2021

NOVEL COMPUTATIONAL METHODS FOR CANCER GENOMICS DATA ANALYSIS

Jinpeng Liu
University of Kentucky, merckey@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2021.215

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Jinpeng Liu, Student

Dr. Zongming Fei, Major Professor

Dr. Zongming Fei, Director of Graduate Studies

NOVEL COMPUTATIONAL METHODS FOR CANCER GENOMICS DATA ANALYSIS

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Engineering at the University of Kentucky

By

Jinpeng Liu

Lexington, Kentucky

Co- Directors: Dr. Zongming Fei, Professor of Computer Science

and Dr. Licong Cui, Assistant Professor of Computer Science

Lexington, Kentucky

2021

ABSTRACT OF DISSERTATION

NOVEL COMPUTATIONAL METHODS
FOR CANCER GENOMICS DATA ANALYSIS

Cancer is a genetic disease responsible for one in eight deaths worldwide. The advancement of next-generation sequencing (NGS) technology has revolutionized the cancer research, allowing comprehensively profiling the cancer genome at great resolution. Large-scale cancer genomics research has sparked the needs for efficient and accurate Bioinformatics methods to analyze the data. The research presented in this dissertation focuses on three areas in cancer genomics: cancer somatic mutation detection; cancer driver genes identification and transcriptome profiling on single-cell level.

NGS data analysis involves a series of complicated data transformation that convert raw sequencing data to the information that is interpretable by cancer researchers. The first project in the dissertation established a robust, reproducible and scalable cancer genomics data analysis workflow management system that automates the best practice mutation calling pipelines to detect somatic single nucleotide polymorphisms, insertion, deletion and copy number variation from NGS data. It integrates mutation annotation, clinically actionable therapy prediction and data visualization that streamlines the sequence-to-report data transformation.

In order to differentiate the driver mutations buried among a vast pool of passenger mutations from a somatic mutation calling project, we developed MEScan in the second project, a novel method that enables genome-scale driver mutations identification based on mutual exclusivity test using cancer somatic mutation data. MEScan implements an efficient statistical framework to *de novo* screen mutual exclusive patterns and in the meantime taking into account the patient-specific and gene-specific background mutation rate and adjusting the heterogenous mutation frequency. It outperforms several existing methods based on simulation studies and real-world datasets. Genome-wide screening using existing TCGA somatic mutation data discovers novel cancer-specific and pan-cancer mutually exclusive patterns.

Bulk RNA sequencing (RNA-Seq) has become one of the most commonly used techniques for transcriptome profiling in a wide spectrum of biomedical and biological research. Analyzing bulk RNA-Seq reads to quantify expression at each gene locus is the first step towards the identification of differentially expressed genes for downstream biological interpretation. Recent advances in single-cell RNA-seq (scRNA-seq)

technology allows cancer biologists to profile gene expression on higher resolution cellular level. Preprocessing scRNA-seq data to quantify UMI-based gene count is the key to characterize intra-tumor cellular heterogeneity and identify rare cells that governs tumor progression, metastasis and treatment resistance. Despite its popularity, summarizing gene count from raw sequencing reads remains the one of the most time-consuming steps with existing tools. Current pipelines do not balance the efficiency and accuracy in large-scale gene count summarization in both bulk and scRNA-seq experiments. In the third project, we developed a light-weight *k*-mer based gene counting algorithm, FastCount, to accurately and efficiently quantify gene-level abundance using bulk RNA-seq or UMI-based scRNA-seq data. It achieves at least an order-of-magnitude speed improvement over the current gold standard pipelines while providing competitive accuracy.

KEYWORDS: Cancer Genomics, Pipeline Framework; Somatic Mutation, Driver Mutations, Single Cell RNA-seq Quantification

Jinpeng Liu

(Name of Student)

06/06/2021

Date

NOVEL COMPUTATIONAL METHODS
FOR CANCER GENOMICS DATA ANALYSIS

By
Jinpeng Liu

Zongming Fei
Co-Director of Dissertation

Licong Cui
Co-Director of Dissertation

Zongming Fei
Director of Graduate Studies

06/06/2021
Date

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to Dr. Zongming Fei and Dr. Licong Cui for their continuous support and encouragement throughout my Ph.D. study. They are not only great researchers with immense knowledge, but also excellent advisors that drive students to the right direction. Without their patient guidance, I would not have been able to finish this dissertation.

I would also like to express my deepest appreciation to Dr. Chi Wang. He not only taught me a lot of knowledge in statistics, but also helped me learn to collaborate with researchers of different backgrounds. I am also extremely grateful to Dr. Dakshnamoorthy Manivannan for his invaluable advice and directions on my dissertation.

I am also extremely grateful to have Dr. Daniela Moga as my Ph.D. Outside Examiner. I wish to thank her for the time and considerations on my Ph.D. defense.

I would like to thank my collaborators from Markey Cancer Center for the valuable discussion we had during the past years. Especially, I would like to thank Dr. Susanne Arnold, Dr. Kathleen O' Connor, Dr. Jill Kolesar, Dr. Chunming Liu and Dr. Hunter Moseley for their suggestion and input in my research.

I wish to thank Dr. Heidi Weiss and all the team members in Biostatistcs and Bioinformatics Shared Resource Facility in Markey Cancer Center. Thank you for their support and constructive suggestions on my Ph.D. study.

I would also like to thank Dr. Jinze Liu and the members of her lab. I cherish the opportunity to work with my fellows and establish a friendship with them: Dr. Xinan Liu, Dr. Ye Yu, Dr. Yi Zhang, Eamonn Manager and Xiaofei Zhang.

My Ph.D. would not have been possible without my family and friends' support and encouragement. I am very grateful to my parents, my wife, and my kids for their love and for providing me a relaxed and supportive environment.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER 1.INTRODUCTION 1

1.1 Cancer Biology 1
1.1.1 Cancer Genome 1
1.1.2 Tumor heterogeneity 2

1.2 Next-generation sequencing 3
1.2.1 DNA Sequencing 4
1.2.2 RNA Sequencing 5

1.3 Current Computational Methods for Cancer Genomics Data Analysis 7
1.3.1 Somatic mutations analysis 7
1.3.2 Driver mutations 9
1.3.3 RNA-seq Analysis 10

1.4 Contribution 12

CHAPTER 2. .....Bioinformatics Framework for Characterization of Squamous Cell Lung Cancers from Appalachian Kentucky 15

2.1 Introduction 15

2.2 Pipeline management 19

2.3 Runtime environment deployment 20

2.4 Best practice workflows 21

2.5 Results 22
2.5.1 Overview of somatic alterations 23
2.5.2 Significantly mutated genes 24
2.5.3 Copy number variation analysis 25
2.5.4 Comparative mutational analysis with other cohorts 26
2.5.5 Clinically actionable mutations assessment 28
2.5.6 Prediction of the effect of IDH1 mutations 30
2.5.7 Localization of PCMTD1 mutations 39

2.6 Conclusion 41

CHAPTER 3. ....MESCAN: A Powerful Statistical Framework for Genome-Scale Mutual Exclusivity Analysis of Cancer Mutations 43

3.1 Introduction 43

3.2 MEScan Framework 47
3.3 Testing mutual exclusivity of a single gene set 50
3.4 Genome-wide screening 53
3.5 Determining a cutoff value to control the FDR 54
3.6 Identifying high-confidence mutually exclusive gene sets 55
3.7 Results 56
3.7.1 Simulation studies 56
3.7.2 Time cost comparison 68
3.7.3 Choosing cutoff values of TG to control FDR 69
3.7.4 Whole genome data analysis 70
3.7.5 Real world validation and comparison 76
3.8 Discussion 77

CHAPTER 4. FastCount: A Fast Gene Count Software for Single Cell and Bulk RNA-seq Data 80

4.1 Introduction 80
4.1.1 Bulk RNA-seq gene quantification 80
4.1.2 Single-cell RNA-seq gene quantification 82
4.2 FastCount algorithm 86
4.2.1 Gene k-mers signatures 87
4.2.2 GeneOthello k-mers index 89
4.2.3 Read assignment to genes 91
4.2.4 FastCount scRNA-seq implementation 94
4.3 Experimental results 98
4.3.1 Bulk RNA-seq simulation datasets 98
4.3.2 scRNA-seq 10X Genomics datasets 99
4.3.3 Comparison with other bulk RNA-seq tools 100
4.3.4 Comparison with other scRNA-seq pipelines 102
4.3.5 Runtime comparison 107
4.4 Conclusion 108

CHAPTER 5 Conclusion 110

REFERENCES 113

VITA 124

LIST OF TABLES

Table 1. Somatic alteration rate comparison between AppKY and TCGA of Lung SQCC. 27
Table 2. Clinically actionable mutations identified for APPKY patients. 29
Table 3. PCMTD1 mutations. The PCMTD1 mutations reported in the literature are in the C-terminal SOCS Box. 40
Table 4. Comparison of computational time. The reported computational time (in seconds) was for analyzing 1000 gene sets of a given size. 68
Table 5 Accuracy of gene count quantification in terms of Pearson and Spearman correlation, MRD and 5% EF using simulated data. 101
Table 6 UMI count concordance between different methods and Cell Ranger in terms of median Pearson, Spearman and MRD for the 6 datasets. 103
Table 7 Runtime comparison on 10X Genomics single cell gene expression datasets in different tools 108

LIST OF FIGURES

Figure 1 Whole exome sequencing data analysis pipeline for somatic mutation calling and copy number variation detection. 22

Figure 2. Significantly mutated genes in lung SQCC. Significantly mutated genes (FDR<0.2) from whole-exome sequencing of 51 samples from Appalachian Kentucky patients. 24

Figure 3 GISTIC amplification (left) and deletion (right) plots of the G-scores (shown at the top of the figure) and q-values (shown at the bottom of the figure) across the entire region analyzed. 25

Figure 4. IDH1 and PCMTD1 mutations. (A) IDH1 mutations and their mutation frequencies (circles). (B) PCMTD1 mutations and their frequencies (circles). 30

Figure 5. Functional analysis of *IDH1* variants. 32

Figure 6. Maximum likelihood phylogenetic tree of *IDH1* and *IDH2* proteins from a representative genome set. 35

Figure 7. Multiple sequence alignment of *IDH1* and *IDH2* proteins from a representative genome set. 36

Figure 8. *IDH1* mutations and *IDH1* associated pathway analysis. 37

Figure 9. Overview of the MEScan framework. 49

Figure 10. Comparison of power for identifying a true mutually exclusive gene set based on simulations. 58

Figure 11. Comparison of power for identifying subsets of a true mutually exclusive gene set based on simulations. 61

Figure 12 Simulation results for applying MEScan, MEGSA, Dendrix, WExT and CoMEt across different sizes (2 to 6) of candidate gene sets. 65

Figure 13 Evaluation of the FDR control based on simulations. 66

Figure 14. **TG** cutoff value estimation. For each size of candidate gene sets, the cutoff value of **TG** for controling FDR < 0.05 was estimated by sampling **107** (red square) or **108** (blue diamond) candidate gene sets. 70

Figure 15. High-confidence mutually exclusive gene sets identified from real data analysis. 72
Figure 16. Cumulative $\boldsymbol{k}$-mer percentage at different $\boldsymbol{k}$-mer occurrence from Human reference genome GRCh38. 89
Figure 17 A toy example illustrating FastCount algorithm. 91
Figure 18 Read assignment procedure. 93
Figure 19 Individual barcode identification and gene assignment for a paired-end read. 95
Figure 20 Cell-level feature summarization. 97
Figure 21 Speed and memory usage of FastCount and 3 other pipelines 102
Figure 22 The scatter plots of the total number of UMI counts per cell 105
Figure 23 Pearson correlation of UMI counts within each cell as a function of total UMI counts per cell 105
Figure 24 Compatible t-SNE plots using the feature-barcode matrices 106

CHAPTER 1. INTRODUCTION

1.1 Cancer Biology

The human body is made of approximately 40 trillion cells (Bianconi et al., 2013). Norm cells live harmoniously to form the basic units of life and together to form more complex tissues and organs. The functions of a cell are determined by the genetic material hosted within the cell in structures called chromosomes. Genes are contained in chromosomes that carry hereditary information stored in long strings of DNA bases, adenine (A), guanine (G), cytosine (C), and thymine (T). The DNA sequence in the gene precisely determines the unique structure and functions of each type of protein in the cell: DNA sequences are copied into RNA in a process called transcription; a gene that is transcribed is said to be actively expressed. The transcription of a gene yields an RNA molecule. And the base sequences in the RNA molecule are translated to synthesize protein. There are over 30,000 genes in the human genome. However, not all genes in a cell are expressed and translated into proteins. The expression of different combination of genes within individual cells creates structurally and functionally diverse cell types. Therefore, both DNA sequences and gene expression patterns control the protein synthesis which in turn determines the phenotypes of the cell.

