figure 4.5. Although SRAMP achieved a good precision of predicted m^6A sites, a large number of true m^6A sites were wrongly classified. RAM-NPPS has more false positives and less true positive predictions than SRAMP and HMpre.

Table 4.5 shows detailed prediction performance. Overall, the precision,

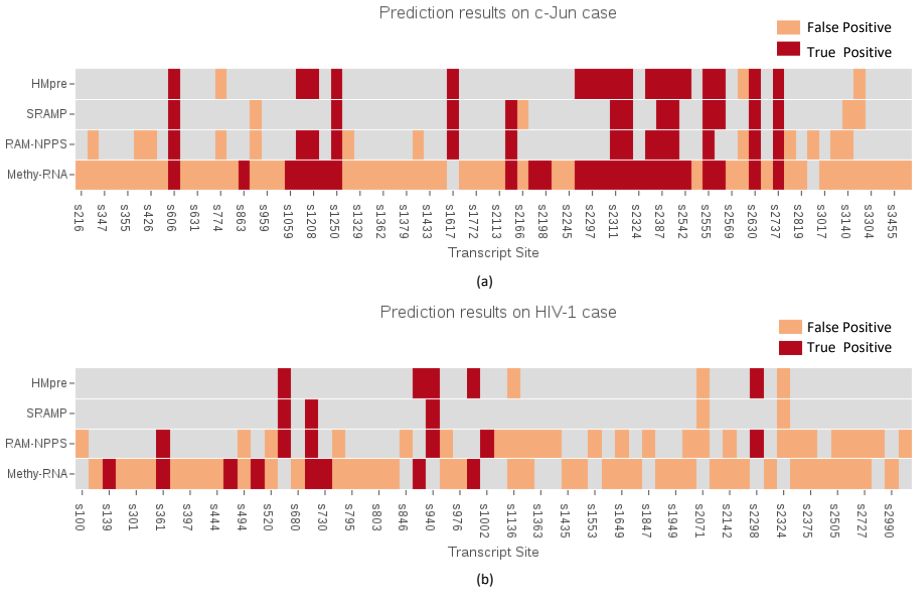

Figure 4.5: **Predicted m^6A sites in the case studies** The x-axis stands for the potential m^6A sites confirming to the sequence motif DRACH and the y-axis indicates the four predictors. All colored blocks are the predicted m^6A sites. Red blocks represent true positive sites, and yellow blocks are false positive ones. (a) the prediction results for the c-Jun case and (b) the predictions for the HIV-1 case.

F1 and MCC of our HMpre method are much higher than the other prediction methods. Although Methy-RNA has a high recall of 0.96, it has the lowest precision, F1 and MCC. The performance of SRAMP is better than RAM-NPPS, but the recall of SRAMP is the lowest 0.48, suggesting a lot of positive samples are predicted to be negative.

4.4.2 m^6A site prediction for a transcript related to HIV-1 infection

The longest transcript ENST00000290858 of the CBFB gene from the Ensembl database was chosen for this case study. There are 62 adenines

Table 4.5: **Results for the c-Jun gene case study** (Methy: Methy-RNA; NPPS: RAM-NPPS)

Case	Methods	Precision	Recall	F1	MCC
c-JUN	Methy	0.3428	**0.96**	0.5052	-0.0542
	NPPS	0.5384	0.56	0.549	0.3019
	SRAMP	0.75	0.48	0.5853	0.4522
	HMpre	**0.8571**	0.72	**0.7826**	**0.6872**
HIV-1	Methy	0.1702	**0.6666**	0.2711	-0.1045
	NPPS	0.1935	0.5	0.279	0
	SRAMP	0.6	0.25	0.3529	0.2727
	HMpre	**0.625**6	0.4166	**0.5**	**0.4203**

(A) conforming to the motif in this transcript. The experimentally validated m^6A sites of the CBFB gene are acquired from RMBase, an online m^6A database (Sun, Li, Liu, Wu, Zhou, Qu & Yang 2015). Based on these data, we constructed a test dataset of 12 positive samples and 50 negative samples.

The predicted m^6A sites are presented in Figure 4.5. HMpre made five true positives and three false-positive predictions, while SRAMP made three true positives and two false-positive predictions. RAM-NPPS and Methy-RNA made more false positives than true positives: RAM-NPPS had six true positives and 16 false positives, and Methy-RNA had 51 false positives and eight true positives. The predicted m^6A sites by SRAMP are mainly correct, but it missed a lot of true m^6A sites.

The detailed results are reported in Table 4.5. Methy-RNA achieves the best recall 0.6666 but the worst precision 0.1702. SRAMP has a high precision 0.7692, but the lowest recall 0.25. Our HMpre method has the best precision 0.56256 and achieves the best performance on the overall metrics F1 0.5 and MCC 0.4203.

4.5 Discussion

In this paper, we adopted an XGBoost classifier as the prediction model. On the one hand, this classifier can learn from imbalanced data, which is similar to data in practical prediction situations, and inner regularization rules can prevent the model from over-fitting. On the other hand, when the scale of training data is quite large, it would cost classifiers like SVM and Random forest much longer time than our method in the training stage.

The efficiency of features is crucial to the performance of predictors. Here, we presented m6A sites with meaningful biological features instead of just using flank window sequence features. In this work, the size of the flanking window is 51-nts, which is the same with existing methods. The influence of sequence size on feature efficiency will be studied in the next stage of research. In addition, some m6A biological characteristics found recently can be taken as new features in the prediction, and we will try them in the future.

4.6 Summary

This chapter has proposed a novel computational method called HMpre to address the research question **Q2** (see **Section 1.2**) of human mRNA m^6A prediction. The key idea is a cost-sensitive learning model. Three types of new features are also introduced to learn more from the imbalanced training data for the further improvement of the prediction performance. Along with the other three existing methods, HMpre was tested on an independent dataset. The results show that our method has better correctness and robustness. The feature importance analysis demonstrates that the new features are exactly meaningful in the prediction. In the detailed case studies, our method also outperforms over the existing predictors. Class imbalance is a long-neglected but important issue in the m^6A prediction problem. Imbalance learning provides a promising way to resolve this issue.

Chapter 5

Identification of lung cancer gene markers through kernel maximum mean discrepancy and information entropy

5.1 Background

As we pointed in Section 2.3, in cancer studies, the histologically normal tissue adjacent to the tumor is usually used to compare with the tumor tissue under the assumption that they are the same with real healthy tissues. This approach allows researchers to compare samples from the same patient and reduce the individual-specific effects. However, recent studies have deepened our understanding of NAT tissue, indicating that NAT is not exactly equal to the real healthy tissue (Aran et al. 2017). In NAT tissues, the specific micro-environment surrounding tumor makes the change of gene expression in various pathways that are related to disease development. In order to identify efficient and meaningful marker genes, we proposed to detect differentially expressed genes (DEGs) from real normal, NAT, and tumor tissues.

Here, we present a novel approach to identify genes markers for lung

cancer with kernel maximum mean discrepancy (MMD) and Information Entropy. As mentioned above, the conventional DEA methods can calculate a p-value to evaluate the expression difference based on certain statistical hypothesis, but it's hard to decide which distribution assumption is correct before calculation. Inspired by the distribution measure method of transfer learning, we use the kernel MMD to detect DEGs between tumor, NAT, and normal tissues. This method can output the maximum mean discrepancy score, which indicates the degree of differential expression without requiring a statistical hypothesis on data distribution. Besides, although the p-value of conventional techniques can identify DEGs, it is essential to define a threshold of expression level to distinguish different types of tissue. Commonly, Researchers would like to take the upper boundary of lower expressed tissue or lower edge of higher expressed tissue as the threshold when there is a distinct expression gap. But this kind of gap is not always existing, and then the threshold is hard to define. As the gene expression level is continuous data and how to choose a definite threshold point is a tough task. Here we applied the information theory to solve this problem.

In this paper, we first evaluate the expression level difference of 23368 genes in normal, normal adjacent tumor and tumor tissues with the kernel maximum mean discrepancy. Then the top-ranked genes selected by the kernel MMD method are compared with genes selected by two conventional DEA methods, t-test and fold change. Then GO and KEGG pathway enrichment analysis are conducted to analyze the top 100 genes ranked by average MMD scores. Lastly, the top 10 genes are selected as marker genes for lung cancer, and their expression boundaries between normal, NAT, and tumor tissues are identified by the proposed information theory method.

5.2 Materials and methods

5.2.1 Dataset

Three gene expression datasets used in this paper are collected from different tissue types in reference (Aran et al. 2017), containing the expression data of 23368 genes. Dataset 1 includes the gene expression data of 373 normal healthy samples. The raw reads file of dataset 1 is obtained from the GTEx program (phs000424.v6.p1, 18 November 2015). Dataset 2 has 59 NAT tissues, while dataset 3 has 541 lung cancer tumor tissues. Their raw feature counts and FPKM values are original from NCBI Gene Expression Omnibus (GEO) (Barrett, Wilhite, Ledoux, Evangelista, Kim, Tomashevsky, Marshall, Phillippy, Sherman, Holko et al. 2012). Since the raw values are from different data sources, the RNA-sequencing raw reads files were processed and normalized with the Rsubread package and aligned to the UCSC hg19 reference genome with the same pipeline. The processed GTEx expression profiles of dataset 1 are available in GEO under an accession number GSE86354, and the other two datasets are deposited as GSE62944.

5.2.2 Gene marker identification framework

With the above three datasets, we apply a novel approach to detect DEGs and determine the expression boundaries between normal, NAT, and tumor cells as the criterion of the lung cancer diagnosis.

In our method, there are mainly four steps: First, the kernel Maximum Mean Discrepancy is used to identify DEGs between two types of tissues respectively, and genes are ranked by the MMD values; Second, the genes with top average MMD rankings are selected from all genes; Third, genes selected from the previous step are put into KEGG pathway analysis and GO enrichment analysis to validate the efficiency of those gene markers; Last, we define the gene expression boundaries for the top 10 marker genes with information gain theory. The whole framework of the proposed approach is

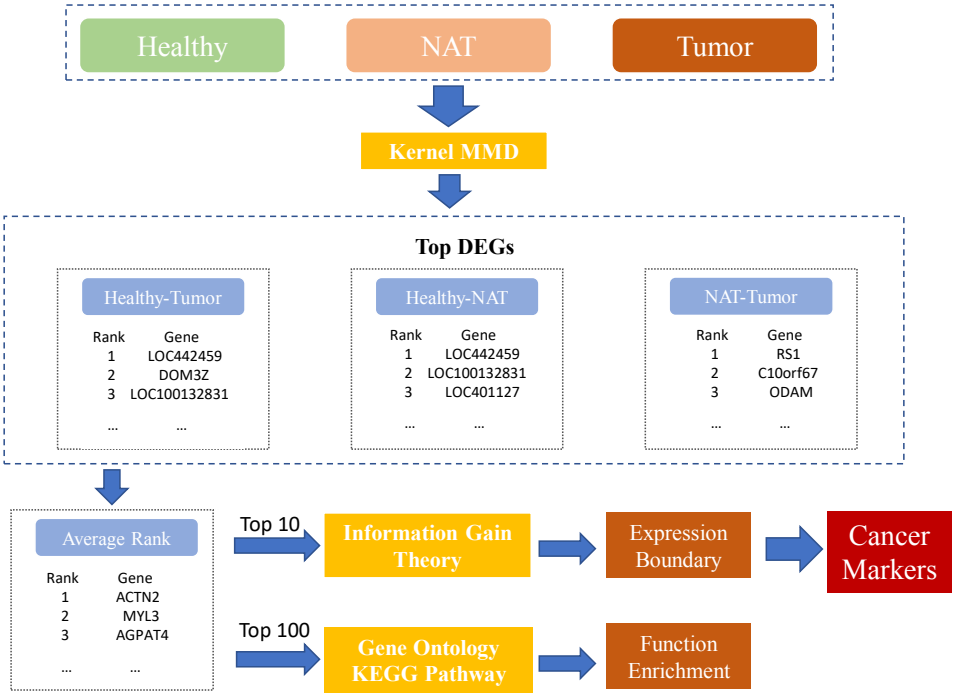

Figure 5.1: **Gene marker identification framework**

illustrated in Figure 5.1.