1.1.1 Cancer Genome

Cancer arises as a result of genomic changes that have occurred in a cell. Similar to normal body cells, a cancer cell carries a copy of the genome from its progenitor fertilized egg. However, cancer cells acquire a set of different DNA sequences from the normal cell genome, called somatic mutations. They occur as a consequence of errors when cells divide or exposure to carcinogenic substances that damage DNA, such as certain

chemicals in tobacco smoke, radiation or ultraviolet rays from the sun. Somatic mutations may have several types of DNA sequence changes: 1) point mutations are single nucleotide variations (SNVs) affecting only one base of the gene; 2) insertions or deletions (INDELs) of segments of DNA; 3) copy number alterations (CNAs) are the loss or gain of genetic material from around 1000 bases of a chromosome to the whole chromosome; 4) chromosome rearrangements when a piece of a chromosome breaks and attacheds to another chromosome. These mutations affect the structure, function, and formation of the corresponding proteins. The abnormal proteins change the behavior of normal cells that cause healthy cells to become cancerous. For example, it iswell known that the p53 tumor suppressor gene is a major guardian of the cancer cell (Petitjean, Achatz, Borresen-Dale, Hainaut, & Olivier, 2007). p53 works actively in normal cell to prevent uncontrolled cell growth. But some types of mutations in the TP53 gene give rise to mutant p53 proteins that lose the tumor suppressive function. Cancerous cells take the advantage of the compromised protein function becoming more invasive, metastatic and resistant.

1.1.2 Tumor heterogeneity

The somatic mutations found in the genome of a cancer cell are the result of continuous acquisition of mutations and nature selection of cells with growth advantage during the lifetime of cancer development, analogous to Darwinian evolution framework. In cancer genomics, a clone is defined as a group of cells from the same ancestral cell. As a nonmalignant cell evolves to a malignant one through the continuous and random accumulation of genetic alterations, the stochastic nature of this process results in clones of cells with diverse phenotypes. Some of the changes are neutral rendering no consequences to the cells, while some may give rise to cell clones with different properties.

For example, a clone carrying errant mutations may activate growth factor signaling to promote survival or proliferation. Such clone can outgrow other surrounding cells causing the increase of the clone population. On the other hand, a clone with senescence signaling might be taken over by other clones leading to the declined or loss of the clone. Therefore, tumors are evolving overtime and space and composed of distinct cell clones, known as tumor heterogeneity (Burrell, McGranahan, Bartek, & Swanton, 2013). Tumor heterogeneity is one of the largest challenges in the cancer therapy development. Although there are many FDA approved cancer therapies, as well as the ones in clinical trials, there is no single drug likely to be effective for any cancer types. In many cases, a cancer initially responses to a therapy but acquires drug resistance clones over time leading to cancer relapse. Therefore, understanding the tumor heterogeneity is important in cancer research to overcome drug resistance and develop personalized medicine.

1.2 Next-generation sequencing

The rapid development in Next-generation sequencing (NGS) technology allows cancer research to comprehensively characterize the cancer somatic mutations and tumor heterogeneity with less cost. Traditional sequencing techniques such as the single-gene or array-based methods only allow the genomic exploration of limited targets in low-throughput fashion (Meyerson, Gabriel, & Getz, 2010). For example, Sanger sequencing only sequences a single DNA fragment at a time. Researchers are limited to sequence small stretches of genomic DNA for a small number of samples due to the high cost and low throughput. NGS technique enables massive parallel sequencing of millions of DNA fragments providing a cost-effective way to screening genetic variants on thousands of gene with higher sensitivity, discovery power and sample throughput.

NGS DNA and RNA Sequencing technologies are complementary to each other in cancer research. Genetic mutations in cancer genome can be detected directly at DNA-level with DNA sequencing. Since mutations on DNA-level have consequences on RNA transcription, gene expression analysis using RNA sequencing technique is often used to predict the functionality of genetic changes.

1.2.1 DNA Sequencing

The goal of DNA sequencing (DNA-seq) is to identify genetic irregularities on the genome, such as somatic SNVs, insertions, deletion, CNAs and structure rearrangement, that drive the growth of cancers.

Depending on the sequencing library preparation procedures, DNA sequencing can be applied to the whole genome sequencing (WGS), whole exome sequencing (WES) and pre-selected regions of interests (gene panels). WGS provides the information of nearly complete DNA sequences of the genome (achieving around 95%-98% (Kamps et al., 2017)). WES requires a library enrichment step for all exons. It offers a cost-effective way to survey all the protein-coding regions of the genome (known as the exome) which covers about 1% regions of WGS.

WGS and WES platforms have been implemented in many large well-known national and international collaborations for the comprehensive characterization of the genomic landscape of human cancer. The Cancer Genome Atlas (Cancer Genome Atlas Research, 2013) (TCGA) has analyzed over 11,000 individuals representing 33 different types of cancers revealing common and cancer-specific somatic mutations and signaling pathways. International Cancer Genome Consortium(International Cancer Genome et al., 2010) (ICGC) have collected and analyzed cancer samples globally, spanning over 76

projects. Such sequencing efforts have revealed genetic aberrations that promote tumor initiation, development, and metastasis, which has substantially advanced our knowledge in cancer biology. However, DNA-seq is limited in accessing the gene expression to evaluate the potential functional changes.

1.2.2 RNA Sequencing

RNA-seq addresses the limitation of DNA-seq on the transcriptome level. The conventional bulk RNA-seq, and most recently, single-cell RNA-seq (scRNA-seq) are used for sample level and cellular level gene expression profiling. The bulk RNA-seq measures the gene expression levels from the bulk population of millions of input cells. The resulting expression value for each gene is the average of all the input cells. It is often used in the comparative transcriptomics to measure the global gene expression changes under different conditions. The expression results can then be used for downstream analysis, such as cancer subtype classification or identifying significantly changed cancer pathways. However, there are also important biological questions that bulk RNA-seq is insufficient to answer. Cancer cells are composed of many distinct cell types and clones. It is hard to differentiate whether the gene expression changes are due to the cellular composition of the tumor sample or due to the underlying phenotype changes based on bulk RNA-seq. In these settings, scRNA-seq quantifying the gene expression at the single-cell level provides high-resolution profiling for cancer studies. It has been increasingly used to discover new types and states of cells, and analyze the evolutionary patterns and resulting heterogeneity in cancer.

The conventional RNA-seq technology has been developed more than a decade. The standard workflow begins with library preparation including RNA extraction, mRNA

enrichment, rRNA depletion and complimentary DNA (cDNA) synthesis, followed by NGS sequencing. It captures the expression levels of thousands of genes at once with high accuracy. Since RNA is extracted from groups of millions of cells, this technology is also known as bulk RNA-seq. Bulk RNA-seq measures the average expression level for each gene across all the cells. They provide vast amount of information for comparing gene expression differences in multiple conditions, e.g., tumor vs normal tissues, treated vs non-treated response as well as identifying cancer biomarkers. Bulk RNA-seq design assumes that cells for a given tissue in question are homogenous and gene expression averages across a pool of cells. This process ignores cell-to-cell variability and drops cell-level information which makes it insufficient for studying the heterogeneous tumor cell system.

In recent years, technical advancement in NGS and cell separation methods has made the gene expression profiling at cellular level possible using single cell RNA-seq (scRNA-seq). Several modern scRNA-seq platforms have been developed, such as 10X Chromium, DropSeq and Fluidigm C1, capable of profiling hundreds to thousands of individual cells at once. Those methods use Unique Molecular Identifiers (UMIs) and cell barcodes and have been optimized for single cell expression profiling with low starting amounts of RNA. One of the key steps in scRNA-seq is cell isolation where cells can be physically separated using fluorescence-activated cell sorting (FACS) or they can be trapped inside hydrogel droplets. Next, within each isolated cell, RNA molecules are extracted and tagged with UMIs and Cell barcodes. Then cells break to release the mRNAs for pooled PCR amplification and sequencing. The UMIs and Cell barcodes are important single cell specific information used in the data analysis. Each mRNA within a cell is tagged with a cell barcode. A cell barcode is unique to each cell, it tells which cell the

mRNA is from. In order to trace the cellular origin, each cDNA molecule from the same cell is labeled with a cell barcode, which is an oligonucleotide sequence unique to the cell, before pooled library preparation. This allows early pooling of thousands of samples and increases the sequencing throughput and allow to computationally recover the mRNAs for a specific cell after the pooled PCR amplification step. Moreover, due to the small amount of starting material in each cell, scRNA-seq requires a PCR amplification step for cDNAs. This procedure increases the sensitivity of the scRNA-seq, but on the other hand, PCR duplicates can not be identified simply based on the reads mapping position. By tagging the cDNAs with UMIs, sequences with the same UMI can be detected as PCR duplicates. Therefore, by adding UMIs and cell barcodes in scRNA-seq library preparation, PCR duplicates and cell origins can be computationally identified enabling sensitively measure the cellular-level expression differences.

1.3 Current Computational Methods for Cancer Genomics Data Analysis

Cancer genomics is a new research area that applies the rapid technological development NGS technology to identify the somatic mutations, cancer driver genes, understand cancer biology and find new methods for cancer diagnosis and treatment. NGS allows researchers to interrogate the cancer genome in great resolution, high accuracy and low costs. In the past decade, many NGS approaches have been developed for cancer study to solve the puzzles in DNA and RNA levels. Parallel with the rapid advancement in NGS technologies is the development of novel algorithms for NGS data analysis.

1.3.1 Somatic mutations analysis

Identifying somatic mutations is the key step in cancer genomics for the characterization of a cancer genome. Majority of the somatic mutation calling protocols

require sequencing of matched tumor and normal samples from the same cancer patient using WES or WGS sequencing technique. Somatic mutation calling consists of mainly four components: read preprocessing, variant calling, variant filtering and variant annotation.

Before mutation calling, the resulting FASTQ files are preprocessed to generate high quality analysis-ready Binary Sequence Alignment Map (BAM) file. The quality control (QC) step is performed on the raw FASTQ files to remove poor quality bases and non-biological sequences (not originated from the sample). Reads passing the QC are mapped to a reference genome using read alignment algorithms. PCR duplication, which are reads originating from the same fragment of DNA, can then be detected based on the alignment position in the BAM file. Reads are ranked by their base-quality scores to determine the primary and duplicate reads. Duplicate reads are marked with the hexadecimal value of 0x0400 in the BAM file and will not be included in the variant calling algorithms. The BAM files from the matched tumor and normal samples are subjected to local realignment around insertions and deletions to correct mapping errors resulted from the read aligner and the regions contain INDELs. Since many variant calling algorithms rely heavily on the per base quality score reported by the sequencer in the BAM files, base quality score recalibration is required to correct the over- or under-estimated base quality due to various sequencing technical errors from the sequencer. Pre-processed BAM files are passed to somatic variant calling algorithm for somatic mutation detection. There are various tools publicly available covering a wide spectrum of application in somatic point mutations, INDELs and CNAs calling. Post-filtering of candidate somatic mutations is often required to reduce false positive calls generated due to the NGS artifacts, read

alignment errors and low-quality samples. For example, reads generated by the Illumina platform are commonly affected by the strand bias artifacts, where the heterozygous genotype can only be observed on one specific strand (Guo et al., 2012). Many variant calling pipelines compute strand bias scores and use it as a filter to improve the specificity.