5.2.3 Kernel maximum mean discrepancy

The problem of comparing the probability distribution between two sample groups, also referred to as the two-sample problem, widely exists in data science areas. In the bioinformatics field, this problem is extensively existing in micro-array data analysis, database attribute matching, data integration from different platforms, and so on. The two-sample problem's key point is to determine if two groups of observations are from the same distribution. Some statistical test methods were applied to address that in previous researches.

However, these methods have different statistical modelings based on specific assumptions of data distribution, which is commonly unknown before

calculation in practical use. In some previous studies, researchers have explored using the kernel Maximum Mean Discrepancy (MMD) method to test the distribution difference in RNA-Transcript expression and pathway differential expression and achieved better performance than traditional statistical tests (Stegle, Drewe, Bohnert, Borgwardt & Rätsch 2010, Vegas, Oller & Reverter 2016). Here, we adopt kernel MMD to identify the DEGs in gene expression data for lung cancer.

Give F to be a class of functions $f : \chi \to R$. Two samples $X = \{x_1, x_2, ...x_m\}$ and $Y = \{y_1, y_2, ...y_n\}$ are drawn form two probability distribution p and q, respectively. The empirical estimation of MMD value is as following (Gretton, Borgwardt, Rasch, Schölkopf & Smola 2007):

$$MMD[F, p, q] := \sup_{f \in F}(E_p[f(x)] - E_q[f(y)]) \tag{5.1}$$

$$MMD[F, p, q] := \sup_{f \in F}(\frac{1}{m}\sum_{i=1}^{m} F(x_i) - \frac{1}{n}\sum_{i=1}^{n} F(y_i)) \tag{5.2}$$

As the definition above, if the function class F is rich enough, the value of MMD will be zero if and only if p=q. But a too rich F will lead to that MMD differs from zero for most finite sample estimates. Thus some restrictions ought to be placed on the function class. One trade-off way is to set F as the unit ball in a universal reproducing kernel Hilbert space H, defined on the compact metric space χ. Since H is a complete inner product space of functions f: $\chi \to R$, the function mapping $f \to f(x)$ can be expressed as an inner product via $f(x) = \langle f, \phi(x)\rangle_H$,where $\phi : \chi \to H$ is the feature space map from x to H. Then MMD can be rewritten as:

$$\begin{aligned} MMD[F, p, q] &= \sup_{\|f\|_H \leq 1} E_p[f(x)] - E_q[f(y)] \\ &= \sup_{\|f\|_H \leq 1} E_p[\langle f, \phi(x)\rangle_H] - E_q[\langle f, \phi(y)\rangle_H] \\ &= \sup_{\|f\|_H \leq 1} \langle \mu_p - \mu_q, f\rangle_H \\ &= \|\mu_p - \mu_q\|_H \end{aligned} \tag{5.3}$$

Then we can calculate like the following function:

$$\begin{aligned} MMD^2 &= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_H \\ &= \langle \mu_p, \mu_p \rangle_H + \langle \mu_q, \mu_q \rangle_H - 2 \langle \mu_p, \mu_q \rangle_H \\ &= E_p \left\langle \phi(x), \phi(x^{'}) \right\rangle_H + E_p \left\langle \phi(y), \phi(y^{'}) \right\rangle_H \\ &- 2E_{p,q} \langle \phi(x), \phi(y) \rangle_H \end{aligned} \tag{5.4}$$

As the inner product can be replaced by Gaussian kernel $k(x, x^{'}) = exp(-\left\|x - x^{'}\right\|^2 / (2\sigma^2))$,the value of MMD^2 can be figured out as:

$$\begin{aligned} MMD^2 = &\frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) \\ &+ \frac{1}{n(n-1)} \sum_{i \neq j}^{m} k(y_i, y_j) \end{aligned} \tag{5.5}$$

In our method, the minimum variance unbiased estimate of MMD value is obtained according to the above functions based on Shogun package in python (Sonnenburg, Henschel, Widmer, Behr, Zien, Bona, Binder, Gehl, Franc et al. 2010). The computational complexity of the MMD method is $O(n^2)$. The MMD score can evaluate the gene expression difference between different sample types, while a higher MMD score means greater gene expression level difference.

5.2.4 Boundary discovery method

As a biomarker, there should be an expression threshold for the marker gene as the indicator for disease diagnosis. If the gene expression level is proven to be different in normal and tumor tissues, it is necessary to define a threshold of expression level as the boundary. When the gene expression level has a distinct gap between normal and tumor samples, the threshold is commonly the lower or upper boundary of this gap. However, the expression level does not have that kind of obvious gap all the time, thus how to define a reliable boundary is challenging in these cases.

Here we propose to identify the threshold with information theory, which has been widely used in decision tree algorithms for classification problems. According to the information theory, the change of information entropy, which is named information gain, can evaluate the classification efficiency of a threshold point. If there is the expression data of a gene from m normal samples and n tumor samples in dataset D, p_m and p_n refer to the proportions of normal and tumor samples in all samples. The original entropy of D is defined as:

$$Ent(D) = -\sum_{k=m,n} p_k \log_2 p_k \tag{5.6}$$

In the boundary identification, all samples are re-classified by the gene expression level with a split point of x and D^v denotes the new dataset re-classed by x. Then the information gain of this split point can be computed as:

$$Gain(D, x) = Ent(D) - \sum_{v=1}^{2} \frac{|D^v|}{|D|} Ent\left(D^v\right) \tag{5.7}$$

Unlike discrete data, the expression level is continuous, and it is inappropriate to use the expression level values in samples as the split points. Besides, as the distribution of the expression level is also unknown, we cannot use the probability function to calculate the entropy. Here, we propose to deal with continuous data like discrete data. First, the expression level values are sorted from small to large, and the middle points between two expression level values are taken as the split points. Second, we calculate the information gain of the split points respectively and choose the point with the highest information gain as the boundary. The algorithm of expression boundary identification with information theory is illustrated in Algorithm S1 in Additional File 7.

5.2.5 GO and KEGG enrichment analysis

The GO enrichment analysis is the major gene-annotation analysis method based on the Gene Ontology resource, describing the gene function at a

molecular level. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis has been widely used to model and simulate the molecular interactions and reaction networks in system biology. In this paper, these two methods are applied to figure out the molecular functions of identified potential marker genes and validate whether these genes are related to lung cancer. Here the enrichment analysis methods are both implemented based on the R package called ClusterProfiler developed by Guangchuang Yu's team (Yu, Wang, Han & He 2012). The GO terms and enriched pathways are all filtered with the p-value < 0.05.

5.2.6 Conventional DEA method and machine learning evaluation metrics

In this work, two conventional differentially expressed gene analysis methods, t-test and fold change, are compared with the proposed kernel MMD. The t-test is completed based on a python package called 'Scipy' (Bressert 2012). The fold change is calculated as below:

$$FoldChange = \left| log_2(\frac{E1}{E2}) \right| \tag{5.8}$$

Where *E1* and *E2* is the average of gene expression level in two different issue types. The p-value, fold change value, and MMD score is calculated for every single gene in our datasets. Then genes are ranked with the same strategy, and top-ranking genes are regarded as potential markers. Here 10-fold cross-validation based on the random forest classifier is applied to evaluate the efficiency of these top genes under four frequently used metrics: recall, F1-score, accuracy, and Matthews correlation coefficient (MCC).

5.3 Results

In the first part, we present the genes ranking with kernel MMD score and analysis the gene expression difference between different issue types. Then

the top-ranked genes are reported and compared with those genes identified by the conventional t-test and fold change methods. The third part shows the results of GO and KEGG pathway analysis of the top-ranked genes. At last, we choose the top ten genes of average ranking as marker genes and identify the expression boundaries of these gene markers with information gain theory.

5.3.1 Gene differential expression between different tissue types

Table 5.1: **Top ranking expressed genes between two type of issues** (NAT: Normal Adjacent Tumor)

Ranking	Normal-NAT	MMD scores	Normal-Tumor	MMD scores	NAT-Tumor	MMD scores
1	LOC442459	81.56	LOC442459	300.89	RS1	67.06
2	DOM3Z	70.85	LOC100132831	293.86	C10orf67	58.85
3	LOC100132831	68.89	LOC401127	288.92	ODAM	57.90
4	LOC401127	67.45	PIN1P1	265.11	LOC100128164	57.16
5	CSNK1A1P1	67.02	CSNK1A1P1	264.75	SH3GL3	56.96
6	MKRN9P	66.54	WNT2B	248.53	JPH4	56.68
7	TPI1P2	65.14	LOC100287632	247.69	SGCG	56.56
8	CYP2D7P1	64.72	CSNK1A1L	247.45	GYPE	55.70
9	CSNK1A1L	63.69	LOC100507373	244.54	LOC643650	53.05
10	PIN1P1	62.24	AOC4	240.66	IHH	52.79

For the three mentioned datasets, kernel MMD values are calculated on each two of them respectively to discover DEGs. For every single gene, we calculate three MMD values from Normal-NAT, Normal-Tumor, and NAT-Tumor groups. The MMD scores indicate the difference in expression levels among three types of samples. The top 10 ranked genes in each group are shown in Table 5.1. As illustrated in the table, the top MMD scores in the Normal-Tumor group are over 200, which are much higher than the other two groups. The Normal-NAT group has comparable MMD scores with the NAT-Tumor group. Gene expression level difference in the normal-tumor group is much greater than the other two groups. More detailed information is listed in Additional file 6.

In addition, the NAT samples have different expression profiles from not only tumor samples but also the real healthy samples. The NAT samples are always considered healthy samples in state-of-art researches, and we test the top 10 ranked genes selected by the NAT-Tumor group, Normal-Tumor group, and their average ranking to explore the influence of regarding NAT as real normal samples. The expression data of the top genes above are applied to classify tumor samples from other samples via 10-fold cross-validation to evaluate the effectiveness of selected genes. The results of the 10-fold cross-validation are reported in Table 5.2.

Table 5.2: **Cross-validation performance of top ten genes from different groups** (NAT: Normal Adjacent Tumor)

Group	Recall	F1	Accuracy	MCC
Normal-Tumor	0.9857	0.9540	0.9476	0.8659
NAT-Tumor	0.9534	0.9670	0.9640	0.9279
Average	0.9885	0.9914	0.9907	0.9816

As shown in Table 5.2, the selected genes from each group can classify tumor samples from other samples. However, the performance of the three groups of genes varies greatly. When considering normal samples and NAT samples together, the top average ranked genes have the best scores under all metrics with an accuracy of 0.9907. The highest F1 score of 0.9914 implies that these genes also have a better classification balance. The results show that the real normal samples and NAT samples are not exactly the same. Researchers should take both of them into consideration in cancer study rather than simply replacing real normal samples with NAT samples.

5.3.2 Identify marker genes in cancer development

In this work, two conventional DEA methods t-test and fold change are compared with our approach. T-test and fold change methods are both applied to identify DEGs between different tissue types. The p-value of the

t-test and fold change values are calculated to evaluate the gene expression difference. Since the ability to detect tumor samples is more significant in clinical application, the top 10 genes of average rankings from the Normal-Tumor group and NAT-tumor group selected by the t-test and fold change are compared with the genes selected by our method. Another 10-fold cross-validation is conducted, and the results are reported in Table 5.3.

Table 5.3: **Cross-validation performance of top ten genes selected by different DEA methods**

Method	Recall	F1	Accuracy	MCC
Fold Change	0.7044	0.7992	0.8048	0.6382
T-test	0.9796	0.9815	0.9794	0.9582
Kernel MMD	0.9885	0.9914	0.9907	0.9816

As shown in Table 5.3, the proposed kernel MMD method outperforms the other two conventional methods under all metrics with the recall of 0.9885, F1 score of 0.9914, the accuracy of 0.9907, and MCC of 0.9816. The fold change method has the worst performance, and the selected genes by fold change method are not efficient enough to classify tumors from other samples. The t-test has a comparable result with the MMD method. Since the t-test and fold change methods have been widely used, the kernel MMD method is promising to improve the differential gene analysis efficiency in practical use.