1.3.2 Driver mutations

Somatic mutation analysis pipeline for WES/WGS on an individual cancer reveals hundreds to thousands of somatic mutations present in the cancer genome. One of the major challenges is to prioritize those somatic mutations to identify the ones casually implicated in cancer. These mutations, known as driver mutations, contribute growth advantage in cancer initiation and development, turning on specific pathways promoting cancer. Deciphering driver mutations is the key to design rational therapeutics aimed at specific cancer phenotypes, predict patient response to traditional treatments, and expanding the pool of patients likely to benefit from existing treatments. However, beside the driver mutations, there are a larger fraction of somatic mutations that do not involve in the development of cancer. These non-functional mutations, often known as passenger mutations, happen randomly in cancer cells that have already acquired driver mutations in the cancer genome. They will be passed to descendants during the cancer cell divisions and present in the final cancer cells. Therefore, differentiating the driver mutations from passenger mutations is the main goal for many cancer studies. Despite a few exceptions, most driver mutations occur in only a small fraction of tumor samples. Therefore, identifying these low recurrent driver mutations that are buried among a vast pool of passenger mutations is challenging.

1.3.3 RNA-seq Analysis

Analyzing RNA-seq reads to quantify expression at each gene locus is the first step towards any downstream biological interpretation. There are two popular gene expression estimation approaches for bulk RNA-seq: gene count and transcript abundance. Gene count is essentially the total number of reads sequenced within a gene. Many popular statistical differential expression methods such as DESeq2 (Love, Huber, & Anders, 2014) and edgeR (Robinson, McCarthy, & Smyth, 2010) take gene count as input. They model it as negative bionomial distribution to deal with biological variability and overdispersion and determines differential expression using exact tests (Seyednasrollah, Laiho, & Elo, 2015). Several tools such as featureCounts (Liao, Smyth, & Shi, 2014) and HTSeq (Anders, Pyl, & Huber, 2015) are used to obtain the gene counts. These tools require several preprocessing steps on the raw FASTQ file from the sequencing before performing read counts: 1) generally, a read trimming step is necessary to remove adapter sequences and low-quality bases from the FASTQ files (Bolger, Lohse, & Usadel, 2014; Martin, 2011a). This improves the mappability of the reads during the downstream alignment step. The quality trimming criteria, such as minimum base quality score or the number of bases to be trimmed on start or end of each read, are selected empirically by the users. 2) trimmed reads are aligned to either the reference genome or the reference transcriptome using RNA-seq mappers, to generate the BAM files. 3) aligned reads in the BAM files are assigned to genes based on the genomic locations provided in the Gene Annotation File (GTF) for gene-/transcript-level read counts. Although there are some efficient algorithms available to summarize read counts from the BAM file, read alignment is computationally heavy, requiring large memory and CPU time. Alternatively, read counts can be estimated from

transcript abundance, by first assigning reads to transcripts using transcript-level annotation, and summarizing read counts at transcript-level. The gene-level expression can then be estimated using tools specially designed for summarizing the transcripts counts to gene counts correcting effective gene length differences across samples (Soneson, Love, & Robinson, 2015).

The scRNA-seq gene expression analysis can safely borrow the tools and pipelines that have been developed for bulk RNA-seq. Similar to bulk RNA-seq, reads are first aligned to the reference genome/transcriptome, and then assigned to genes or transcripts depending on the selected algorithms. However, after the read assignment, there are additional single-cell specific steps required for carefully analyzing the data. By introducing UMIs for each cDNA molecule, reads originated from the same molecule will have the same UMI sequence, allowing computationally deduplicate reads to reduce the amplification bias. So instead of counting the number of reads assigned to each gene, scRNA-seq aims at counting the number of unique UMIs. The UMI collapsing procedure, in which reads assigned to the same gene with identical UMIs are only counted once, is performed. However, the UMI counts are often overestimated due to the sequencing/PCR errors within the UMI sequences. Therefore, error correction for UMIs is necessary before collapsing reads to UMIs. Another aspect that complicates the analysis is cell barcoding. cDNAs from the same cell are tagged with a cell-specific sequence, cell barcode, to identify cells and at the same time, allow the sequencing of thousands of cells in a single run. Sequencing errors on the cell barcodes not only overestimates the actual number of cells in the sample, but also may underestimate the UMI counts for the affected cell. Current

analysis pipelines will filter the low sequencing quality barcodes and barcodes with low read count to reduce such errors in the downstream analysis.

1.4 Contribution

The aims of the dissertation are to develop novel bioinformatics methods for resolving several challenges in large-scale cancer genomics study covering sample-level gene mutation analysis and single-cell level gene transcriptome analysis. We assess the performance of methods in the completed projects and demonstrate their application in cancer genomics research.

High-throughput, robust and reproducible Bioinformatics framework for somatic mutation calling. We developed user-friendly Snakepipe framework for systematically processing the raw whole -exome/-genome sequencing data for somatic/germline mutation detection. Snakepipe abstracts the complexity of the analytical pipeline, parameters selection and computation environment deployment from the users. It ships with automated NGS best practices pipelines enabling direct "sequence-to-report" data transformation and requires minimun user configurations. All the analytical softwares are precompiled and packaged into Docker containers allowing cross-platform pipeline execution and built-in version controls for reproducibility. Snakepipe scales well to cloud and HPC infrastrutture for processing large-scale genomics data using distributed computation resources and parallel computing. Snakepipe provides automatic failure recovery in cases of unplanned hardware errors or system downtime. Moreover, new analytical modules can be easily developed and integrated to the existing pipelines.

Efficent mututally exclusive testing for genome-wide drive mutation detection. We developed MEScan, one of the first tools to allow *de novo* screening of mutually

exclusive patterns at the genome scale. The core component of MEScan is a test statistic directly quantifies the discrepancy between the observed level of mutual exclusivity and the expected value due to the background, taking into accout for the backgroud mutation rate heterogeneity and unbalanced mutation patterns. Comparing to other methods, MEScan offers more power in detection of true mutually exclusive patterns even under the conditions with low read coverage and is at least two orders of magnitude faster than the existing methods. We extend MEScan with Markov chain Monte Carlo (MCMC) algorithm to efficiently screen for mutually exclusive gene sets at the genome scale with a summarization procedure to select high-confidence findings. MEScan has been applied to GBM, BRCA, LUSC, OV and PanCancer to identify cancer-specific driver mutations.

Alignment-free gene count quantification for bulk and single cell RNA-seq. RNA-seq has been widely used in cancer research for differential gene expression between samples using bulk RNA-seq and more recently for characterizing differential gene expression at cellular level and studying tumor heterogeneity using scRNA-seq. Although many algorithms have been published, they either have high computational costs in terms of time and resources, or they are designed for transcript-level abudance estimation which requires additional downstream processing for gene-level abudance conversion. Moreover, since most scRNA-seq library protocols have strong 5' or 3'-end bias and are sequenced with low coverage, assigning reads to transcript-level features in scRNA-seq is much more difficult than for gene-level. Therefore, we developed a novel alignment-free gene expression quantification algorithm FastCount that performs gene-level expression analysis for both bulk and single cell RNA-seq . It avoids the time-consuming base-wise alignment step and classifies reads to gene using gene-specic *k*-mer signatures. Comparing

with other methods, it is over an order-of-magnitude faster than the existing gold standard algorithms while achieves very competitive accuracy.

CHAPTER 2. Bioinformatics Framework for Characterization of Squamous Cell Lung Cancers from Appalachian Kentucky

Comprehensive characterization of cancer genomics relies heavily on the bioinformatics pipelines to analyze the massive production of genomic, transcriptomic, epigenomic and proteomic NGS data. Currently pipelines are either not executable cross-platform or not easily customizable to extend for new analysis modules. In this project, we present an open-source, modular computational framework to perform high-throughput, robust and reproducible bioinformatics analyses of cancer genomic data. It automates best practice data analysis pipelines, requires minimum configurations from the users and provides publication-ready figures. We have applied this framework for the characterization of squamous cell lung cancers from Appalachian Kentucky using WES data and have identified distinct genomic landscape and potential therapeutic markers.

2.1 Introduction

NGS has a broad spectrum of applications in cancer genomics, however, the bioinformatic analysis which involves in the transformation of the raw "ATGC" sequence to meaningful genomic information such as gene expression abundance or gene mutations is a non-trivial work. NGS generates millions of DNA sequences for a single sample. The raw DNA sequences are used as source input for cancer biologist to answer various biological questions. Bioinformatics pipelines, which consist of a series of computational software to systematically process the large number of genomic data, have become the power horse for cancer research. However, most researchers have no capacity to perform large-scale analyses on the NGS data sets using appropriate tools and pipelines. This has sparked the need for the development of various analysis pipelines and platforms. In the past few years, there are a variety of analysis pipelines being published such as Galaxy

(Goecks, Nekrutenko, Taylor, & Galaxy, 2010), bcbio-nextgen (https://github.com/bcbio/bcbio-nextgen), Taverna (Wolstencroft et al., 2013), Toil (https://github.com/bd2kgenomics/toil), The Cancer Genomics Cloud (Lau et al., 2017), DNAnexus (http://dnanexus.com), Firehose (http://firebrowse.org/) and many others. They differ in the analysis procedures, tools selection, parameter configurations as well as computational environment. Although current bioinformatics pipeline platforms provide good support to perform data analyses using the built-in modules, customizing and extending the pipelines to meet various research requirements are very difficult. Galaxy (Goecks et al., 2010) platform is a web-based approach that enables researchers with Internet access to perform genomic data analyses through a web page interface. Galaxy users can create analysis pipelines using the interactive, graphical editor by simply connecting software modules pre-wrapped by Galaxy. A similar web-based tool, Taverna, is a workflow management platform, that allows users to define and execute workflows from a web portal. Those platforms, though very helpful for scientists without programming or informatics expertise, have several limitations. For example, users are limited to the number of tools and analyses collected by the platforms; non-programmer users must wait for the platform updates to apply new algorithms on their data; programmatic access to service is not available for advanced users. bcbio-nextgen is a powerful python toolkit for users with extensive programming knowledge. bcbio-nextgen optimizes each analytical pipeline and software for improved computational performance handling job distribution, idempotent processing restarts and safe transactional steps. However, the pipeline customization and development are challenging and complicated even for users with programming background. Cloud-based commercial service providers,

such as The Cancer Genomic Cloud and DNAnexus, offer flexible high-performance computing resources for large-scale collaborative research national and international wise. Both web interfaces and programming APIs are available for users for conveniently submitting and monitoring automated large batch analyses. Users can rent the computing resources such as virtualized computers, data storage and bandwidth on demand and pay for the exact resources used. However, the cloud-based service is currently not an option for small labs or institutes with limited funding resources.

NGS data analysis has been a very active research field. A large number of bioinformatics tools have been published covering every step in the data analysis pipeline. Novel algorithms have continuingly been developed to aid in discovery of new findings in the data, or to improve the performance of existing algorithms. It is challenging for caner biologists to appreciate all the steps and select appropriate tools necessary to conduct the data analysis properly. A basic somatic mutation calling pipeline based on WES would contain as few as 12 steps from raw sequence preprocessing to the final somatic mutation calling. For each individual step in the pipeline, there are collections of tools specially designed for accomplishing the data transformation. The number of bioinformatics tools to choose from can be overwhelming. In the past decades, over 30 somatic SNV callers have been published by different research groups (C. Xu, 2018). They differ considerably in terms of the core algorithms, filtering criteria, and output. Samtools (H. Li, 2011) uses Bayesian approaches to calculate the log-likelihood ratio of tumor and normal samples having the same genotype. Varscan2 (Koboldt et al., 2012) relies on heuristic approaches to identify variants with supporting reads meeting certain thresholds. Then it applies Fisher's exact test on the contingency table of read counts to isolate somatic variants based

on the p-value. Samtools and Varscan2 can report both SNVs and INDELs. Varscan2 can also infer the relative copy number changes in tumor sample by performing pairwise comparison of the read depth between the tumor and its matched normal samples. Mutect (Kristian Cibulskis et al., 2013) is based on the Bayesian classifier approach, but instead of modeling the joint genotypes in the tumor and normal samples, it uses joint allele frequencies to take into account the presence of heterogenous subclones in the tumor sample. Therefore, a pipeline that employs the standardized best practices workflows and analytic tools provide a guide for cancer researchers.

Lastly, analyzing data in large scale and collaborative studies requires a cross-platform, scalable and reproducible pipeline management system. From 2012 to 2017, the amount of genomic data in the Sequence Read Archive (SRA) has doubled four times. Such data archives are comprehensive enough to allow researchers to ask and answer a broad range of sophisticated questions without generating new data. Re-analyzing large collection of data requires pipelines to be easily adapted on different commercial clouds such as Google Cloud Platform, Amazon AWS and Microsoft Azure or academic high-performance computing (HPC) clusters. Large collaboration projects and consortiums require standardized and reproducible pipelines to make sure that each participant will produce the same outputs given the same input data. ICGC, for example, is a large collaboration on cancer research. More than 25,000 cancer omics data at the genomic, epigenomic and transcriptomic levels will be collected and analyzed globally to reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies (International Cancer Genome et al., 2010).

Reproducible analyses among the collaborators are non-trival. It requires a standard way to specify all the pipeline dependencies and execution environments for simplifying deployment, sharing and reusing of tools between research groups.

Snakepipe provides a user friendly bioinformatics pipeline framework. It ships with automated NGS best practices pipelines for analysing WES and WGS data for germline/somatical mutation calling and requires mimimun configurations on the user end. Snakepipe enables direct "sequence-to-results" transformation. It natually determines the dependencies and tools for each individual steps and jobs are distributed to the work nodes for serialized or parallel processing. Once the analyses are complete, Snakepipe automatically collect the input source files, user configurations and results in a compressed format for backup and reproducible research collabration. New pipelines and functions can be easily developed and integerated to the existing ones in the Snakepipe framework by containerization of bioinformatics tools and job definitions in simple python-like scripts.

The overall workflow for mutation calling is comprised of raw seqeuecning read preprocessing, mutation calling, significantly mutated genes identification, CNA calling, cohort comparison, clinical actionable mutation prediction as well as comprehensive results visualization.

2.2 Pipeline management

The pipeline framework adopts a modern workflow management engine, Snakemake (Koster & Rahmann, 2018) and docker containers. The Snakemake engine defines each pipeline in a "Snakefile" using a domain-specific language. It adopts the rule concept used by the GNU Make (Stallman RM, 1991), with extended functionalities and flexibilities. The analysis steps of a pipeline are composed of corresponding rule

definitions. A regular rule expresses 1) input files 2) output files and 3) a shell command or scripts using other programming languages (such as Python or R scripts) which describes how the output files are generated given the input files. In order to naturally represent the plan of job executions and job dependencies in a workflow, Snakemake uses a directed acyclic graph (DAG), where a vertex is the execution of a job defined by a rule and a directed edge indicates the execution sequence of the 2 jobs. A job on the ending vertex of each edge requires the input from the job on the starting vertex of the edge. Therefore, a path which is the sequence of edges in the DAG serializes the execution order of the individual jobs in the workflow. Snakemake has the following properties: i) it automatically detects the rules required for the completion of the final workflow; ii) jobs on disjoint paths can be run in parallel; iii) it only executes rules with missing output files or changes of the input file modification time to avoid re-running the completed jobs and for failure recovery; and iv) jobs can be executed locally and distributed to accessible computing resources, such as cloud or HPC.

2.3 Runtime environment deployment

The analytical tools required in the workflow are managed using docker containers to simplify the deployment of the pipeline under different computing environments, for version control and research reproducibility. Each program and all its dependencies are packaged into a docker container tagged with a unique version id. The docker images are hosted on Docker Hub and can be easily shipped to any machine, either cloud or local. Setting up the tools from scratch only needs minimal configuration of Docker or Singularity. Snakepipe will infer the necessary container images based on the types of the NGS data and automatically pulls the images from the Docker Hub. The pipeline is portable

and robust that enables reproducibility, transparency and shareability. It can be easily deployed and executed cross-platform such as local workstations, HPC and cloud platforms.

2.4 Best practice workflows

Snakepipe WES/WGS data analysis pipeline for germline/somatic mutation calling (**Error! Reference source not found.**) is developed based on the recommended best practice from Broad Institute using the Genome Analysis Toolkit (Aaron McKenna et al., 2010) (GATK). Sequencing reads were trimmed and filtered using Cutadapt (v1.4.1) (Martin, 2011b), then aligned to human reference genome b37/hg19 using BWA (v.0.7.9a) (H. Li & Durbin, 2010). PCR duplicates were removed using Picard (http://broadinstitute.github.io/picard/, v1.115). The Genome Analysis Toolkit (GATK v3.1-1) (A. McKenna et al., 2010) was used for local indels realignment and base quality recalibration. Somatic point mutations and indels were detected using MuTect (v1.1.4) (K. Cibulskis et al., 2013) and SomaticIndelDetector (GATK v2.3-9), respectively, with default settings. Mutations were annotated using Oncotator (v1.4.1.0) (Ramos et al., 2015). Significantly mutated genes were identified using MutSigCV (v1.4) (Michael S. Lawrence et al., 2013). Somatic copy-number alterations (SCNA) analysis was conducted using ExomeCNV (Sathirapongsasuti et al., 2011), an R statistical package. Exonic CNAs were inferred based on the depth-of-coverage ratio between matched tumor and normal samples. Then, CNAs calls were combined into larger segments using circular binary segmentation in DNAcopy (Venkatraman E. Seshan). Gistic2.0 (Mermel et al., 2011) with a confidence level of 0.95 was used on the copy ratio profiles to identify significantly amplified/deleted regions. To evaluate the clinical relevance of the somatic genomic alterations identified in

our cohort, we downloaded the OncoKB database (Chakravarty et al., 2017) (accessed in December 2017) to identify FDA approved drugs for the FDA-recognized and standard care biomarkers.

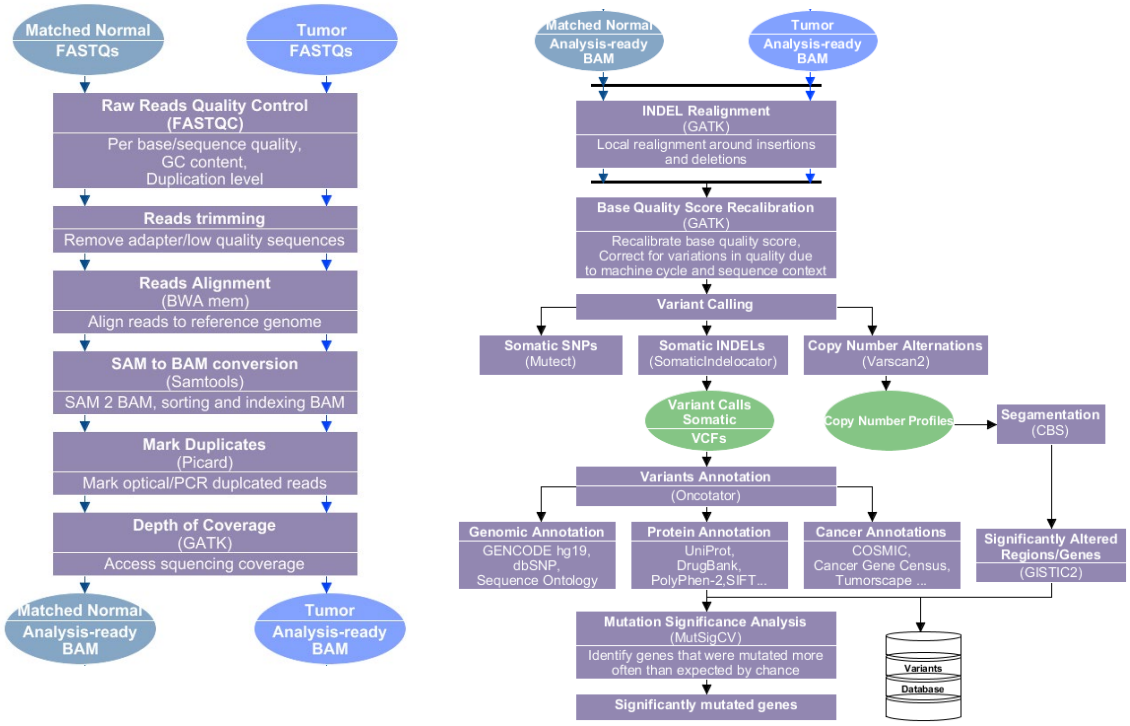


Figure 1 Whole exome sequencing data analysis pipeline for somatic mutation calling and copy number variation detection. The pipeline contains two major parts: 1) preprocessing to prepare analysis-ready bam files from tumor and matched normal samples; 2) variant calling step.

2.5 Results

We applied the Snakepipe to analyze the WES of 51 SQCC patients from AppKY, which includes an overview of somatic alterations and copy-number variations, explores unique mutational patterns, and provides a clinically actionable assessment of mutations in this population. Essential to this effort was the full sharing of the comprehensive genomic profile of lung SQCC in TCGA (The Cancer Genome Atlas Research Network, 2012),

which provided the comparison of the initial 178 subjects from a US genomic profile that does not focus on Central Appalachians.

2.5.1 Overview of somatic alterations

The mean coverage of WES across the targeted regions was 104× with 92% of targeted bases being covered at ≥ 30×. Raw sequencing data are available at dbGaP (Accession: phs001651.v1.p1). We identified 16,005 somatic single-nucleotide variants and 217 somatic insertions or deletions (indels) across 51 matched tumor and normal pairs in the protein coding regions. Of those mutations, 12,117 were predicted to be non-silent mutations resulting in an amino acid change. The mean mutation rate in our cohort was 237 non-silent mutations per patient, corresponding to 8.5 mutations per megabases (Mb). Among non-silent mutations, transitions and transversions at CpG sites were the most commonly observed mutation types, with rates of 11.5 per Mb and 15.5 per Mb, respectively. For non-CpG sites, transitions were more frequently observed at C:G sites (3.2 per Mb) than at A:T sites (1.8 per Mb). Similarly, transversions were more frequently observed at C:G sites (8.0 per Mb) than at A:T sites (2.0 per Mb).

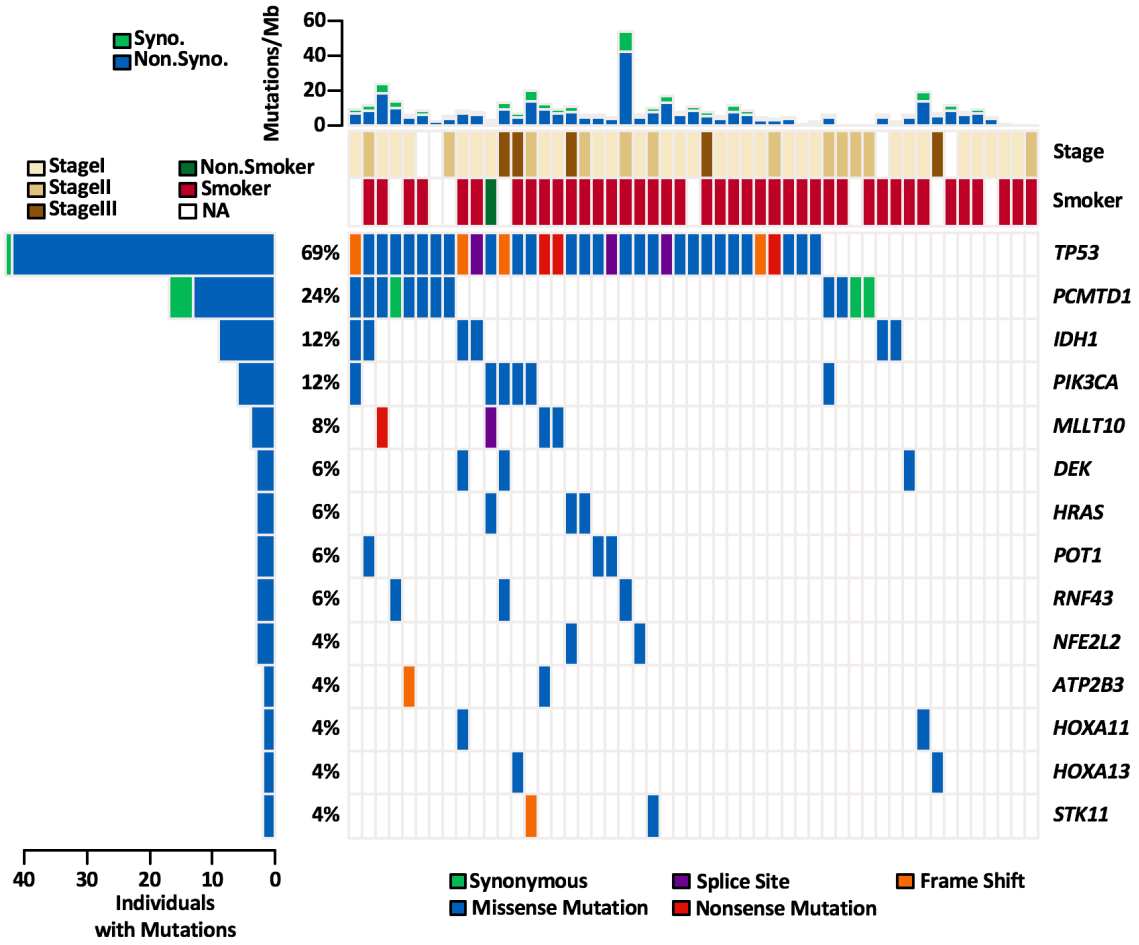


Figure 2. Significantly mutated genes in lung SQCC. Significantly mutated genes (FDR<0.2) from whole-exome sequencing of 51 samples from Appalachian Kentucky patients.The number and percentage of samples with mutations in each gene are shown on the left. Samples are displayed as columns, with the overall number of mutations, smoking status, and tumor stage plotted at the top.