From Table 5.1, we can see there are some overlapping genes like LOC442459, LOC100132831, LOC401127, CSNK1A1P1, CSNK1A1L, and PIN1P1 in Normal-NAT group and Normal-Tumor group. These genes can distinguish normal samples from not only NAT samples, but also tumor samples. Inspired by the previous part, the average ranking of all groups can identify more significant genes. Thus, the gene average ranking of the three groups is calculated, and top genes of average ranking are chosen to be potential marker genes to diagnose lung cancer. In Figure 5.2, expression levels in normal, NAT, and tumor samples of the top 4 genes of average ranking are presented. From the figure, the four genes exactly have distinct

expression levels in different types of tissues.

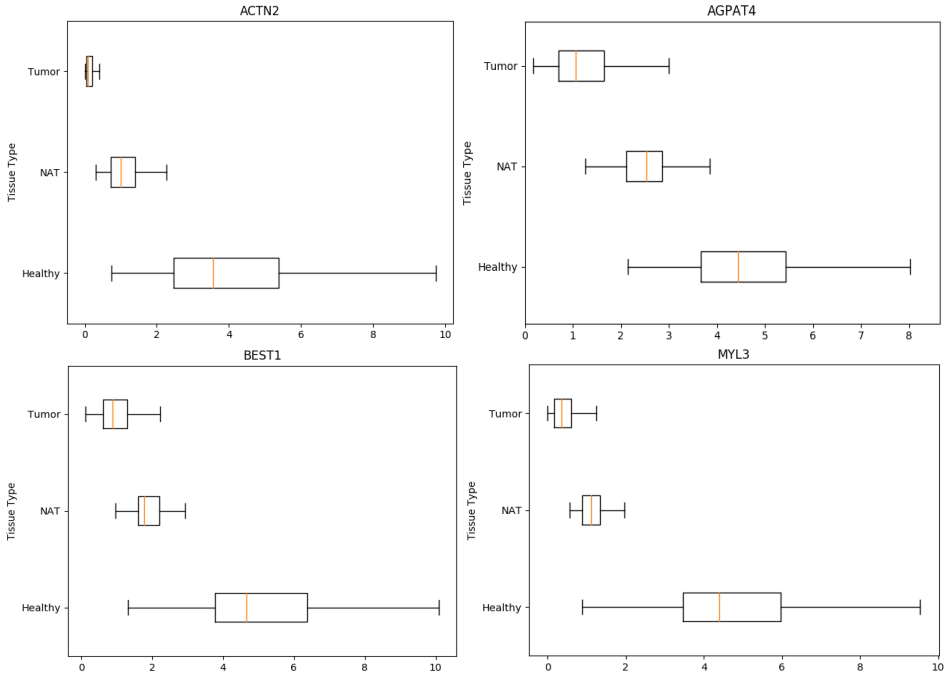

Figure 5.2: **Box-plot of gene expression levels in three tissue types.**The X-axis is the FPKM expression level; the Y-axis is the tissue type.

5.3.3 GO and KEGG pathway enrichment

From the average ranking gene list, we choose the top 100 genes to conduct the GO and KEGG pathway enrichment analysis. In the GO enrichment analysis, we select 'Biology Process' as the enrichment target, and there are 12 GO terms with p-value $<$ 1.0e-04 and count $\geq$ 5. As shown in Table 5.4, the top two terms, 'GO:0051480' and 'GO:0007204', are both related to the regulation factors of cytosolic calcium ion concentration while term No.5 and No.6 are also involved in cellular calcium ion homeostasis. The influence of calcium ion channels on lung cancer has been studied for a long time (Moody, Murphy, Mahmoud & Fiskum 1987, Moody, Staley, Zia, Coy & Jensen 1992,

Arbabian, Brouland, Apáti, Pászty, Hegedűs, Enyedi, Chomienne & Papp 2013), and the cellular calcium ion level change has been explored in lung cancer development (Arbabian et al. 2013). It suggests that these genes related to calcium ion regulations are significant in lung cancer.

Table 5.4: **Go function analysis for the top ranking genes** (p-value $<$ 1.0e-04 and count $\geq$ 5).

No.	GOBPID	p-Value	Count	Term
1	GO:0051480	7.6032e-07	10	regulation of cytosolic calcium ion concentration
2	GO:0007204	3.0453e-06	9	positive regulation of cytosolic calcium ion concentration
3	GO:0019229	4.4969e-06	5	regulation of vasoconstriction
4	GO:0007200	6.6689e-06	6	phospholipase C-activating G-protein coupled receptor signaling pathway
5	GO:0006874	7.5060e-06	10	cellular calcium ion homeostasis
6	GO:0055074	9.4074e-06	10	calcium ion homeostasis
7	GO:0042310	1.4462e-05	5	vasoconstriction
8	GO:0072503	1.5632e-05	10	cellular divalent inorganic cation homeostasis
9	GO:0072507	2.1785e-05	10	divalent inorganic cation homeostasis
10	GO:0097756	2.3563e-05	5	negative regulation of blood vessel diameter
11	GO:0007189	6.5898e-05	5	adenylate cyclase-activating G-protein coupled receptor signaling pathway
12	GO:0019932	7.4403e-05	8	second-messenger-mediated signaling

The results of KEGG pathway enrichment analysis are illustrated in Figure 5.3. There are 20 pathways with a p-value below 0.05 and count number over 2. The adrenergic signaling pathway and the cGMP-PKG signaling pathway are the most significant pathways. Currently, the role of adrenergic signaling pathways plays in lung cancer development has not been fully studied. However, the β-adrenergic signaling has been found to be a possible novel cancer therapy in tumor cells (Schuller 2010). Besides, some researches have made some explorations about that (Schuller & Cekanova 2005). The second top significant pathway is the cGMP-PKG signaling pathway, which mediates the action of cellular ion concentration and sensitivity, influencing cell proliferation. The regulation relationship between the cGMP-PKG signaling pathway and lung cancer has been studied in (Wong, Bathina & Fiscus 2012). The results of GO and KEGG pathway enrichment analysis show that the top gene selected by the MMD method is indeed highly related to lung cancer.

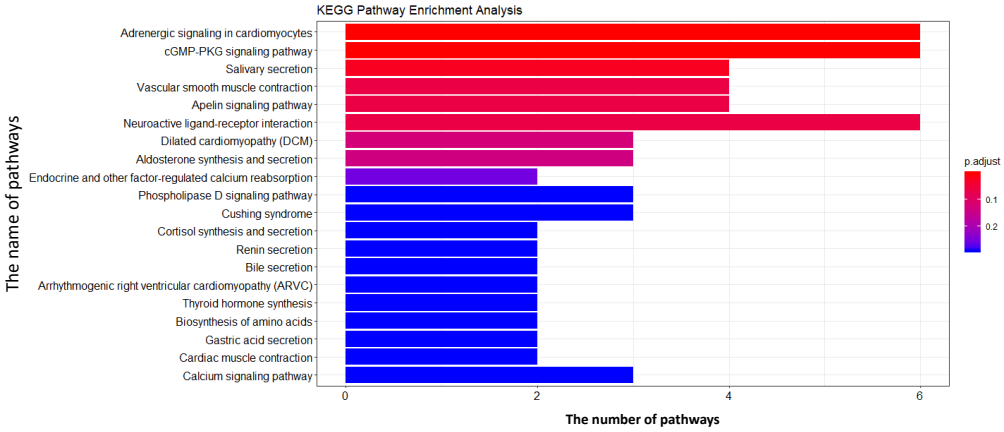

Figure 5.3: **KEGG pathway enrichment analysis for top ranking genes.**

5.3.4 Expression boundary identification

Although the conventional methods can detect the differential expressed gene, they can only manually define the expression boundary when there is a distinct expression level gap. After selecting lung cancer marker genes, we identify the expression boundaries between normal, NAT, and tumor with the mentioned information theory method. Here the top 10 genes in MMD average ranking list are chosen as the lung cancer marker genes, and the expression boundaries of them are illustrated in Table 5.5. The detailed results are presented in Additional file 6.

As shown in Table 5, the ten gene markers have a distinct expression range in normal, NAT, and tumor samples, which can be an indicator of lung cancer development. Additionally, in practical clinic applications, the boundary between tumor and other tissues is the most significant for disease diagnosis. The boundary between normal samples and NAT samples also implied that there would be some gene expression changes in the disease development, and the NAT samples may serve to detect cell carcinogenesis, which can help to understand the lung cancer mechanisms.

Table 5.5: **Expression boundary of lung cancer biomarkers** (e : FPKM expression level)

Gene Name	Normal	Normal Adjacent Tumor	Tumor
ACTN2	$e \geq 3.5247$	$0.7146 < e < 3.5247$	$e \leq 0.7146$
MYL3	$e \geq 5.3223$	$4.9211 < e < 5.3223$	$e \leq 4.9211$
AGPAT4	$e \geq 4.3722$	$2.3052 < e < 4.3722$	$e \leq 2.3052$
BEST1	$e \geq 3.7487$	$1.6216 < e < 3.7487$	$e \leq 1.6216$
TWIST2	$e \geq 4.2450$	$1.2030 < e < 4.2450$	$e \leq 1.2030$
LINC00472	$e \geq 3.4721$	$0.8045 < e < 3.4721$	$e \leq 0.8045$
MYO7B	$e \geq 4.1723$	$0.7450 < e < 4.1723$	$e \leq 0.7450$
CCNF	$e \leq 16.4506$	$16.4506 < e < 20.5656$	$e \geq 20.5656$
NECAB1	$0.9961 < e < 4.7770$	$e \geq 4.7770$	$e \leq 0.9961$
NOTCH4	$e \geq 4.6829$	$1.9808 < e < 4.6829$	$e \leq 1.9808$

5.4 Discussion

Since the early-diagnosis of lung cancer has been a long-term critical problem in clinical practice, researchers have explored various types of biomarkers, like genetic mutations, blood proteins. Here, this paper proposed a novel method to identify genes markers for lung cancer. There are two main problems in efficient gene markers identification: first, how to evaluate the gene expression difference; second, how to find the reliable expression boundary between tumor and other samples. The most existing DEA methods were built to solve the first problem, but they can only give out a p-value to assess the differential expressing gene without defining the expression boundary. The motivation of this research is to address both of the problems in biomarker identifications.

Although the gene markers are given out based on the existing lung cancer dataset, we think there are two limitations to our work. First, a larger dataset can help to obtain more accurate results. Second, a threshold of MMD value to define the differentially expressed gene can be set with a large dataset, while here we just take the top-ranked genes as potential marker genes.

5.5 Summary

In this chapter, we proposed a more efficient method, kernel MMD to evaluate the expression changes, and an information theory-based algorithm to identify the gene expression boundary. This method addressed the research question **Q3** (see **Section 1.2**). The experiment results show our method can select more significant genes than traditional methods and give out the expression boundary of the marker gene. Through the GO and KEGG pathway enrichment analysis, the function of marker genes in lung cancer is studied, and these marker genes are indeed related to lung cancer development. We will collect more gene expression data related to lung cancer and calculate more accurate results in the future. Besides, we will explore the application of our method of biomarker discovery for other diseases.

Chapter 6

Conclusions and future work

6.1 Conclusions

In this thesis, three research problems are addressed, namely DNA N^4-methylcytosine site prediction, mRNA N^6-methyladenosine site prediction, and lung cancer gene markers identification. The proposed computational methods for the three problems are discussed in Chapter2-5 and have been represented in my three published journal papers (see the **list of Publications**). Chapter 3 describes the method to predict the 4mC site in the DNA sequence, and Chapter 4 presents the technique designed to predict m^6A site in imbalanced mRNA situations. Chapter 5 introduce the novel gene marker identification method with kernel statistics and information theory. The work, motivation, and contributions of this thesis are concluded below:

In Chapter 3, a novel method is developed for accurate prediction of DNA N^4-methylcytosine site via a boost-learning of various types of sequence features. This method has advantages over the existing methods. Firstly, the adjacent nucleotide patterns are discovered in the sequence logos, which is important for the feature representation. Secondly, three sequence features, such as k-nucleotide frequency, k-spectrum nucleotide pair frequency, and PseDNC, are employed to extract the local contiguous nucleotides sequence

characteristics. Together with two global features, the mentioned features are integrated into a 292-dimensional feature space. Lastly, an embedding feature selection scheme based on the XGBoost machine is applied before training the SVM prediction model. Compared with the existing F-score scheme, the embedding feature selection is more meaningful. The optimizations of the feature space and feature selection scheme solve the problems of the existing methods. The independent test, 10-fold cross-validation, and two case studies all confirm the reliability and accuracy of the proposed method.

Chapter 4 presents an imbalance learning method named HMPre to identify m^6A site in human mRNA. This method improves the model computation performance in several aspects. Rather than represent the sequence only with text feature, this method utilizes three novel biological features: site location, information entropy gain, and SNP variants. Compared with the existing methods, these biological features are from the latest research and proposed for m^6A prediction the first time. In the feature construction process, a feature selection algorithm combining Fisher exact test and MRMR approach is built to select significant SNP positions, improve feature efficiency, and reduce noise from feature space. The existing prediction methods are all trained on balanced data to avoid over-fitting on negative samples, t, where the negative samples are selected randomly. However, the proposed method with a weighted XGBoost machine learns from the imbalance dataset containing all negative data, which is more similar to the practical situations. In the model evaluation, four metrics are adopted, such as precision, recall, F1, and MCC. The comparison with three existing methods in an independent test dataset and two case studies reveal the correctness and robustness of the HMPre method.

The problem of gene marker identification is addressed in Chapter 5. This part solves two main problems in gene marker identification: select differential expressing marker genes, and identify the expression range of marker genes. The innovations of this work lie in: (1) Since the traditional research considered the norm tissue adjacent to the tumor as the control

group, true normal tissue is introduced in the lung cancer biomarker research for the first time. The tissues are divided into three types: true normal, normal adjacent to tumor (NAT), and tumor tissue. (2) Unlike the existing DEA methods with specific statistic hypothesis, this method proposes to apply a more general kernel approach to evaluate the distribution difference. The kernel MMD calculates the maxi-mum mean discrepancy score without presupposing of data distribution, indicating the degree of differential expression. (3) This method presents a gene expression boundary detection algorithm based on information theory, which solves the boundary identification problem when there is no distinct gene expression gap between groups. Compared with the traditional DEA method, the kernel MMD method has better efficiency, and results from KEGG and GO enrichment shows the select marker genes are meaningful. The expression boundaries detected by the method can be applied in the early diagnosis of lung cancer.

6.2 Future work

This decade has witnessed the fast development of computational biology and bioinformatics. The rapid progress of information technology promotes research on epigenetic modification and gene expression significantly. However, there are still many issues to be addressed in these fields. The NGS technologies have accelerated the application of machine learning techniques in genomics studies, and the developed biological tools should aim at practical situations. Under this background and trend, our future work will focus on the following problems.

- **Sequence feature space optimization**
  Currently, the computational epigenetic modification site identification methods are mainly based on base segments in DNA or RNA sequence. In the existing methods, the target site is in the central position, and a fixed-length flanking window is taken out as the input of machine learning. However, there are several aspects to be considered in

the future: (1). The length of the sequence fragment should be evaluated before the model construction. The modifications have the specific binding site of methylation enzymes near the modified site, and the flanking window should contain these binding sites. (2) In Chapter 4, the proposed biological features immensely improve performance. In future work, more biological characteristics of the epigenetic modifications can be employed. For example, the SNP variants can be used in the prediction of the 4mC site. The diversity of sequence features can extract the pattern more accurately. (3) The feature selection scheme is an efficient way to reduce the feature dimension and improve model performance. For the classification problem, more schemes of feature selection should be explored.

- **Data driven machine learning algorithm optimization**
  The machine learning-based prediction model in the existing method is trained on the experiment validated datasets. However, there are still several problems that are also widely existing in other bioinformatics fields. First, the size of training data is limited in these tools. With the development of relative technology, more data should be collected to expand the benchmark datasets to improve model accuracy and generalization ability. Second, the lack of reliable negative samples. Since the positive samples in training data are validated from the wet-lab experiment, the negative samples are usually randomly selected, and reliable negative sample selection algorithms should be studied in the future. Last, learning from unbalanced data. The positive samples are much less than negative samples in practical situations, and the data imbalance issue needs to be considered in the 4mC prediction problem.

- **Regulation mechanism mining for epigenetic modifications**
  In future work, we will mine the co-regulation relationship between epigenetic modifications and other regulatory factors. Then, the

molecular mechanism of disease development related to epigenetic modifications can be studied. For example, since the dynamic modification process requires proteins to binding to mRNA, miRNA may work cooperatively with m6A to regulate an individual gene or a cohort of genes that participate in similar processes. As the m6A expression level in genes is applied, we can analyze the relationship of m6A-related proteins and miRNA binding sites to verify the results. With the m6A and miRNAs co-regulation pairs, we can build m6A-gene-miRNA regulation networks for a certain disease. With network analysis algorithms, we can mine more details about the molecular mechanisms of post-transcript modifications that affect the development of diseases and potential medicine targets for treatment.

- **Disease gene marker identification**

  The proposed identification method for the lung cancer gene marker addresses the two fundamental problems: select marker genes and define the expression range for marker genes. The gene markers are given out based on the current gene expression profiles. In the future work, we will focus on three folds to improve the method: First, a larger dataset containing more expression profiles will be built to obtain more accurate results; Second, a threshold of MMD value will be studied for the differentially expressed genes, rather than just taking the top-ranked genes as potential marker genes; Third, the selected marker genes are validated with KEGG and GO pathway enrichment analysis, and the identified gene marker should still be evaluated in the practical situations, such as disease early-diagnosis and prognosis.

Appendix A

Appendix: Methodology foundation

A.1 Applied statistical methods

A.1.1 Information entropy

The information entropy is a basic quantity related to a random variable to measure the uncertainty for an event with a probability distribution (Shannon 1948, Borda 2011). In Chapter 4 and Chapter 5, the information entropy is applied to represent the sequence feature and detect the expression boundary of the marker gene.

For a given variable X with possible outcome x_1, x_2, ..., x_n, the information entropy of X is defined as:

$$E(X) = -\sum_{i=1}^{n} f_{x_i} \log\left(f_{x_i}\right) \tag{A.1}$$

where f_{x_i} is the frequency of x_i in the outcome of variable X. In the feature representation of Chapter 4, $x_i \in$ (A,G,U,C), while in Chapter 5, $x_i \in$ (group1, group2) and the group refers to Normal, NAT or Tumor.

A.1.2 Fisher's exact test

Proposed by Ronald Fisher, fisher's exact test is a statistical significance test for the analysis of contingency tables (Fisher 1922, Agresti et al. 1992). The test is applied to cope with a small number of observations, which are usually presented with a 2*2 contingency table (Bower 2003). In Chapter 4, fisher's exact test is adopted to select the specific SNP positions, which helps construct the novel feature.

Table A.1: **The example of 2*2 contingency table**

Groups	category A	category B
Observation 1	a	b
Observation 2	c	d

The output of fisher's exact test is p-value, indicating the significance of the statistic deviation from the hyper geometric distribution. The example of observed 2*2 contingency table is shown in Table A.1, and the p-value can be calculated with the formula:

$$[oddsratio, p\_value] = stats.fisher\_exact([a, b], [c, d]) \tag{A.2}$$

In the formula, the function stats.fisher_exact is from Python package SciPy (vision 1.4.1). The output p_value is calculated with the default settings.

A.2 Adopted machine learning algorithms

A.2.1 Support vector machine

A support vector machine is a supervised machine learning model with associated learning algorithms, coping with classification, or regression analysis. For a dataset containing n samples (x_1,x_2,...,x_n), the label of samples $y_i \in 0, 1$. The SVM is optimized with the following formulation:

Given a dataset with l samples x_i ($x_i \in R^n, i \in \{1, 2, ..., l\}$) and their labels y_i ($y_i \in \{1, -1\}$), the SVM solves the following primal optimization problem (Chang & Lin 2011):

$$\min_{\alpha}(L(\alpha)) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{n}\alpha_j \tag{A.3}$$

$$\text{subject to} \sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leqslant \alpha_i \leqslant C \tag{A.4}$$

where α_i is the Lagrange multipliers, and C is the parameter to regulate the support vector. $K(x_i, x_j)$ is the kernel fucntion and the Radial Basis Function (RBF) kernel is used in Chapter 3 as the prediction model. The RBF kernel is defined as:

$$k_{\sigma}^{RBF} = \exp\left(-\frac{1}{\sigma}\left\|x_i - x_{j'}\right\|^2\right) \tag{A.5}$$

Where θ is the parameter to regulate the radial basis width, and it is optimized by cross-validation in the thesis along with parameter C.

A.2.2 XGBoost

XGBoost (eXtreme Gradient Boosting) is a tree boosting algorithm, which is an advanced implementation of gradient boosting algorithm developed by Chen (Chen & Guestrin 2016). XGBoost has several advantages over other machine learning classifiers: Firstly, there is a regularization process, effectively preventing model over-fitting. Secondly, embedded parallel processing allows faster-learning speed. Thirdly XGBoost is of high flexibility, and users can define customized optimization objectives and evaluation criteria.

In this thesis, XGBoost is used in Chapter 3 for feature importance calculation, and in Chapter 4 for imbalance learning. In Chapter 3, we use the default setting of XGBoost, and 'auc' is the learning metric. In Chapter 4, XGBoost classifier learns from unbalanced training data with class

weight, and 'roc' is taken as evaluation criteria. We implement the model with a python package named xgboost (vision 0.6a2), which is available at https://github.com/dmlc/xgboost.

A.3 Cross validation and evaluation metrics

A.3.1 Cross validation

In data mining, the model evaluation is an important process, and cross-validation is one of the evaluation strategies. In this thesis, cross-validation is employed in several situations of Chapter 3 and Chapter 4: 1.selected feature subset evaluation; 2. model parameter optimization; 3. model performance evaluation.

The cross-validation on existing data can estimate the model performance and generalization ability on independent data. According to the fold number n, it's also named n-fold cross-validation, in which the data is divided into n equal subgroups. In each round, n-1 subgroups are taken as the training data, and the rest one is tested. After n rounds, each sample in the dataset has a predicted label. Then the evaluation metrics are calculated with the predicted label and original label for the whole dataset.

A.3.2 Performance evaluation metrics

The following evaluation indices are used to evaluate the performance of the prediction/classification models:

$$Sensitivity(Sn) = \frac{TP}{TP + FN} \times 100\% \tag{A.6}$$

$$Specificity(Sp) = \frac{TN}{TN + FP} \times 100\% \tag{A.7}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{A.8}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{A.9}$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{A.10}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{A.11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{A.12}$$

Where TP and FP are the counts of correctly and falsely predicted positive samples, respectively; TN and FN are the counts of correctly and falsely predicted negative samples, respectively. Except for the above metrics, the AUC (area under the receiver operating characteristic curve) is used as evaluation metrics as well. Sensitivity, Specificity, Precision, recall, accuracy, F1-score, and MCC can be calculated using the above formulas. AUC can be calculated with the ROC curve using scientific tool packages (e.g., scikit-learn in Python).

Appendix B

Appendix: List of Supplementary files

The Additional file list and the corresponding download links

name	chapter	description	link
Additional file 1	3	supplementary tables for chapter 3	download
Additional file 2	3	feature importance scores for Chapter 3	download
Additional file 3	3	case study results for Chapter 3	download
Additional file 4	4	human mature mRNA m^6A data sets	download
Additional file 5	4	supplementary materials for Chapter 4	download
Additional file 6	5	EDA and boundary results for Chapter 5	download
Additional file 7	5	boundary detection algorithm for Chapter 5	download
Additional file 8	3	code and data for Chapter 3	open
Additional file 9	4	code and data for Chapter 4	open
Additional file 10	5	code and data for Chapter 5	open

Appendix C

Appendix: List of Symbols

The following list is neither exhaustive nor exclusive, but may be helpful.