2.5.2 Significantly mutated genes

We identified 3 genes that were significantly mutated (i.e., non-silent mutation rates higher than background mutation rates) in the AppKY cohort with an FDR < 0.2 using MutSigCV (Michael S. Lawrence et al., 2013): *TP53*, *PCMTD1* and *IDH1*. To increase the statistical power of our analysis, we followed the approach of the TCGA SQCC report (2012) and performed a secondary MutSigCV (M. S. Lawrence et al., 2013; Michael S.

Lawrence et al., 2013) analysis to only consider genes causally implicated in cancer according to the COSMIC database (Futreal et al., 2004). This approach enabled us to identify 11 additional genes that were significantly mutated with an FDR < 0.2: *PIK3CA*, *RNF43*, *MLLT10*, *STK11*, *NFE2L2*, *DEK*, *POT1*, *ATP2B3*, *HRAS*, *HOXA11* and *HOXA13* (Figure 2).

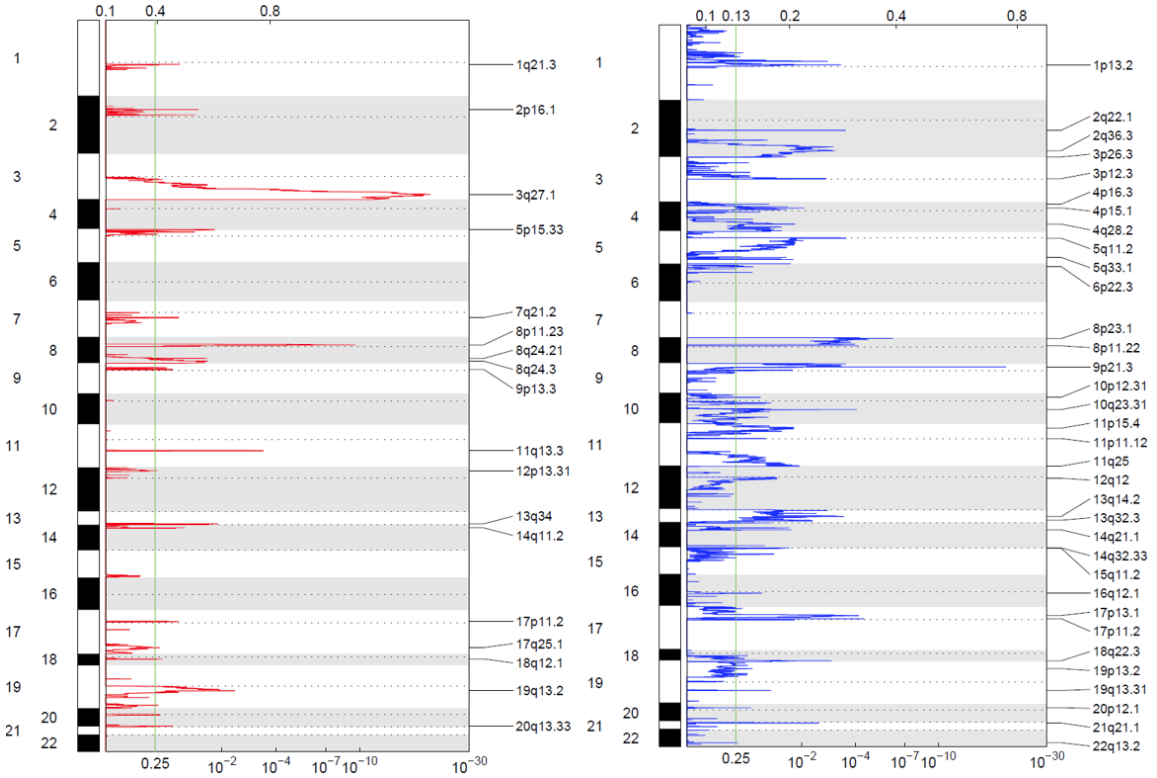


Figure 3 GISTIC amplification (left) and deletion (right) plots of the G-scores (shown at the top of the figure) and q-values (shown at the bottom of the figure) across the entire region analyzed.

2.5.3 Copy number variation analysis

SCNAs were analyzed using WES data. We identified regions with significant SCNAs using Gistic2.0 (Mermel et al., 2011). There were 18 peaks of significant amplification and 34 peaks of significant deletions (FDR<0.25). Significantly amplified regions were 3q27 (*MCF2L2*), 8p11 (*FGFR1*, *TACC1*, *WHSC1L1*, *LETM2*, *RNF5P1*),

11q13 (*CCND1*-oncogene), 7q21.2 *(CDK6*), 19q13, 13q34, 5p15, 8q24 (*MYC*-oncogene) and deleted regions were 9p21 (*CDKN2A*-tumor suppressor, *CDKN2B*), 8p23, 10q23 (*PTEN*, *CFL1P1*, *KLLN*), 17p13, 4q28.2 (*VEGFC*), 22q13.2 (*CHEK2*). Consistent amplification patterns were seen in certain related sets of genes, such as stem cell renewal genes.

2.5.4 Comparative mutational analysis with other cohorts

We first compared somatic mutations and SCNAs of AppKY lung SQCC to TCGA cohort (Campbell et al., 2016; Kim et al., 2014; C. Li et al., 2015; The Cancer Genome Atlas Research Network, 2012). We focused our comparison on significantly mutated genes in at least one cohort by the MutSigCV (M. S. Lawrence et al., 2013) analysis. Our comparative analysis presented here (Table 1) included somatic mutations (point mutations and indels) only in the calculation of gene alteration rate. Both cohorts showed similar rates of alterations for *TP53* (68.6% AppKY, 80.9% TCGA, FDR q-value=1.000), *PIK3CA* (11.8% AppKY, 15.7% TCGA, FDR q-value=1.000), *NOTCH1* (11.8% AppKY, 8.4% TCGA, FDR q-value=1.000) and *PTEN* (5.9% AppKY, 7.9% TCGA, FDR q-value=1.000).

Table 1. Somatic alteration rate comparison between AppKY and TCGA of Lung SQCC. The comparison focuses on genes that were identified as significantly mutated based on the MutSigCV analysis in at least one of the two cohorts.

Hugo Symbol*	AppKY (%)	TCGA (%)	p-value§	q-value¥
IDH1Ⓚ	11.80%	1.10%	0.002	**0.039**
PCMTD1Ⓚ	17.60%	3.90%	0.002	**0.045**
DEK	5.90%	0.00%	0.011	0.200
NFE2L2⊕	3.90%	15.20%	0.032	0.584
CDKN2AⓉ	3.90%	14.60%	0.050	0.830
HOXA11	3.90%	0.00%	0.049	0.830
TP53⊕	68.60%	80.90%	0.082	1.000
PTENⓉ	5.90%	7.90%	0.770	1.000
PIK3CA⊕	11.80%	15.70%	0.655	1.000
KEAP1Ⓣ	9.80%	12.40%	0.806	1.000
KMT2DⓉ	9.80%	19.70%	0.142	1.000
HLA-AⓉ	7.80%	3.40%	0.236	1.000
NOTCH1Ⓣ	11.80%	8.40%	0.424	1.000
RB1Ⓣ	2.00%	6.70%	0.307	1.000
RNF43	5.90%	1.70%	0.126	1.000
MLLT10	7.80%	3.90%	0.269	1.000
STK11	3.90%	1.70%	0.309	1.000
POT1	5.90%	2.20%	0.186	1.000
ATP2B3	3.90%	2.20%	0.617	1.000
HRAS	5.90%	2.80%	0.381	1.000
HOXA13	3.90%	0.60%	0.125	1.000

*Ⓚ: significantly mutated in AppKY only; Ⓣ: significantly mutated in TCGA only; ⊕: significantly mutated in both cohorts

§ The p-value was based on the Fisher's exact test to compare percentages of samples that had somatic alterations (somatic mutations or SCNAs) in the two cohorts.

¥ The q-value was based on the Benjamini–Hochberg procedure. Genes with significant differences (FDR<0.2) in the alteration rate are shown in bold.

Significant differences in mutation rates between the AppKY and TCGA cohorts were observed. The *IDH1* mutations were observed in 11.8% of patients in the AppKY cohort. In contrast, only 1.1% of patients in the TCGA cohort had *IDH1* mutations (FDR

q-value=0.039). Similarly, the AppKY cohort also showed a higher rate of mutations in *PCMTD1* (17.6% AppKY vs. 3.9% TCGA, FDR q-value=0.045). Even after adjusting for age, gender, stage, and smoking via exact logistic regression, mutation frequencies are still significantly different between the AppKY and TCGA cohorts for *IDH1* (p-value=0.0024) and *PCMTD1* (p-value=0.019).

2.5.5 Clinically actionable mutations assessment

We investigated the somatic mutations/SCNAs observed in our cohort in association with FDA approved agents or published or ongoing clinical trials for non-small-cell lung carcinoma (NSCLC) or other tumor types. 5 subjects (10%) had actionable mutations, defined as FDA approved drugs (either for this indication or another cancer type), with a total of 8 somatic mutations/SCNAs events found in these 5 individuals. Additionally, we found that 33 out of 51 subjects (65%) had high (>20 mut/MB) or intermediate (6-20 mut/MB) tumor mutation burden (TMB), indicating an additional group of therapeutic choices for this population using checkpoint inhibitors. Overall, 65% of subjects had actionable mutations with FDA approved drugs and/or TMB that was high or intermediate. Many others had mutations that are under clinical investigation (Table 2).

Table 2. Clinically actionable mutations identified for APPKY patients.

Gene	Patient	Mutation	Drug	DrugLevel
ERBB2	MCC-51	Amplification	Trastuzumab;Neratinib; Lapatinib + Trastuzumab, Pertuzumab + Trastuzumab, Ado-trastuzumab emtansine, Lapatinib, Trastuzumab	1;3A;1
KIT	MCC-12	Amplification	Imatinib;Sunitinib, Sorafenib; Regorafenib, Imatinib, Sunitinib	2A;2A;1
KIT	MCC-36	Amplification	Imatinib;Sunitinib, Sorafenib; Regorafenib, Imatinib, Sunitinib	2A;2A;1
KIT	MCC-47	Amplification	Imatinib;Sunitinib, Sorafenib; Regorafenib, Imatinib, Sunitinib	2A;2A;1
PDGFRA	MCC-12	Amplification	Imatinib	2A
PDGFRA	MCC-36	Amplification	Imatinib	2A
PDGFRA	MCC-47	Amplification	Imatinib	2A
TSC2	MCC-7	Deletion	Everolimus	2A
HRAS	MCC-7	c.181C>A	Tipifarnib	4
NF1	MCC-2	c.55G>T	LTT462, Binimetinib, BVD523, Trametinib	4
HRAS	MCC-12	c.34G>A	Tipifarnib	4
BRAF	MCC-19	c.1391G>T	LTT462, BVD-523, KO-947	4
KRAS	MCC-25	c.40G>A	LY3214996, KO-947, GDC-0994; Binimetinib, Trametinib	4;4
PTEN	MCC-47	c.367C>T	GSK2636771, AZD8186	4
HRAS	MCC-49	c.37G>C	Tipifarnib	4
NOTCH1	MCC-21	Deletion	PF-03084014	4
NOTCH1	MCC-29	Deletion	PF-03084014	4
HRAS	MCC-49	Amplification	Tipifarnib	4
KRAS	MCC-37	Amplification	LY3214996, KO-947, GDC-0994; Binimetinib, Trametinib	4;4
PTEN	MCC-2	Deletion	GSK2636771, AZD8186	4
KRAS	MCC-12	Amplification	LY3214996, KO-947, GDC-0994; Binimetinib, Trametinib	4;4
KRAS	MCC-10	Amplification	LY3214996, KO-947, GDC-0994; Binimetinib, Trametinib	4;4
HRAS	MCC-40	Amplification	Tipifarnib	4

PTEN	MCC-9	Deletion	GSK2636771, AZD8186	4
BRAF	MCC-19	Amplification	LTT462, BVD-523, KO-947	4
PTEN	MCC-33	Deletion	GSK2636771, AZD8186	4

Note: The drug level is defined by OncoKB. Levels 1 and 2 are FDA approved drugs.