$4mC$	DNA N4-methylcytosine
OHB	One hot binary encode
SNF	Sequential nucleotide frequency
KNF	K-nucleotide frequency
$KSNPF$	K-spectrum nucleotide pair frequency
ROC	Receiver operating characteristic curve
MCC	Matthews correlation coefficient
ACC	Accuracy
$m6A$	N^6-Methylation
$MRMR$	Max-Relevance Min-Redundancy algorithm
SNP	singe nucleotide polymorphism
PTM	post-transcription modifications
NGS	next generation sequence

PCA	Principal component analysis
CPD	chemical property with density feature
AUC	Area under the receiver operating characteristic curve
DEA	differentially expressed analysis
MMD	maximum mean discrepancy
NAT	normal adjacent to the tumor
GO	Gene Ontology
$KEGG$	Kyoto Encyclopedia of Genes and Genomes
CV	Cross validation
SVM	Support vector machine
RBF	Radial basis function

Bibliography

Agresti, A. et al. (1992), 'A survey of exact inference for contingency tables', *Statistical science* **7**(1), 131–153.

Anders, S. & Huber, W. (2010), 'Differential expression analysis for sequence count data', *Nature Precedings* pp. 1–1.

Andre, F., Schartz, N. E., Movassagh, M., Flament, C., Pautier, P., Morice, P., Pomel, C., Lhomme, C., Escudier, B., Le Chevalier, T. et al. (2002), 'Malignant effusions and immunogenic tumour-derived exosomes', *The Lancet* **360**(9329), 295–305.

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M. & Butte, A. J. (2017), 'Comprehensive analysis of normal adjacent to tumor transcriptomes', *Nature communications* **8**(1), 1077.

Arbabian, A., Brouland, J.-P., Apáti, Á., Pászty, K., Hegedűs, L., Enyedi, Á., Chomienne, C. & Papp, B. (2013), 'Modulation of endoplasmic reticulum calcium pump expression during lung cancer cell differentiation', *The FEBS journal* **280**(21), 5408–5418.

Auluck, P. K., Chan, H. E., Trojanowski, J. Q., Lee, V. M.-Y. & Bonini, N. M. (2002), 'Chaperone suppression of α-synuclein toxicity in a drosophila model for parkinson's disease', *Science* **295**(5556), 865–868.

Baker, D. & Sali, A. (2001), 'Protein structure prediction and structural genomics', *Science* **294**(5540), 93–96.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M. et al. (2012), 'Ncbi geo: archive for functional genomics data sets—update', *Nucleic acids research* **41**(D1), D991–D995.

Baxevanis, A. D., Bader, G. D. & Wishart, D. S. (2020), *Bioinformatics*, John Wiley & Sons.

Berger, S. L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. (2009), 'An operational definition of epigenetics', *Genes & development* **23**(7), 781–783.

Bestor, T., Laudano, A., Mattaliano, R. & Ingram, V. (1988), 'Cloning and sequencing of a cdna encoding dna methyltransferase of mouse cells: the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases', *Journal of molecular biology* **203**(4), 971–983.

Bhattacharya, S. K., Ramchandani, S., Cervoni, N. & Szyf, M. (1999), 'A mammalian protein with specific demethylase activity for mcpg dna', *Nature* **397**(6720), 579–583.

Bhutani, N., Burns, D. M. & Blau, H. M. (2011), 'Dna demethylation dynamics', *Cell* **146**(6), 866–872.

Bird, A. (2001), 'Methylation talk between histones and dna', *Science* **294**(5549), 2113–2115.

Bokar, J. A. (2005), The biosynthesis and functional roles of methylated nucleosides in eukaryotic mrna, *in* 'Fine-tuning of RNA functions by modification and editing', Springer, Berlin,Heidelberg, pp. 141–177.

Borda, M. (2011), *Fundamentals in information theory and coding*, Springer Science & Business Media.

Bower, K. M. (2003), When to use fisher's exact test, *in* 'American Society for Quality, Six Sigma Forum Magazine', Vol. 2, pp. 35–37.

Bressert, E. (2012), *SciPy and NumPy: an overview for developers*, " O'Reilly Media, Inc.".

Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. (2010), 'Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments', *BMC bioinformatics* **11**(1), 94.

Capelozzi, V. L. (2009), 'Role of immunohistochemistry in the diagnosis of lung cancer', *Jornal Brasileiro de Pneumologia* **35**(4), 375–382.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G. & Wang, W. (2006), 'Data mining curriculum: A proposal (version 1.0)', *Intensive Working Group of ACM SIGKDD Curriculum Committee* **140**.

Chan, C. T., Dyavaiah, M., DeMott, M. S., Taghizadeh, K., Dedon, P. C. & Begley, T. J. (2010), 'A quantitative systems approach reveals dynamic control of trna modifications during cellular stress', *PLoS genetics* **6**(12).

Chang, C.-C. & Lin, C.-J. (2011), 'Libsvm: A library for support vector machines', *ACM transactions on intelligent systems and technology (TIST)* **2**(3), 1–27.

Chen, C., Chen, H., He, Y. & Xia, R. (2018), 'Tbtools, a toolkit for biologists integrating various biological data handling tools with a user-friendly interface', *BioRxiv* p. 289660.

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., Su, J., de Magalhaes, J. P., Rigden, D. J. & Meng, J. (2019), 'Whistle: a high-accuracy map of the human n 6-methyladenosine (m6a) epitranscriptome predicted using a machine learning approach', *Nucleic acids research* **47**(7), e41–e41.

Chen, K., Zhao, B. S. & He, C. (2016), 'Nucleic acid modifications in regulation of gene expression', *Cell chemical biology* **23**(1), 74–85.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chen, W., Feng, P., Ding, H. & Lin, H. (2016), 'Identifying n6-methyladenosine sites in the arabidopsis thaliana transcriptome', *Molecular Genetics and Genomics* **291**(6), 2225–2229.

Chen, W., Feng, P., Ding, H., Lin, H. & Chou, K.-C. (2015), 'irna-methyl: identifying n6-methyladenosine sites using pseudo nucleotide composition', *Analytical Biochemistry* **490**, 26–33.

Chen, W., Feng, P.-M., Lin, H. & Chou, K.-C. (2013), 'irspot-psednc: identify recombination spots with pseudo dinucleotide composition', *Nucleic acids research* **41**(6), e68–e68.

Chen, W., Tang, H. & Lin, H. (2017), 'Methyrna: a web server for identification of n6-methyladenosine sites', *Journal of Biomolecular Structure and Dynamics* **35**(3), 683–687.

Chen, W., Tang, H., Ye, J., Lin, H. & Chou, K.-C. (2016), 'irna-pseu: Identifying rna pseudouridine sites', *Molecular Therapy—Nucleic Acids* **5**(7), e332.

Chen, W., Tran, H., Liang, Z., Lin, H. & Zhang, L. (2015), 'Identification and analysis of the n 6-methyladenosine in the saccharomyces cerevisiae transcriptome', *Scientific Reports* **5**, 13859.

Chen, W., Xing, P. & Zou, Q. (2017), 'Detecting n6-methyladenosine sites from rna transcriptomes using ensemble support vector machines', *Scientific Reports* **7**, 40242.

Chen, W., Yang, H., Feng, P., Ding, H. & Lin, H. (2017), 'idna4mc: identifying dna n4-methylcytosine sites based on nucleotide chemical properties', *Bioinformatics* **33**(22), 3518–3523.

Cheng, X. (1995), 'Dna modification by methyltransferases', *Current opinion in structural biology* **5**(1), 4–10.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003), 'Multiple sequence alignment with the clustal series of programs', *Nucleic acids research* **31**(13), 3497–3500.

Choi, J., Ieong, K.-W., Demirci, H., Chen, J., Petrov, A., Prabhakar, A., O'leary, S. E., Dominissini, D., Rechavi, G., Soltis, S. M. et al. (2016), 'N6-methyladenosine in mrna disrupts trna selection and translation-elongation dynamics', *Nature Structural & Molecular Biology* **23**(2), 110–115.

Chow, C. S., Lamichhane, T. N. & Mahto, S. K. (2007), 'Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications', *ACS chemical biology* **2**(9), 610–619.

Clifton, C. (2010), 'Encyclopædia britannica: definition of data mining', *Retrieved on* **9**(12), 2010.

Cohen, A. M. & Hersh, W. R. (2005), 'A survey of current work in biomedical text mining', *Briefings in bioinformatics* **6**(1), 57–71.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. & Jacobsen, S. E. (2008), 'Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning', *Nature* **452**(7184), 215–219.

Cowley, A. W. (2006), 'The genetic dissection of essential hypertension', *Nature Reviews Genetics* **7**(11), 829–840.

Crews, K. R., Hicks, J. K., Pui, C.-H., Relling, M. V. & Evans, W. E. (2012), 'Pharmacogenomics and individualized medicine: translating science into practice', *Clinical Pharmacology & Therapeutics* **92**(4), 467–475.

Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. (2004), 'Weblogo: a sequence logo generator', *Genome research* **14**(6), 1188–1190.

Daly, A. K. (2010), 'Genome-wide association studies in pharmacogenomics', *Nature Reviews Genetics* **11**(4), 241–246.

Das, P. M. & Singal, R. (2004), 'Dna methylation and cancer', *Journal of clinical oncology* **22**(22), 4632–4642.

Davis, B. M., Chao, M. C. & Waldor, M. K. (2013), 'Entering the era of bacterial epigenomics with single molecule real time dna sequencing', *Current opinion in microbiology* **16**(2), 192–198.

Deng, X., Chen, K., Luo, G.-Z., Weng, X., Ji, Q., Zhou, T. & He, C. (2015), 'Widespread occurrence of n6-methyladenosine in bacterial mrna', *Nucleic Acids Research* **43**(13), 6557–6567.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M. et al. (2012), 'Topology of the human and mouse m6a rna methylomes revealed by m6a-seq', *Nature* **485**(7397), 201.

Doseeva, V., Colpitts, T., Gao, G., Woodcock, J. & Knezevic, V. (2015), 'Performance of a multiplexed dual analyte immunoassay for the early detection of non-small cell lung cancer', *Journal of translational medicine* **13**(1), 55.

Dupont, C., Armant, D. R. & Brenner, C. A. (2009), Epigenetics: definition, mechanisms and clinical perspective, *in* 'Seminars in reproductive medicine', Vol. 27, © Thieme Medical Publishers, pp. 351–357.

Edgar, R. C. (2004), 'Muscle: multiple sequence alignment with high accuracy and high throughput', *Nucleic acids research* **32**(5), 1792–1797.

Esteller, M. (2008), 'Epigenetics in evolution and disease', *The Lancet* **372**, S90–S96.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI magazine* **17**(3), 37–37.

Feany, M. B. & Bender, W. W. (2000), 'A drosophila model of parkinson's disease', *Nature* **404**(6776), 394–398.

Feng, Z., Li, W., Ward, A., Piggott, B. J., Larkspur, E. R., Sternberg, P. W. & Xu, X. S. (2006), 'A c. elegans model of nicotine-dependent behavior: regulation by trp-family channels', *Cell* **127**(3), 621–633.

Fisher, R. A. (1922), 'On the interpretation of χ 2 from contingency tables, and the calculation of p', *Journal of the Royal Statistical Society* **85**(1), 87–94.

Flannick, J. & Florez, J. C. (2016), 'Type 2 diabetes: genetic data sharing to advance complex disease research', *Nature Reviews Genetics* **17**(9), 535.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. & Turner, S. W. (2010), 'Direct detection of dna methylation during single-molecule, real-time sequencing', *Nature methods* **7**(6), 461.

Fontana, R. S., Sanderson, D. R., Taylor, W. F., Woolner, L. B., Miller, W. E., Muhm, J. R. & Uhlenhopp, M. A. (1984), 'Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the mayo clinic study', *American Review of Respiratory Disease* **130**(4), 561–565.

Frémont, M., Siegmann, M., Gaulis, S., Matthies, R., Hess, D. & Jost, J.-P. (1997), 'Demethylation of dna by purified chick embryo 5-

methylcytosine-dna glycosylase requires both protein and rna', *Nucleic acids research* **25**(12), 2375–2380.

Frost, J. K., Ball Jr, W. C., Levin, M. L., Tockman, M. S., Baker, R. R., Carter, D., Eggleston, J. C., Erozan, Y. S., Gupta, P. K., Khouri, N. F. et al. (1984), 'Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the johns hopkins study', *American Review of Respiratory Disease* **130**(4), 549–554.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012), 'Cd-hit: accelerated for clustering the next-generation sequencing data', *Bioinformatics* **28**(23), 3150–3152.

Fu, Y., Dominissini, D., Rechavi, G. & He, C. (2014), 'Gene expression regulation mediated through reversible m6a rna methylation', *Nature Reviews Genetics* **15**(5), 293–306.

Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, L. C. et al. (2015), 'N6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas', *Cell* **161**(4), 879–892.

Goetsch, C. M. (2011), Genetic tumor profiling and genetically targeted cancer therapy, *in* 'Seminars in oncology nursing', Vol. 27, Elsevier, pp. 34–44.

Gómez, J., García, L. J., Salazar, G. A., Villaveces, J., Gore, S., García, A., Martín, M. J., Launay, G., Alcántara, R., Del-Toro, N. et al. (2013), 'Biojs: an open source javascript framework for biological data visualization', *Bioinformatics* **29**(8), 1103–1104.

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.-H., Aravind, L., He, C. & Shi, Y. (2015), 'Dna methylation on n6-adenine in c. elegans', *Cell* **161**(4), 868–878.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. & Smola, A. J. (2007), A kernel method for the two-sample-problem, *in* 'Advances in neural information processing systems', pp. 513–520.

Griffith, J. S. & Mahler, H. R. (1969), 'Dna ticketing theory of memory', *Nature* **223**(5206), 580–582.

Grosjean, H., Benne, R. et al. (1998), *Modification and Editing of RNA*, ASM press Washington, DC.

Grosjean, H. & Grosjean, H. (2005), *Fine-tuning of RNA functions by modification and editing*, Vol. 12, Springer.

Group, B. D. W., Atkinson Jr, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T. et al. (2001), 'Biomarkers and surrogate endpoints: preferred definitions and conceptual framework', *Clinical pharmacology & therapeutics* **69**(3), 89–95.

Han, J., Pei, J. & Kamber, M. (2011), *Data mining: concepts and techniques*, Elsevier.

Hardcastle, T. J. & Kelly, K. A. (2010), 'bayseq: empirical bayesian methods for identifying differential expression in sequence count data', *BMC bioinformatics* **11**(1), 1–14.

Hasan, M. M., Manavalan, B., Khatun, M. S. & Kurata, H. (2019), 'i4mc-rose, a bioinformatics tool for the identification of dna n4-methylcytosine sites in the rosaceae genome', *International journal of biological macromolecules* .

Hasan, M. M., Manavalan, B., Shoombuatong, W., Khatun, M. S. & Kurata, H. (2020), 'i4mc-mouse: Improved identification of dna n4-methylcytosine sites in the mouse genome using multiple encoding schemes', *Computational and Structural Biotechnology Journal* .

He, H. & Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284.

He, W., Jia, C. & Zou, Q. (2019), '4mcpred: machine learning methods for dna n4-methylcytosine sites prediction', *Bioinformatics* **35**(4), 593–601.

Helm, M. (2006), 'Post-transcriptional nucleotide modification and alternative folding of rna', *Nucleic acids research* **34**(2), 721–733.

Heyn, H. & Esteller, M. (2015), 'An adenine code for dna: a second life for n6-methyladenine', *Cell* **161**(4), 710–713.

Holzinger, A. & Jurisica, I. (2014), Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions, *in* 'Interactive knowledge discovery and data mining in biomedical informatics', Springer, pp. 1–18.

Hotchkiss, R. D. (1948), 'The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography', *Journal of Biological Chemistry* **175**(1), 315–332.

Hu, J., He, X., Yu, D.-J., Yang, X.-B., Yang, J.-Y. & Shen, H.-B. (2014), 'A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction', *PloS One* **9**(9), e107676.

Huang, W., Xiong, J., Yang, Y., Liu, S.-M., Yuan, B.-F. & Feng, Y.-Q. (2015), 'Determination of dna adenine methylation in genomes of mammals and plants by liquid chromatography/mass spectrometry', *RSC Advances* **5**(79), 64046–64054.

Huang, Y., He, N., Chen, Y., Chen, Z. & Li, L. (2018), 'Bermp: a cross-species classifier for predicting m6a sites by integrating a deep learning algorithm and a random forest approach', *International journal of biological sciences* **14**(12), 1669.

Hussain, A., Khatri, M., Casali, G., Batchelor, T. & West, D. (2014), '194 follow up after lung cancer surgery: plain chest x ray does not increase diagnostic accuracy', *Lung Cancer* **83**, S72.

Indovina, P., Marcelli, E., Maranta, P. & Tarro, G. (2011), 'Lung cancer proteomics: recent advances in biomarker discovery', *International journal of proteomics* **2011**.

Isla, D., Sarries, C., Rosell, R., Alonso, G., Domine, M., Taron, M., Lopez-Vivanco, G., Camps, C., Botia, M., Nunez, L. et al. (2004), 'Single nucleotide polymorphisms and outcome in docetaxel–cisplatin-treated advanced non-small-cell lung cancer', *Annals of oncology* **15**(8), 1194–1203.

Jantus-Lewintre, E., Usó, M., Sanmartín, E. & Camps, C. (2012), 'Update on biomarkers for the detection of lung cancer', *Lung Cancer: Targets and Therapy* **3**, 21.

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C. & Thun, M. J. (2006), 'Cancer statistics, 2006', *CA: a cancer journal for clinicians* **56**(2), 106–130.

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y.-G. et al. (2011), 'N6-methyladenosine in nuclear rna is a major substrate of the obesity-associated fto', *Nature chemical biology* **7**(12), 885.

Jiang, X., Zhao, Y., Wang, X., Malin, B., Wang, S., Ohno-Machado, L. & Tang, H. (2014), 'A community assessment of privacy preserving techniques for human genomes', *BMC medical informatics and decision making* **14**(S1), S1.

Jin, B., Li, Y. & Robertson, K. D. (2011), 'Dna methylation: superior or subordinate in the epigenetic hierarchy?', *Genes & cancer* **2**(6), 607–617.

Jones, P. A. (2012), 'Functions of dna methylation: islands, start sites, gene bodies and beyond', *Nature Reviews Genetics* **13**(7), 484–492.

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P. V., Snaedal, J., Bjornsson, S., Huttenlocher, J., Levey, A. I., Lah, J. J. et al. (2013), 'Variant of trem2 associated with the risk of alzheimer's disease', *New England Journal of Medicine* **368**(2), 107–116.

Kang, S.-M., Sung, H.-J., Ahn, J.-M., Park, J.-Y., Lee, S.-Y., Park, C.-S. & Cho, J.-Y. (2011), 'The haptoglobin β chain as a supportive biomarker for human lung cancers', *Molecular BioSystems* **7**(4), 1167–1175.

Kantardzic, M. (2011), *Data mining: concepts, models, methods, and algorithms*, John Wiley & Sons.

Kanwal, R. & Gupta, S. (2012), 'Epigenetic modifications in cancer', *Clinical genetics* **81**(4), 303–311.

Katoh, K. & Standley, D. M. (2013), 'Mafft multiple sequence alignment software version 7: improvements in performance and usability', *Molecular biology and evolution* **30**(4), 772–780.

Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M. J., Park, C. Y. et al. (2015), 'A majority of m6a residues are in the last exons, allowing the potential for 3'utr regulation', *Genes & Development* **29**(19), 2037–2053.

Keith, G. (1995), 'Mobilities of modified ribonucleotides on two-dimensional cellulose thin-layer chromatography', *Biochimie* **77**(1-2), 142–144.

Klose, R. J., Kallin, E. M. & Zhang, Y. (2006), 'Jmjc-domain-containing proteins and histone demethylation', *Nature reviews genetics* **7**(9), 715.

Kohli, R. M. & Zhang, Y. (2013), 'Tet enzymes, tdg and the dynamics of dna demethylation', *Nature* **502**(7472), 472.

Korlach, J. & Turner, S. W. (2012), 'Going beyond five bases in dna sequencing', *Current opinion in structural biology* **22**(3), 251–261.

Krug, R. M., Morgan, M. A. & Shatkin, A. J. (1976), 'Influenza viral mrna contains internal n6-methyladenosine and 5'-terminal 7-methylguanosine in cap structures.', *Journal of virology* **20**(1), 45–53.

Kvam, V. M., Liu, P. & Si, Y. (2012), 'A comparison of statistical methods for detecting differentially expressed genes from rna-seq data', *American journal of botany* **99**(2), 248–256.

Leonhardt, H., Page, A. W., Weier, H.-U. & Bestor, T. H. (1992), 'A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei', *Cell* (5), 865–873.

Li, G.-Q., Liu, Z., Shen, H.-B. & Yu, D.-J. (2016), 'Targetm6a: Identifying n6-methyladenosine sites from rna sequences via position-specific nucleotide propensities and a support vector machine', *IEEE transactions on Nanobioscience* **15**(7), 674–682.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. (2017), 'Feature selection: A data perspective', *ACM Computing Surveys (CSUR)* **50**(6), 1–45.

Lichinchi, G., Gao, S., Saletore, Y., Gonzalez, G. M., Bansal, V., Wang, Y., Mason, C. E. & Rana, T. M. (2016), 'Dynamics of the human and viral m6a rna methylomes during hiv-1 infection of t cells', *Nature Microbiology* **1**, 16011.

Lichinchi, G., Zhao, B. S., Wu, Y., Lu, Z., Qin, Y., He, C. & Rana, T. M. (2016), 'Dynamics of human and viral rna methylation during zika virus infection', *Cell Host & Microbe* **20**(5), 666–673.

Lin, C., Chen, W., Qiu, C., Wu, Y., Krishnan, S. & Zou, Q. (2014), 'Libd3c: ensemble classifiers with a clustering and dynamic selection strategy', *Neurocomputing* **123**, 424–435.

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E. & Jaffrey, S. R. (2015), 'Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome', *Nature Methods* **12**(8), 767–772.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. & Ecker, J. R. (2008), 'Highly integrated single-base resolution maps of the epigenome in arabidopsis', *Cell* **133**(3), 523–536.

Liu, B., Liu, F., Fang, L., Wang, X. & Chou, K.-C. (2015), 'repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects', *Bioinformatics* **31**(8), 1307–1309.

Liu, G.-H., Shen, H.-B. & Yu, D.-J. (2016), 'Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures', *The Journal of Membrane Biology* **249**(1-2), 141–153.

Liu, H., Flores, M. A., Meng, J., Zhang, L., Zhao, X., Rao, M. K., Chen, Y. & Huang, Y. (2014), 'Met-db: a database of transcriptome methylation in mammalian cells', *Nucleic Acids Research* **43**(D1), D197–D203.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X. et al. (2014), 'A mettl3-mettl14 complex mediates mammalian nuclear rna n6-adenosine methylation', *Nature Chemical Biology* **10**(2), 93–95.

Liu, L., Lei, X., Fang, Z., Tang, Y., Meng, J. & Wei, Z. (2020), 'Lithophone: improving lncrna methylation site prediction using an ensemble predictor', *Frontiers in genetics* **11**, 545.

Liu, L., Lei, X., Meng, J. & Wei, Z. (2020), 'Witmsg: Large-scale prediction of human intronic m6a rna methylation sites from sequence and genomic features', *Current Genomics* **21**(1), 67–76.

Liu, Z., Xiao, X., Yu, D.-J., Jia, J., Qiu, W.-R. & Chou, K.-C. (2016), 'prnam-pc: Predicting n6-methyladenosine sites in rna sequences via physical–chemical properties', *Analytical Biochemistry* **497**, 60–67.

Lock, C., Hermans, G., Pedotti, R., Brendolan, A., Schadt, E., Garren, H., Langer-Gould, A., Strober, S., Cannella, B., Allard, J. et al. (2002), 'Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis', *Nature medicine* **8**(5), 500–508.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. et al. (2013), 'The genotype-tissue expression (gtex) project', *Nature genetics* **45**(6), 580.

Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I. & Xiong, M. (2010), 'Genome-wide gene and pathway analysis', *European Journal of Human Genetics* **18**(9), 1045–1053.

Lv, Z., Wang, D., Ding, H., Zhong, B. & Xu, L. (2020), 'Escherichia coli dna n-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology', *IEEE Access* **8**, 14851–14859.

Ma, Q. & Lu, A. Y. (2011), 'Pharmacogenetics, pharmacogenomics, and individualized medicine', *Pharmacological reviews* **63**(2), 437–459.