2.5.6 Prediction of the effect of IDH1 mutations

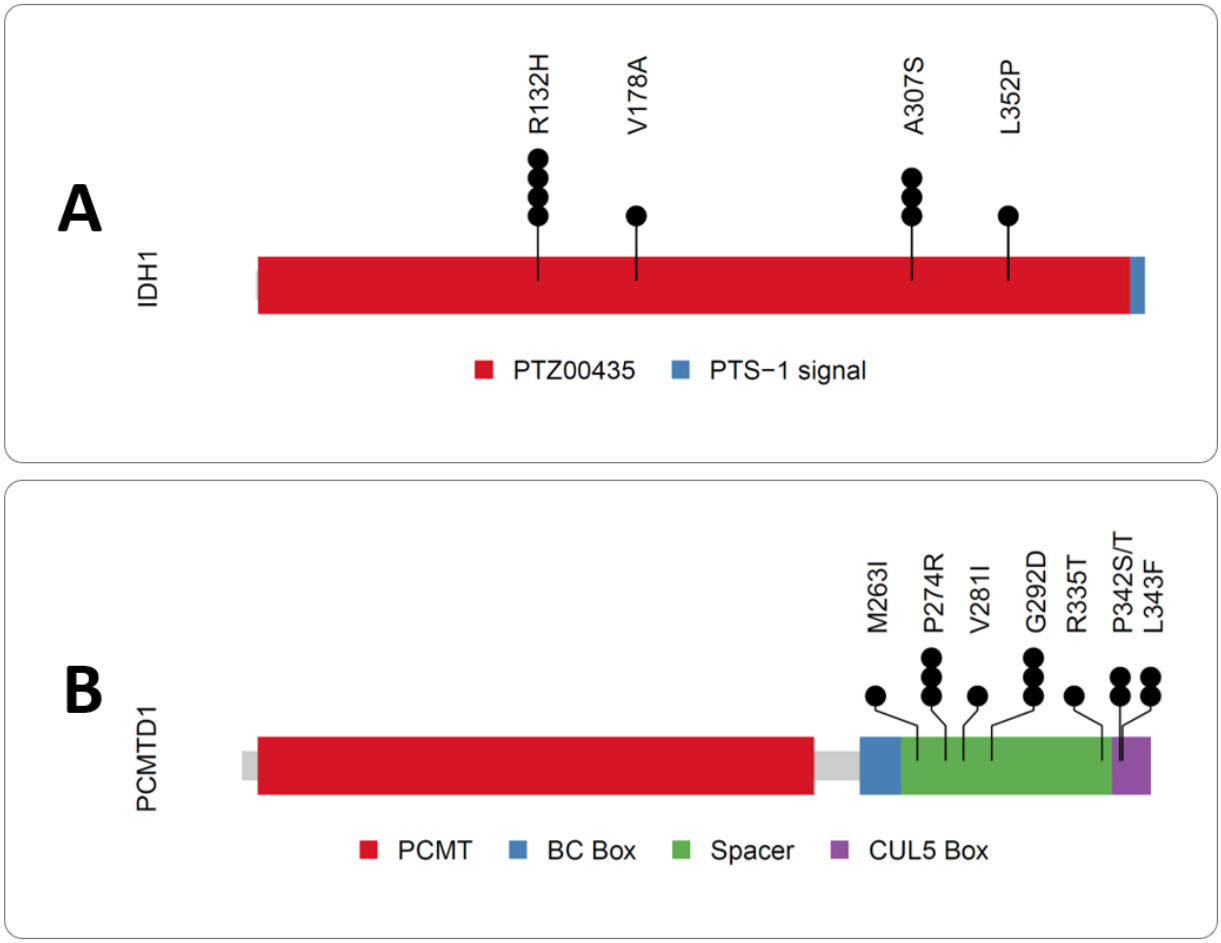


Figure 4. IDH1 and PCMTD1 mutations. (A) IDH1 mutations and their mutation frequencies (circles). (B) PCMTD1 mutations and their frequencies (circles).

Mutations in *IDH1* and its homolog *IDH2* coding for cytosolic and mitochondrial isocitrate dehydrogenases, correspondingly, are common in gliomas (Yan et al., 2009) and myeloid neoplasms (Molenaar et al., 2015), but rare in lung cancer. We observed multiple *IDH1* variants: R132H, V178A, A307S and L352P (Figure 4A), and the R132H variant

was confirmed by immunohistochemistry. The *IDH1* variant R132H is reported in a variety of cancers and the role of various R132 missense substitutions has been studied extensively. These mutations are generally heterozygous, suggesting a gain-of-function by the enzyme, and supported by mechanistic studies demonstrating that the R132H variant protein has an aberrant enzymatic activity, converting α-ketoglutarate (2OG) to (R)-2-hydroxyglutarate (2HG) (Dang et al., 2009). This enantiomer of 2HG acts as an oncometabolite and interferes with cell differentiation (Lu et al., 2012).

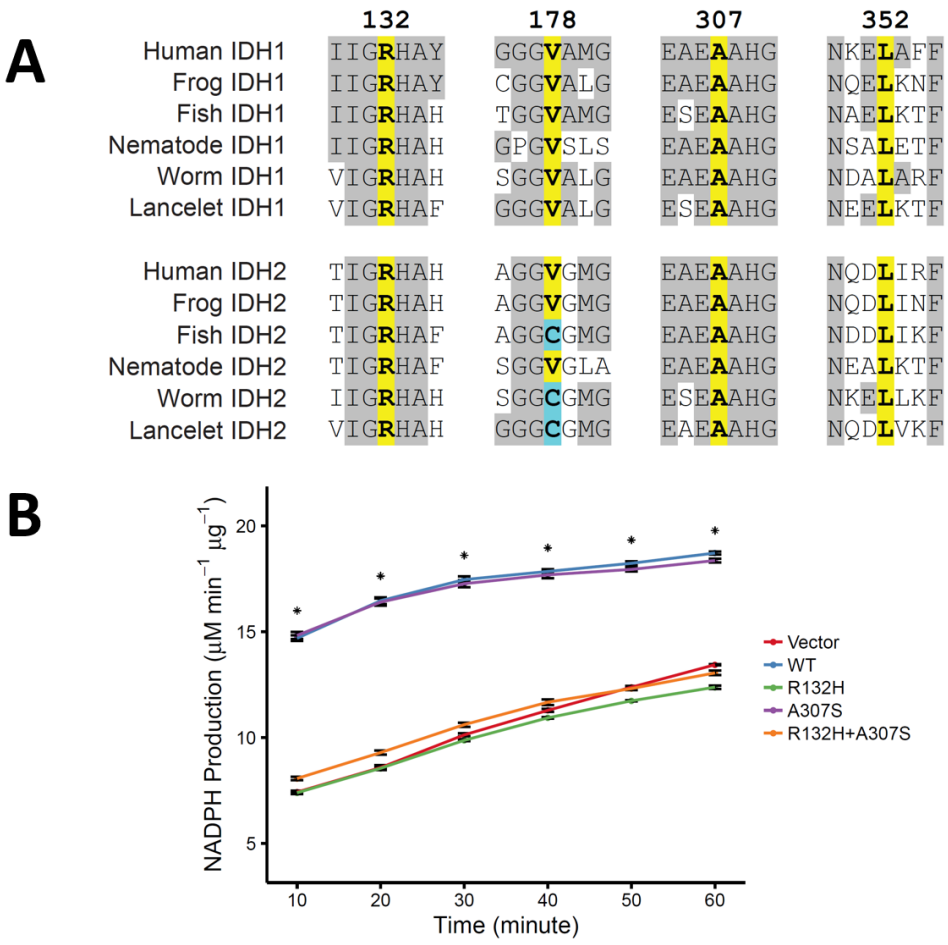


Figure 5. Functional analysis of *IDH1* variants. (A) Segments of multiple sequence alignment for representative IDH1 (upper set) and IDH2 (lower set) orthologs, showing conservation of Arg132, Val178, Ala207, and Leu352. Numbers are provided for a human *IDH1* protein. A complete alignment and sequence accession numbers are shown in Figure 7. Positions 132, 178, 307, and 352 are marked and highlighted in yellow, whereas substitutions in these positions are highlighted in blue. For all other positions, residues that are identical to those in the human *IDH1* are highlighted in gray. Human, Homo sapiens; Frog, Xenopus tropicalis; Fish, Takifugu rubripes; Nematode, Caenorhabditis elegans, Worm, Saccoglossus kowalevskii; Lancelet, Branchiostoma floridae. (**B**) Effect of IDH1 variants on enzyme activity. Left: effect of R132H and A307S mutants; Right: effect of V178A and L352P mutants. The two-sample t-test was performed to compare each *IDH1* mutant versus the wild type and the Bonferroni correction was used for multiple comparison adjustment. • Statistically significant reductions of NADPH production comparing *IDH1* R132H versus wild type; ♦Statistically significant reductions of NADPH production comparing IDH1 L352P versus wild type.

To understand potential consequences of the other detected *IDH1* variants (V178A, A307S, and L352P), we applied a recently developed evolutionary approach (Adebali, Reznik, Ory, & Zhulin, 2016), based on the principle that most deleterious, and hence potentially disease-promoting mutations, result in reduced evolutionary fitness and thus are selected against during evolution. Homologous genes derive from a common ancestor gene, while orthologous genes diverge after a speciation event in two different species; paralogous genes occur within a single species and diverge after a duplication event. Unlike orthologous genes, a paralogous gene evolves new function(s), making the distinction between the roles of orthologous and paralogous genes in disease critical for estimating disease risk using molecular conservation (Adebali et al., 2016). We have identified both *IDH1* and *IDH2* orthologs in representative genomes from all major eukaryotic supergroups and built a maximum-likelihood phylogenetic tree (Figure 6) from their multiple sequence alignment (Figure 7). Satisfactorily, we found that position corresponding to R132 in the human *IDH1* protein is absolutely invariant, not only in orthologous sequences, but in all IDH homologs (Figure 6), which is consistent with deleterious effects of its substitution. Similar to R132, both A307 and L352 are also invariant residues in all *IDH1* and *IDH2* orthologs and all other *IDH1* homologs with uncertain evolutionary history from all major eukaryotic supergroups (Figure 5A and Figure 7). Because no substitutions in these positions occurred since the last eukaryotic common ancestor, any changes in these positions were predicted to be disease-promoting. While position V178 is not invariable among all homologs, the only allowable substitutions are V178I (occasionally found in both *IDH1* and *IDH2*) and V178C (occasionally found only in IDH2) (Figure 5A and Figure 7). No V178A substitution was ever detected in any

IDH homologs, including the most distant ones, and might be cancer-promoting. We therefore tested the activity of these mutations using an enzymatic activity assay.

Figure 6. Maximum likelihood phylogenetic tree of *IDH1* and *IDH2* proteins from a representative genome set. Clades of definite *IDH1* and *IDH2* orthologs are highlighted in green and magenta, correspondingly. Multiple sequence alignment was used to construct the tree and sequence accession numbers are shown in Figure 7. Human proteins are marked by a star.

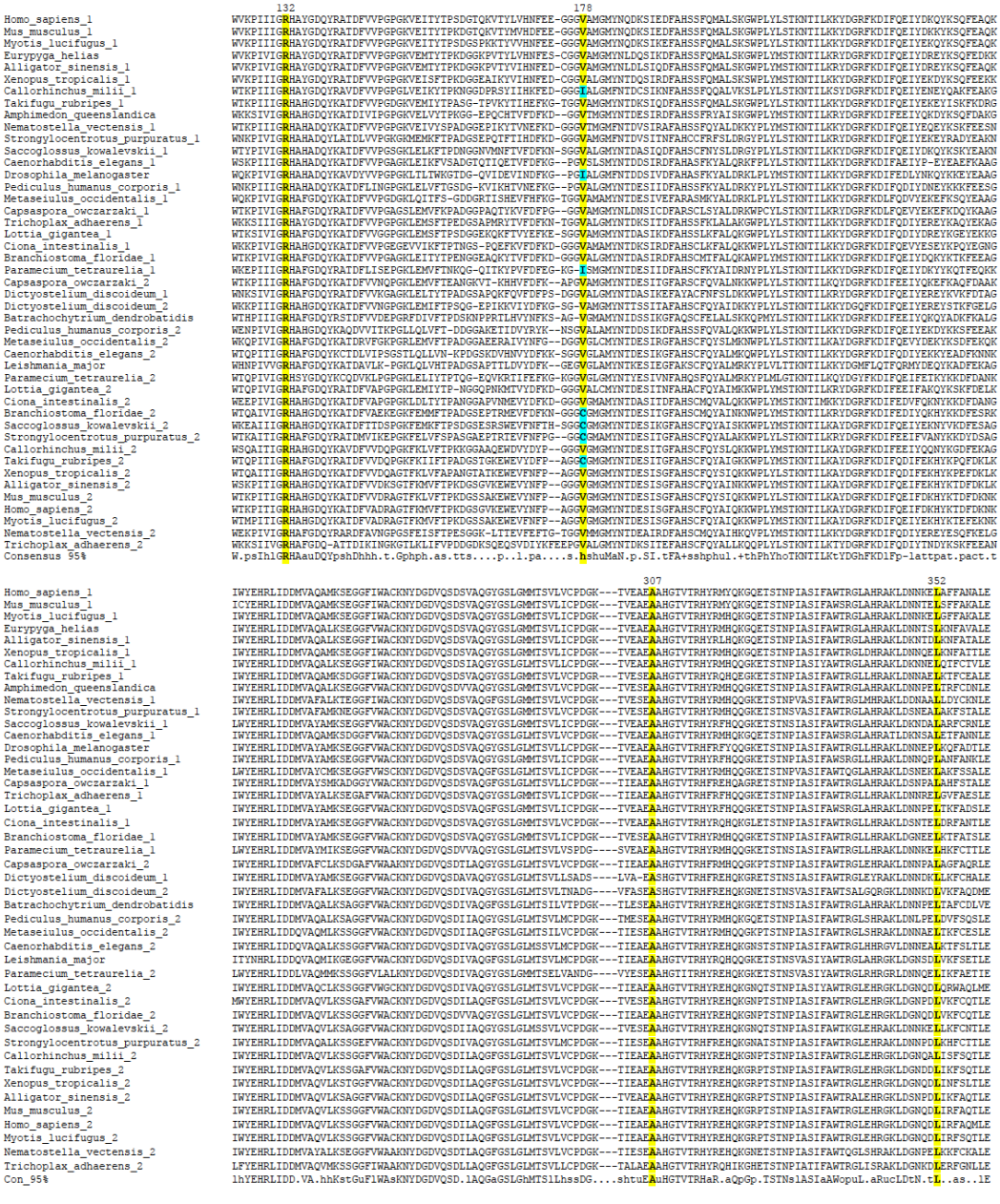


Figure 7. Multiple sequence alignment of *IDH1* and *IDH2* proteins from a representative genome set. Sequence labels and the order of sequences correspond to that of the phylogenetic tree shown in Figure 6. Positions corresponding to K132, V178, A307, and L352 in the human *IDH1* are highlighted.

To test the function of *IDH1* and the effect of different variants on *IDH1* functions, we constructed plasmids with wildtype (WT) *IDH1* and mutant *IDH1* genes (pcDNA3.1-

IDH1-A307S; pcDNA3.1-IDH1-R132H; pcDNA3.1-IDH1-V178A; and pcDNA3.1-IDH1-L352P). We tested the enzymatic activity of the WT and each IDH1 variant by analysis of isocitrate dehydrogenase activity that directly tests NADPH production. We found that R132H and L352P mutations significant attenuated net NADPH production of *IDH1* (Figure 5B), while A307S and V178 mutations had no significant effect. In the context of other R132 IDH1 studies, attenuation of net NADPH production by the R132H variant enzyme implies that production of 2HG in the oncogenic reaction consumes NADPH. These results suggest that R132H is a point mutation that disables or attenuates some enzymatic activity of *IDH1*.

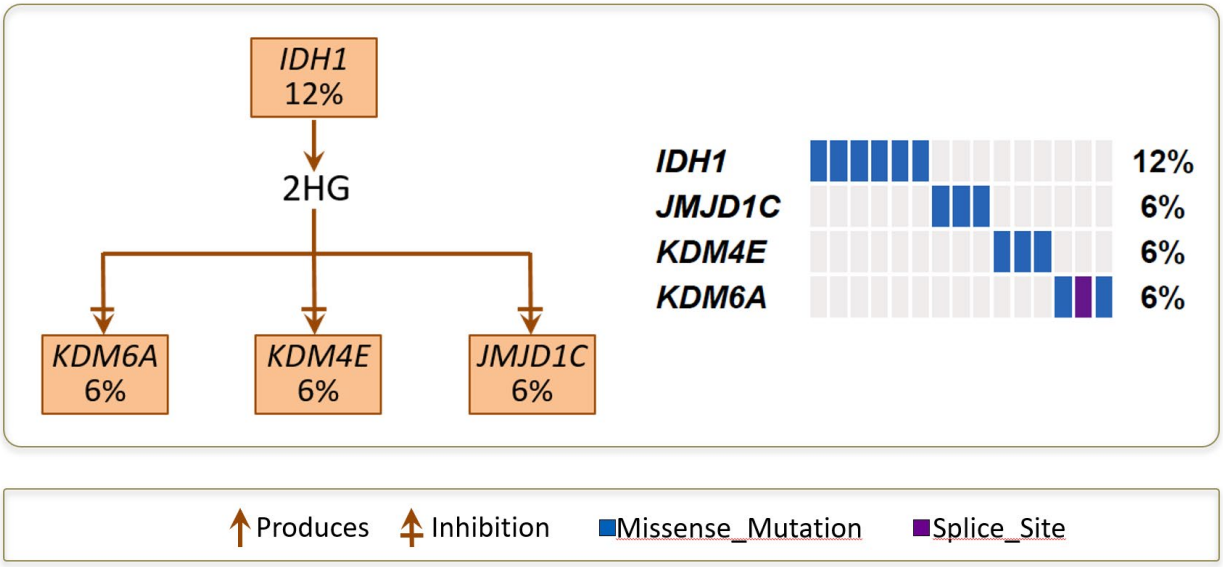


Figure 8. *IDH1* mutations and *IDH1* associated pathway analysis. Variant IDH1 may produce the oncometabolite 2HG that inhibits 2OG-dependent dioxygenases; the 2OG-dependent dioxygenases are highly sensitive to inhibition by 2HG. Mutations in *IDH1* and 2OG dependent enzymes are mutually exclusive. The number and percentage of samples with mutations in each gene are shown on the left. Samples are displayed as columns.

As previously mentioned, certain variants of IDH1 are known to produce the oncometabolite 2HG (Cairns & Mak, 2013; Ward et al., 2012), which showed inhibitory

effects on 2OG-dependent enzymes, with the histone demethylases (KDM) most sensitive to inhibition (Joberty et al., 2016). There are two classes of KDMs: 2OG-dependent and FAD-dependent. The biochemical function of both classes of KDMs is to demethylate specific lysine residues in histones, leading to regulation of gene expression (Labbe, Holowatyj, & Yang, 2013). KDMs may also regulate gene expression via demethylation of other residues on histones (Walport et al., 2016). Based on this information and our discovery of mutually exclusive mutational patterns between certain histone demethylases and methyl transferases, we proceeded to ask if mutations in *IDH1* share a mutually exclusive pattern with 2OG-dependent enzymes in this lung SQCC population. We found that mutations in 2OG-dependent KDMs are mutually exclusive with *IDH1* (Figure 8), suggesting that mutations in either *IDH1* or the 2OG-dependent KDMs lead to a common inhibition of histone demethylation. The mutually exclusive mutational pattern involving *IDH1* is statistically significant (P=0.018 based on the MEGSA (X. Hua et al., 2016) method). This mutual exclusion is a novel observation in lung SQCC, which has not previously been reported. More than 35% of AppKY patients have mutations in 2OG-dependent protein demethylases, the vast majority of them in KDMs. Furthermore, when all lysine demethylases are included in the analyses, only one FAD-dependent, *KDM1A,* is found to be mutated in one case. These data suggest that *IDH1* mutations may regulate gene expression via inhibition of 2OG-dependent KDMs. We further evaluated the mutations in the KDMs to see if they had functional consequences and found mutations possibly affecting a variety of specific regions in each of the different KDMs. The mutations in the KDMs are not localized to a specific region, are highly dispersive across each gene, and functionally affect protein-protein interactions, post-translational

modification sites, and metal-binding, suggesting a general loss-of-function. This loss-of-function interpretation is further strengthened by the fact that *IDH1* mutations responsible for the production of 2HG, which is inhibitory to KDMs (Joberty et al., 2016), are mutually exclusive with mutations in the above mentioned KDMs (Figure 8). The mutational patterns observed between *IDH1* and KDMs suggest that restoring the KDM 2HG-inhibited function in cases with certain *IDH1* mutations may prevent cancer signaling through *IDH1* (Mondesir, Willekens, Touat, & de Botton, 2016).

2.5.7 Localization of PCMTD1 mutations

PCMTD1 has an N-terminal canonical iso-aspartate methyl transferase (PCMT) domain, which in another protein has been shown to methylate iso-aspartate and aspartate residues on proteins including histone H4, and suggests a role in protein repair or turnover (Biterge, Richter, Mittler, & Schneider, 2014; McFadden & Clarke, 1982). *PCTMD1*'s C-terminal domain is not well characterized, and the cellular function(s) of the gene-product are not known. In the AppKY dataset, mutations in *PCMTD1* were always observed in the C-terminus coding region of the protein and never in the N-terminus region. These results are similar to other cancer studies including pancreatic cancer, melanoma, aggressive rhabdomyosarcoma and others (Figure 4B and Table 3). Therefore, the C-terminus coding region of *PCMTD1* appears to be a mutation hotspot.

Table 3. *PCMTD1* mutations. The *PCMTD1* mutations reported in the literature are in the C-terminal SOCS Box. *PCMTD1* mutations in cancers are rarely found in the PCMT domain. The vast majority of mutations (except 1 case in TCGA Lung SQCC and 1 case in Glioblastoma) occur in the SOCS Box**.**

			SOCS Box (240-356)			
Study PMID	**Cancer**	**PCMT (1-239)**	**BC (~16)**	**Spacer (~82)**	**Cul5 Box (~15)**	**% of cases**
22960745	Lung SQCC	Yes	No	Yes	Yes	4%
24793135	Aggressive Rhabdomyosarcoma	No	No	Yes	No	65%
22622578	Melanoma	No	Yes	Yes	Yes	28%
22610119	Prostrate	No	No	No	Yes	1%
24816255	Gastric Carcinoma	No	No	Yes	No	7%
25855536	Pancreatic Cancer	No	Yes	Yes	No	7%
24120142	Glioblastoma	Yes	No	Yes	No	1%
AppKY	Lung SQCC	No	No	Yes	Yes	18%

A recent report indicates that lysine methyltransferases (*KMTs*), *KMT2A* and *KMT2D*, are upregulated by gain-of-function *TP53* mutations (mutations in the DNA binding domain) (Zhu et al., 2015). *PCMTD1* is also a methyltransferase (MT). As mentioned earlier, isoaspartate residues of *TP53* have been shown to be methylated, and this in turn has been shown to regulate levels of *TP53* as well as its function during DNA damage (Lee et al., 2012). *CUL5*, a *PCMTD1* interacting protein is recruited to target the *TP53* protein for proteasomal degradation (Okumura, Joo-Okumura, Nakatsukasa, & Kamura, 2016). We explored the connections between *PCMTD1* and *TP53*, the most frequently mutated gene in the AppKY dataset (69%). *TP53* mutations in this cohort showed a strong signature for a smoking-associated mutational pattern, with frequent mutations in the protein regions 157-159 and 192-193 (Halvorsen et al., 2016). We also

found that the mutations within the smoking signature, specifically the 157-159 region frequently co-occur with mutations in *PCMTD1*.