Machnicka, M. A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K. M. et al. (2012), 'Modomics: a database of rna modification pathways—2013 update', *Nucleic Acids Research* **41**(D1), D262–D267.

Maciel, C. M., Junqueira, M., Paschoal, M. E. M., Kawamura, M. T., Duarte, R. L. M., Carvalho, M. d. G. d. C. & Domont, G. B. (2005), 'Differential proteomic serum pattern of low molecular weight proteins expressed

by adenocarcinoma lung cancer patients.', *Journal of experimental therapeutics & oncology* **5**(1).

Madden, T. (2013), The blast sequence analysis tool, *in* 'The NCBI Handbook [Internet]. 2nd edition', National Center for Biotechnology Information (US).

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., Lee, G. et al. (2019), '4mcpred-el: an ensemble learning framework for identification of dna n4-methylcytosine sites in the mouse genome', *Cells* **8**(11), 1332.

Manavalan, B., Basith, S., Shin, T. H., Wei, L. & Lee, G. (2019), 'Meta-4mcpred: a sequence-based meta-predictor for accurate dna 4mc site prediction using effective feature representation', *Molecular Therapy-Nucleic Acids* **16**, 733–744.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. (2008), 'Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays', *Genome research* **18**(9), 1509–1517.

Marshall, H. M., Bowman, R. V., Yang, I. A., Fong, K. M. & Berg, C. D. (2013), 'Screening for lung cancer with low-dose computed tomography: a review of current status', *Journal of thoracic disease* **5**(Suppl 5), S524.

McCarthy, J. J. & Hilfiker, R. (2000), 'The use of single-nucleotide polymorphism maps in pharmacogenomics', *Nature biotechnology* **18**(5), 505–508.

McGuffin, L. J., Bryson, K. & Jones, D. T. (2000), 'The psipred protein structure prediction server', *Bioinformatics* **16**(4), 404–405.

Messer, W. & Noyer-Weidner, M. (1988), 'Timing and targeting: the biological functions of dam methylation in e. coli', *Cell* **54**(6), 735–737.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E. & Jaffrey, S. R. (2012), 'Comprehensive analysis of mrna methylation

reveals enrichment in 3'utrs and near stop codons', *Cell* **149**(7), 1635–1646.

Minna, J. D., Roth, J. A. & Gazdar, A. F. (2002), 'Focus on lung cancer', *Cancer cell* **1**(1), 49–52.

Mitas, M., Hoover, L., Silvestri, G., Reed, C., Green, M., Turrisi, A. T., Sherman, C., Mikhitarian, K., Cole, D. J., Block, M. I. et al. (2003), 'Lunx is a superior molecular marker for detection of non-small lung cell cancer in peripheral blood', *The Journal of molecular diagnostics* **5**(4), 237–242.

Mizuguchi, S., Nishiyama, N., Iwata, T., Nishida, T., Izumi, N., Tsukioka, T., Inoue, K., Uenishi, T., Wakasa, K. & Suehiro, S. (2007), 'Serum sialyl lewisx and cytokeratin 19 fragment as predictive factors for recurrence in patients with stage i non-small cell lung cancer', *Lung cancer* **58**(3), 369–375.

Modrich, P. (1991), 'Mechanisms and biological effects of mismatch repair', *Annual review of genetics* **25**(1), 229–253.

Montani, F., Marzi, M. J., Dezi, F., Dama, E., Carletti, R. M., Bonizzi, G., Bertolotti, R., Bellomi, M., Rampinelli, C., Maisonneuve, P. et al. (2015), 'mir-test: a blood test for lung cancer early detection', *JNCI: Journal of the National Cancer Institute* **107**(6).

Moody, T. W., Murphy, A., Mahmoud, S. & Fiskum, G. (1987), 'Bombesin-like peptides elevate cytosolic calcium in small cell lung cancer cells', *Biochemical and biophysical research communications* **147**(1), 189–195.

Moody, T. W., Staley, J., Zia, F., Coy, D. H. & Jensen, R. T. (1992), 'Neuromedin b binds with high affinity, elevates cytosolic calcium and stimulates the growth of small-cell lung cancer cell lines.', *Journal of Pharmacology and Experimental Therapeutics* **263**(1), 311–317.

Morris, J. et al. (2001), 'Genes, genetics, and epigenetics: a correspondence', *Science* **293**(5532), 1103–1105.

Motorin, Y. & Helm, M. (2011), 'Rna nucleotide methylation', *Wiley Interdisciplinary Reviews: RNA* **2**(5), 611–631.

Mount, D. W. (2007), 'Using the basic local alignment search tool (blast)', *Cold Spring Harbor Protocols* **2007**(7), pdb–top17.

Nagrath, S., Sequist, L. V., Maheswaran, S., Bell, D. W., Irimia, D., Ulkus, L., Smith, M. R., Kwak, E. L., Digumarthy, S., Muzikansky, A. et al. (2007), 'Isolation of rare circulating tumour cells in cancer patients by microchip technology', *Nature* **450**(7173), 1235.

Noé, L. & Kucherov, G. (2005), 'Yass: enhancing the sensitivity of dna similarity search', *Nucleic acids research* **33**(suppl_2), W540–W543.

Okada, M., Nishio, W., Sakamoto, T., Uchino, K., Yuki, T., Nakagawa, A. & Tsubota, N. (2004), 'Effect of histologic type and smoking status on interpretation of serum carcinoembryonic antigen value in non–small cell lung carcinoma', *The Annals of thoracic surgery* **78**(3), 1004–1009.

Parkin, D. M. (2001), 'Global cancer statistics in the year 2000', *The lancet oncology* **2**(9), 533–543.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), 'Scikit-learn: Machine learning in python', *Journal of machine learning research* **12**(Oct), 2825–2830.

Peng, H., Long, F. & Ding, C. (2005), 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(8), 1226–1238.

Peters, K. E., Walters, C. C. & Moldowan, J. (2005), *The biomarker guide*, Vol. 1, Cambridge University Press.

Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y.-S. et al. (2014), 'Mammalian wtap is a regulatory subunit of the rna n6-methyladenosine methyltransferase', *Cell Research* **24**(2), 177.

Portela, A. & Esteller, M. (2010), 'Epigenetic modifications and human disease', *Nature biotechnology* **28**(10), 1057.

Pujol, J.-L., Grenier, J., Daurès, J.-P., Daver, A., Pujol, H. & Michel, F.-B. (1993), 'Serum fragment of cytokeratin subunit 19 measured by cyfra 21-1 immunoradiometric assay as a marker of lung cancer', *Cancer research* **53**(1), 61–66.

Qiang, X., Chen, H., Ye, X., Su, R. & Wei, L. (2018), 'M6amrfs: robust prediction of n6-methyladenosine sites with sequence-based features in multiple species', *Frontiers in genetics* **9**, 495.

Quackenbush, J. (2006), 'Microarray analysis and tumor classification', *New England Journal of Medicine* **354**(23), 2463–2472.

Rabinowits, G., Gerçel-Taylor, C., Day, J. M., Taylor, D. D. & Kloecker, G. H. (2009), 'Exosomal microrna: a diagnostic marker for lung cancer', *Clinical lung cancer* **10**(1), 42–46.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P. & Jaenisch, R. (2000), 'Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a', *Proceedings of the National Academy of Sciences* **97**(10), 5237–5242.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. & Betel, D. (2013), 'Comprehensive evaluation of differential gene expression analysis methods for rna-seq data', *Genome biology* **14**(9), 3158.

Ratel, D., Ravanat, J.-L., Berger, F. & Wion, D. (2006), 'N6-methyladenine: the other methylated base of dna', *Bioessays* **28**(3), 309–315.

Rathi, P., Maurer, S. & Summerer, D. (2018), 'Selective recognition of n 4-methylcytosine in dna by engineered transcription-activator-like effectors', *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**(1748), 20170078.

Raza, K. (2012), 'Application of data mining in bioinformatics', *arXiv preprint arXiv:1205.1125* .

Riquelme Barrios, S. A., Pereira-Montecinos, C., Valiente-Echeverría, F. & Soto-Rifo, R. (2018), 'Emerging roles of n6-methyladenosine on hiv-1 rna metabolism and viral replication', *Frontiers in Microbiology* **9**, 576.

Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. (2015), 'Rebase—a database for dna restriction and modification: enzymes, genes and genomes', *Nucleic acids research* **43**(D1), D298–D299.

Robertson, K. D. (2002), 'Dna methylation and chromatin–unraveling the tangled web', *Oncogene* **21**(35), 5361–5379.

Robertson, K. D. (2005), 'Dna methylation and human disease', *Nature Reviews Genetics* **6**(8), 597–610.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2010), 'edger: a bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics* **26**(1), 139–140.

Rodríguez-Paredes, M. & Esteller, M. (2011), 'Cancer epigenetics reaches mainstream oncology', *Nature medicine* **17**(3), 330.

Saeys, Y., Inza, I. & Larrañaga, P. (2007), 'A review of feature selection techniques in bioinformatics', *bioinformatics* **23**(19), 2507–2517.

Sauna, Z. E. & Kimchi-Sarfaty, C. (2011), 'Understanding the contribution of synonymous mutations to human disease', *Nature Reviews Genetics* **12**(10), 683–691.

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M. & Lyko, F. (2010), 'Rna methylation by dnmt2 protects transfer rnas against stress-induced cleavage', *Genes & development* **24**(15), 1590–1595.

Schibler, U., Kelley, D. E. & Perry, R. P. (1977), 'Comparison of methylated sequences in messenger rna and heterogeneous nuclear rna from mouse l cells', *Journal of molecular biology* **115**(4), 695–714.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E. & Zipursky, S. L. (2000), 'Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity', *Cell* **101**(6), 671–684.

Schnabel, P. & Junker, K. (2015), 'Pulmonary neuroendocrine tumors in the new who 2015 classification: Start of breaking new grounds?', *Der Pathologe* **36**(3), 283–292.

Schuller, H. M. (2010), 'Beta-adrenergic signaling, a novel target for cancer therapy?', *Oncotarget* **1**(7), 466.

Schuller, H. M. & Cekanova, M. (2005), 'Nnk-induced hamster lung adenocarcinomas over-express β2-adrenergic and egfr signaling pathways', *Lung Cancer* **49**(1), 35–45.

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T. S., Satija, R., Ruvkun, G. et al. (2013), 'High-resolution mapping reveals a conserved, widespread, dynamic mrna methylation program in yeast meiosis', *Cell* **155**(6), 1409–1421.

Seyednasrollah, F., Laiho, A. & Elo, L. L. (2013), 'Comparison of software packages for detecting differential expression in rna-seq studies', *Briefings in bioinformatics* **16**(1), 59–70.

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell system technical journal* **27**(3), 379–423.

Shen, F., Huang, W., Huang, J.-T., Xiong, J., Yang, Y., Wu, K., Jia, G.-F., Chen, J., Feng, Y.-Q., Yuan, B.-F. et al. (2015), 'Decreased n6-methyladenosine in peripheral blood rna from diabetic patients is associated with fto expression rather than alkbh5', *The Journal of Clinical Endocrinology & Metabolism* **100**(1), E148–E154.

Shi, Y. (2007), 'Histone lysine demethylases: emerging roles in development, physiology and disease', *Nature reviews genetics* **8**(11), 829–833.

Smyth, D. J., Plagnol, V., Walker, N. M., Cooper, J. D., Downes, K., Yang, J. H., Howson, J. M., Stevens, H., McManus, R., Wijmenga, C. et al. (2008), 'Shared and distinct genetic variants in type 1 diabetes and celiac disease', *New England Journal of Medicine* **359**(26), 2767–2777.

Soneson, C. & Delorenzi, M. (2013), 'A comparison of methods for differential expression analysis of rna-seq data', *BMC bioinformatics* **14**(1), 91.

Song, L., Li, D., Zeng, X., Wu, Y., Guo, L. & Zou, Q. (2014), 'ndna-prot: identification of dna-binding proteins based on unbalanced classification', *BMC bioinformatics* **15**(1), 298.

Sonnenburg, S., Henschel, S., Widmer, C., Behr, J., Zien, A., Bona, F. d., Binder, A., Gehl, C., Franc, V. et al. (2010), 'The shogun machine learning toolbox', *Journal of Machine Learning Research* **11**(Jun), 1799–1802.

Sozzi, G., Boeri, M., Rossi, M., Verri, C., Suatoni, P., Bravi, F., Roz, L., Conte, D., Grassi, M., Sverzellati, N. et al. (2014), 'Clinical utility of

a plasma-based mirna signature classifier within computed tomography lung cancer screening: a correlative mild trial study', *Journal of clinical oncology* **32**(8), 768.

Sozzi, G., Conte, D., Leon, M., Cirincione, R., Roz, L., Ratcliffe, C., Roz, E., Cirenei, N., Bellomi, M., Pelosi, G. et al. (2003), 'Quantification of free circulating dna as a diagnostic marker in lung cancer', *Journal of clinical oncology* **21**(21), 3902–3908.

Squires, J. E. & Preiss, T. (2010), 'Function and detection of 5-methylcytosine in eukaryotic rna', *Epigenomics* **2**(5), 709–715.

Stegle, O., Drewe, P., Bohnert, R., Borgwardt, K. & Rätsch, G. (2010), 'Statistical tests for detecting differential rna-transcript expression from read counts'.

Stoiber, M. H., Quick, J., Egan, R., Lee, J. E., Celniker, S. E., Neely, R., Loman, N., Pennacchio, L. & Brown, J. B. (2016), 'De novo identification of dna modifications enabled by genome-guided nanopore signal processing', *BioRxiv* p. 094672.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z. & Cedar, H. (2009), 'Developmental programming of cpg island methylation profiles in the human genome', *Nature structural & molecular biology* **16**(5), 564.

Sujansky, W. (2001), 'Heterogeneous database integration in biomedicine', *Journal of biomedical informatics* **34**(4), 285–298.

Sun, W.-J., Li, J.-H., Liu, S., Wu, J., Zhou, H., Qu, L.-H. & Yang, J.-H. (2015), 'Rmbase: a resource for decoding the landscape of rna modifications from high-throughput sequencing data', *Nucleic Acids Research* **44**(D1), D259–D265.

Sung, H.-J. & Cho, J.-Y. (2008), 'Biomarkers for the lung cancer diagnosis and their advances in proteomics', *BMB reports* **41**(9), 615–625.

Suzuki, M. M. & Bird, A. (2008), 'Dna methylation landscapes: provocative insights from epigenomics', *Nature Reviews Genetics* **9**(6), 465–476.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L. et al. (2009), 'Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1', *Science* **324**(5929), 930–935.

Tarailo-Graovac, M., Shyr, C., Ross, C. J., Horvath, G. A., Salvarinova, R., Ye, X. C., Zhang, L.-H., Bhavsar, A. P., Lee, J. J., Drögemöller, B. I. et al. (2016), 'Exome sequencing and the management of neurometabolic disorders', *New England Journal of Medicine* **374**(23), 2246–2255.

Tomczak, K., Czerwińska, P. & Wiznerowicz, M. (2015), 'The cancer genome atlas (tcga): an immeasurable source of knowledge', *Contemporary oncology* **19**(1A), A68.

Torkamani, A., Topol, E. J. & Schork, N. J. (2008), 'Pathway analysis of seven common diseases assessed by genome-wide association', *Genomics* **92**(5), 265–272.

Tsai, K., Courtney, D. G. & Cullen, B. R. (2018), 'Addition of m6a to sv40 late mrnas enhances viral structural gene expression and replication', *PLoS Pathogens* **14**(2), e1006919.

Valenti, R., Huber, V., Filipazzi, P., Pilla, L., Sovena, G., Villa, A., Corbelli, A., Fais, S., Parmiani, G. & Rivoltini, L. (2006), 'Human tumor-released microvesicles promote the differentiation of myeloid cells with transforming growth factor-β–mediated suppressive activity on t lymphocytes', *Cancer research* **66**(18), 9290–9298.

Van Ham, T. J., Thijssen, K. L., Breitling, R., Hofstra, R. M., Plasterk, R. H. & Nollen, E. A. (2008), 'C. elegans model identifies genetic modifiers of α-synuclein inclusion formation during aging', *PLoS genetics* **4**(3).

Vargas, A. J. & Harris, C. C. (2016), 'Biomarker development in the precision medicine era: lung cancer as a case study', *Nature Reviews Cancer* **16**(8), 525.

Vazquez, M. F., Koizumi, J. H., Henschke, C. I. & Yankelevitz, D. F. (2007), 'Reliability of cytologic diagnosis of early lung cancer', *Cancer Cytopathology: Interdisciplinary International Journal of the American Cancer Society* **111**(4), 252–258.

Vegas, E., Oller, J. M. & Reverter, F. (2016), 'Inferring differentially expressed pathways using kernel maximum mean discrepancy-based test', *BMC bioinformatics* **17**(5), 205.

Waddington, C. H. et al. (1939), 'An introduction to modern genetics.', *An introduction to modern genetics.* .

Wan, S., Duan, Y. & Zou, Q. (2017), 'Hpslpred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source', *Proteomics* .

Wan, Y., Tang, K., Zhang, D., Xie, S., Zhu, X., Wang, Z. & Lang, Z. (2015), 'Transcriptome-wide high-throughput deep m6a-seq reveals unique differential m6a methylation patterns between three organs in arabidopsis thaliana', *Genome Biology* **16**(1), 272.

Wang, X. & Yan, R. (2018), 'Rfathm6a: a new tool for predicting m6a sites in arabidopsis thaliana', *Plant Molecular Biology* pp. 1–11.

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H. & He, C. (2015), 'N6-methyladenosine modulates messenger rna translation efficiency', *Cell* **161**(6), 1388–1399.

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z. & Zhao, J. C. (2014), 'N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells', *Nature Cell Biology* **16**(2), 191.

Wang, Z., Gerstein, M. & Snyder, M. (2009), 'Rna-seq: a revolutionary tool for transcriptomics', *Nature reviews genetics* **10**(1), 57.

Wei, H.-L. & Billings, S. A. (2006), 'Feature subset selection and ranking for data dimensionality reduction', *IEEE transactions on pattern analysis and machine intelligence* **29**(1), 162–166.

Wei, L., Luan, S., Nagai, L. A. E., Su, R. & Zou, Q. (2019), 'Exploring sequence-based features for the improved prediction of dna n4-methylcytosine sites in multiple species', *Bioinformatics* **35**(8), 1326–1333.

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q. & Shi, X. (2019), 'Iterative feature representations improve n4-methylcytosine site prediction', *Bioinformatics* **35**(23), 4930–4937.

Wei, Z.-S., Han, K., Yang, J.-Y., Shen, H.-B. & Yu, D.-J. (2016), 'Protein–protein interaction sites prediction by ensembling svm and sample-weighted random forests', *Neurocomputing* **193**, 201–212.

Wong, J. C., Bathina, M. & Fiscus, R. R. (2012), 'Cyclic gmp/protein kinase g type-iα (pkg-iα) signaling pathway promotes creb phosphorylation and maintains higher c-iap1, livin, survivin, and mcl-1 expression and the inhibition of pkg-iα kinase activity synergizes with cisplatin in non-small cell lung cancer cells', *Journal of cellular biochemistry* **113**(11), 3587–3598.

Woodcock, D., Crowther, P. & Diver, W. (1987), 'The majority of methylated deoxycytidines in human dna are not in the cpg dinucleotide', *Biochemical and biophysical research communications* **145**(2), 888–894.

Wu, R., Jiang, D., Wang, Y. & Wang, X. (2016), 'N6-methyladenosine (m6a) methylation in mrna with a dynamic and reversible epigenetic modification', *Molecular Biotechnology* **58**(7), 450–459.

Xiang, S., Yan, Z., Liu, K., Zhang, Y. & Sun, Z. (2016), 'Athmethpre: a web server for the prediction and query of mrna m 6 a sites in arabidopsis thaliana', *Molecular BioSystems* **12**(11), 3333–3337.

Xing, P., Su, R., Guo, F. & Wei, L. (2017), 'Identifying n6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine', *Scientific Reports* **7**.

Yan, D., Wu, Z., Chisholm, A. D. & Jin, Y. (2009), 'The dlk-1 kinase promotes mrna stability and local translation in c. elegans synapses and axon regeneration', *Cell* **138**(5), 1005–1018.

Yang, Y., Huang, W., Huang, J.-T., Shen, F., Xiong, J., Yuan, E.-F., Qin, S.-s., Zhang, M., Feng, Y.-Q., Yuan, B.-F. et al. (2016), 'Increased n6-methyladenosine in human sperm rna as a risk factor for asthenozoospermia', *Scientific Reports* **6**, 24345.

Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C. & Xie, Z. (2016), 'Methsmrt: an integrative database for dna n6-methyladenine and n4-methylcytosine generated by single-molecular real-time sequencing', *Nucleic acids research* p. gkw950.

Yu, D.-J., Hu, J., Huang, Y., Shen, H.-B., Qi, Y., Tang, Z.-M. & Yang, J.-Y. (2013), 'Targetatpsite: a template-free method for atp-binding sites prediction with residue evolution image sparse representation and classifier ensemble', *Journal of Computational Chemistry* **34**(11), 974–985.

Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. (2012), 'clusterprofiler: an r package for comparing biological themes among gene clusters', *Omics: a journal of integrative biology* **16**(5), 284–287.

Yu, M., Ji, L., Neumann, D. A., Chung, D.-h., Groom, J., Westpheling, J., He, C. & Schmitz, R. J. (2015), 'Base-resolution detection of

n 4-methylcytosine in genomic dna using 4mc-tet-assisted-bisulfite-sequencing', *Nucleic acids research* **43**(21), e148–e148.

Zamay, T., Zamay, G., Kolovskaya, O., Zukov, R., Petrova, M., Gargaun, A., Berezovski, M. & Kichkailo, A. (2017), 'Current and prospective protein biomarkers of lung cancer', *Cancers* **9**(11), 155.

Zhang, C., Samanta, D., Lu, H., Bullen, J. W., Zhang, H., Chen, I., He, X. & Semenza, G. L. (2016), 'Hypoxia induces the breast cancer stem cell phenotype by hif-dependent and alkbh5-mediated m6a-demethylation of nanog mrna', *Proceedings of the National Academy of Sciences* **113**(14), E2047–E2056.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J. et al. (2015), 'N6-methyladenine dna modification in drosophila', *Cell* **161**(4), 893–906.

Zhang, M., Sun, J.-W., Liu, Z., Ren, M.-W., Shen, H.-B. & Yu, D.-J. (2016), 'Improving n 6-methyladenosine site prediction with heuristic selection of nucleotide physical–chemical properties', *Analytical Biochemistry* **508**, 104–113.

Zhang, Y. & Hamada, M. (2018), 'Deepm6aseq: prediction and characterization of m6a-containing sequences using deep learning', *BMC bioinformatics* **19**(19), 524.

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C.-M., Li, C. J., Vgb, C. B., Shi, Y., Wang, W.-L., Song, S.-H. et al. (2013), 'Alkbh5 is a mammalian rna demethylase that impacts rna metabolism and mouse fertility', *Molecular Cell* **49**(1), 18–29.

Zheng, Y., Nie, P., Peng, D., He, Z., Liu, M., Xie, Y., Miao, Y., Zuo, Z. & Ren, J. (2017), 'm6avar: a database of functional variants involved in m6a modification', *Nucleic Acids Research* .

Zhou, Y., Zeng, P., Li, Y.-H., Zhang, Z. & Cui, Q. (2016), 'Sramp: prediction of mammalian n6-methyladenosine (m6a) sites based on sequence-derived features', *Nucleic Acids Research* **44**(10), e91–e91.

Zou, Q., Wan, S., Ju, Y., Tang, J. & Zeng, X. (2016), 'Pretata: predicting tata binding proteins with novel features and dimensionality reduction strategy', *BMC systems biology* **10**(4), 114.

Zou, Q., Zeng, J., Cao, L. & Ji, R. (2016), 'A novel features ranking metric with application to scalable visual and bioinformatics data classification', *Neurocomputing* **173**, 346–354.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K. B. (2007), 'Frontiers of biomedical text mining: current progress', *Briefings in bioinformatics* **8**(5), 358–375.