2.6 Conclusion

From our analyses and other studies, there is growing evidence that numerous pathways converge on protein modification enzymes, including MTs and protein demethylases, that function via direct protein modification, and in the regulation of gene expression via chromatin modification. Therefore, regulation of protein MTs and demethylases affects the methylation status of histones and other substrates such as signaling proteins50. For example, mutations in PI3K/AKT signaling regulate H3K4 methylation through *KDM5A* (Hamamoto, Saloura, & Nakamura, 2015), and *PIK3CA* and *AKT* phosphorylate KDMs and KMTs, which alters their functions and renders them oncogenic5 (Hamamoto et al., 2015; K. Xu et al., 2012). Thus, these methyltransferases and demethylases may be promising targets in cancer therapy.

The observation of a smoking-associated mutational signature in TP53 is not surprising (Schoenberg, Huang, Seshadri, & Tucker, 2015) given the high rate of smoking in AppKY, and this signature appears to frequently co-occur with mutations in *PCMTD1*. We hypothesize that *PCMTD1* could function as a regulator of *TP53*, although further study will be needed to examine this hypothesis. In the AppKY population, concentrations of arsenic, chromium and nickel are higher than the US national levels (Johnson et al., 2011). The toxicity of carcinogenic metals has been shown to be mediated by altering histone methylation via 2OG-dependent enzymes (Arita et al., 2012; Chervona, Arita, & Costa, 2012). In addition to the known link to tobacco exposure, we hypothesize that environmental exposures relevant to AppKY may be contributing to the development of

this (R)-2-hydoxyglutarate-specific cancer mechanism in our cohort. This could help explain the *IDH1* and 2OG-dependent KDMs mutually exclusive pattern seen only in the AppKY cohort.

This study is the first characterization of the genomic alterations in lung SQCC from AppKY residents. Our data shares several findings with the TCGA, namely high rates of *TP53*, *NOTCH1*, *PTEN* and *PI3KCA*, the complexity of genomic patterns, and well-recognized pathways upregulated in SQCC lung cancer. However, the AppKY SQCC has a specific genetic signature characterized by an increased number of *IDH1* and *PCMTD1* mutations, as compared to the TCGA. The findings in this study have important mechanistic implications for how SQCC lung cancers develop in AppKY residents and provide insights into treatment. The 10% potentially actionable mutations/SCNAs observed in our AppKY cohort (based on FDA-approved drugs) coupled with 65% of subjects with high or intermediate mutation burden indicates that a majority of these patients have potential molecular targets for treatment including *ERBB2* amplification with FDA approved monoclonal antibodies and tyrosine kinase inhibitors; *PDGFRA*, and *TSC2* where targeted agents are approved in other tumor types; as well as other mutations with targeted therapies under active investigation (*HRAS*, *KRAS*, *PTEN*, *NOTCH1*, *NF1*, *BRAF*). The current study adds to the body of literature that supports drug development based on mutations in lung SQCC and highlights genomic population differences that are relevant. By utilizing therapies specific to actionable mutations that are common in our AppKY population, we can provide a more personalized approach through directed drug discovery targeting highly mutated genes, such as *IDH1* and *PCMTD1*.

CHAPTER 3. MESCAN: A Powerful Statistical Framework for Genome-Scale Mutual Exclusivity Analysis of Cancer Mutations

3.1 Introduction

Cancer arises from somatically acquired genetic and epigenetic alterations. While large consortia like TCGA and ICGC have profiled genomic somatic mutations of thousands of tumor samples from various cancer types based on whole-genome/-exome sequencing, meaningful mechanistic interpretation of these gene variation results are still very limited. One basic yet challenging task is to distinguish driver mutations, which are causally implicated in cancer development, from passenger mutations, which occur randomly with neutral effect. Despite a few exceptions, most driver mutations occur in only a small fraction of tumor samples (Tamborero et al., 2013). Therefore, identifying these low recurrent driver mutations that are buried among a vast pool of passenger mutations is challenging. Tremendous efforts have been spent on identifying driver mutations ((L. Ding et al., 2018); ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). It has been suggested that assessing mutations in a set of related genes may enhance the power of the detection, since genes act together in various biological (regulatory, signaling, and metabolic) pathways (Leiserson, Blokh, Sharan, & Raphael, 2013; M. D. M. Leiserson, H.-T. Wu, F. Vandin, & B. J. Raphael, 2015; Szczurek & Beerenwinkel, 2014; Vandin, Upfal, & Raphael, 2012). Mutations associated with genes within a pathway often show a mutually exclusive pattern across a cohort of patients, meaning that each patient carries just one mutation in the pathway, which is often sufficient to perturb the function of that pathway. Although the mutation rate for each gene in the pathway is often low, the mutually exclusive mutations among genes in the pathway provide a stronger combined signal that is easier to detect. This is due to the increased

mutation rate by considering the set of genes as a whole as well as the mutually exclusive pattern across genes that provides an additional signal for detection.

Several bioinformatics methods have been developed for *de novo* discovery of mutually exclusive gene sets (Constantinescu, Szczurek, Mohammadi, Rahnenfuhrer, & Beerenwinkel, 2016; Li Ding et al., 2018; Xing Hua et al., 2016; Y.-A. Kim, S. Madan, & T. M. Przytycka, 2017; Leiserson et al., 2013; M. D. M. Leiserson, M. A. Reyna, & B. J. Raphael, 2016; M. D. M. Leiserson et al., 2015; Mina et al., 2017; Szczurek & Beerenwinkel, 2014; Vandin et al., 2012). However, there are still three major challenges. Firstly, the heterogeneity in background (or passenger) mutation rate needs to be adjusted. Lawrence et.al (Michael S. Lawrence et al., 2013) demonstrated large variation in the background mutation rate across genes and across patients of the same cancer type from TCGA data. Adjusting for a patient- and gene-specific background mutation rate has been shown as the key to reducing artifactual findings and improving the identification of driver genes (Korthauer & Kendziorski, 2015; Michael S. Lawrence et al., 2013; Youn & Simon, 2011) . This is also true for dN/dS-style tests, where dS represents a proxy for background mutation rate (Nik-Zainal et al., 2016; Zhao et al., 2019). The heterogeneity in the background mutation rate can also affect identification of mutually exclusive mutational patterns, because spurious patterns are more likely to occur in genes and patients with high background mutation rates. However, only a few mutual exclusivity analysis methods have taken into account the heterogeneity in the background mutation rate, and adjustment approaches are limited. Hua et al. (X. Hua et al., 2016) used a likelihood-based approach to directly adjust for the background mutation rate. However, the method is based on the assumption that the relative mutation frequencies of genes in a mutually exclusive gene set

are proportional to the background mutation frequencies of those genes. It also assumed the same background mutation rate of a gene across patients. A few other methods (Y. A. Kim, S. Madan, & T. M. Przytycka, 2017; M. D. Leiserson, M. A. Reyna, & B. J. Raphael, 2016; M. D. Leiserson, H. T. Wu, F. Vandin, & B. J. Raphael, 2015) used a conditional technique to indirectly adjust for the mutation rate heterogeneity. These methods used either permutation or a hypergeometric distribution method to make inferences by conditioning on the observed mutation frequencies of genes and patients. However, the conditional technique was unable to distinguish whether the observed mutation frequencies were due to random background noise or true signals that drive cancer development.

Secondly, as pointed out by Leiserson et al. (M. D. Leiserson et al., 2015), a gene with a very high mutation rate plus a few other genes with very low mutation rates may show a mutually exclusive mutational pattern by random chance. The highly mutated gene, e.g. TP53 in several cancer types, could be a driver gene by itself. But other genes in this spurious mutually exclusive set may not be drivers and may be biologically unrelated to the highly mutated gene. Therefore, such an unbalanced pattern, which is dominated by the highly mutated gene, is less of interest as compared to a more balanced pattern, where each gene in the gene set has a non-negligible contribution to the overall pattern. Note that adjusting for the background mutation rate does not solve this problem. The highly mutated gene could be a driver whose mutation rate is much higher than the background so that the pattern would still be significant even after the background mutation rate adjustment. Many bioinformatics methods do not distinguish unbalanced and more balanced patterns, and therefore can lead to spurious results. Although a conditional method (M. D. Leiserson et

al., 2015) has been proposed to favor more balanced patterns, its power could still be affected by the presence of a highly mutated gene based on our simulations.

Thirdly, computational efficiency is a major hurdle for genome-scale screening of mutual exclusive gene sets. Most methods (Constantinescu et al., 2016; X. Hua et al., 2016; Y. A. Kim et al., 2017; M. D. Leiserson et al., 2016; M. D. Leiserson et al., 2015; Szczurek & Beerenwinkel, 2014) are based on statistical tests to examine mutual exclusivity of gene sets. However, current statistical tests have high computational burden because they involve computationally intensive statistical modeling and/or require permutation to calculate p-values. Furthermore, the computational burden increases dramatically as the size of the candidate gene set increases. A few methods have been proposed to reduce this computational burden. WExT (M. D. Leiserson et al., 2016) used a saddlepoint algorithm to approximate the permutation test, but its computational efficiency was not sufficiently high. WeSME (Y. A. Kim et al., 2017) proposed a weighted sampling algorithm instead of permutation, but the algorithm was limited to examining two genes at a time. As a compromise, most methods only focused on genes with relatively high mutation rates and/or known to be cancer drivers. The number of genes they considered was typically less than 1000, or even less than 100, which limited their ability to perform genome-scale screening.

Due to these major hurdles, current mutual exclusivity analysis methods have limited ability of analyzing the whole genome to identify novel driver genes, especially those with low mutation frequencies. In this project, we explore methods for removing those hurdles so as to unleash the power of mutual exclusivity analysis for genome-wide screening of driver gene mutations. To address the challenges mentioned in the cancer

driver mutation discovery, MEScan is developed based on a statistical test to *de novo* screen mutually exclusive patterns at the genome scale. The framework has the following key component: 1) the test statistic directly quantifies the discrepancy between the observed level of mutual exclusivity and the expected value due to the background, where a patient-specific and gene-specific background mutation rate is taken into account.; 2) it incorporates a gene-specific weight to adjust for gene mutation frequencies, favoring more balanced rather than unbalanced patterns; 3) test statistic only involves simple algebra, and thus is very fast to calculate. Equipped with this very fast test, MEScan implement a Markov chain Monte Carlo (MCMC) algorithm to efficiently scan for mutually exclusive gene sets at the genomic scale, a false discovery rate (FDR) adjustment procedure to control false positives, and a summarization procedure to select high-confidence findings. We demonstrate our test statistics outperforms several existing methods based on simulation studies. And our algorithm has been applied to TCGA data for genome-scale screening of mutually exclusive gene sets.

3.2 MEScan Framework

The overview of the MEScan framework is in Figure 9. Overview of the MEScan framework.. We propose a test statistic, T_G, to examine whether a candidate gene set G pertains to a mutually exclusive mutational pattern. The T_G quantifies the difference between the observed potential of mutual exclusivity in G with its expected value due to background noise. A T_G larger score indicates that the gene set is more likely to be mutually exclusive. As the background mutation rate varies across patients and genes, the T_G incorporates a patient- and gene-specific background mutation rate in the calculation to adjust for the background noise. In addition, T_G includes a gene-specific weight to down-

weigh genes with very high mutation rates, which could lead to spurious unbalanced mutually exclusive patterns. As illustrated in the figure, the candidate gene set $G = (g_2, g_4, g_5)$ appears to show a mutually exclusive mutational pattern. However, most of the mutations are from gene g_2 while the other two genes, g_4 and g_5 , have very few mutations. The apparent mutually exclusive pattern is highly unbalanced and dominated by g_2. To balance the impact of each individual gene on the overall pattern, our T_G statistic includes a gene-specific weight, which is inversely correlated with the gene's mutation rate, to reduce the impact of g_2. Furthermore, T_G is very fast to calculate, which is critical for genome-scale screening over a vast number of candidate gene sets.

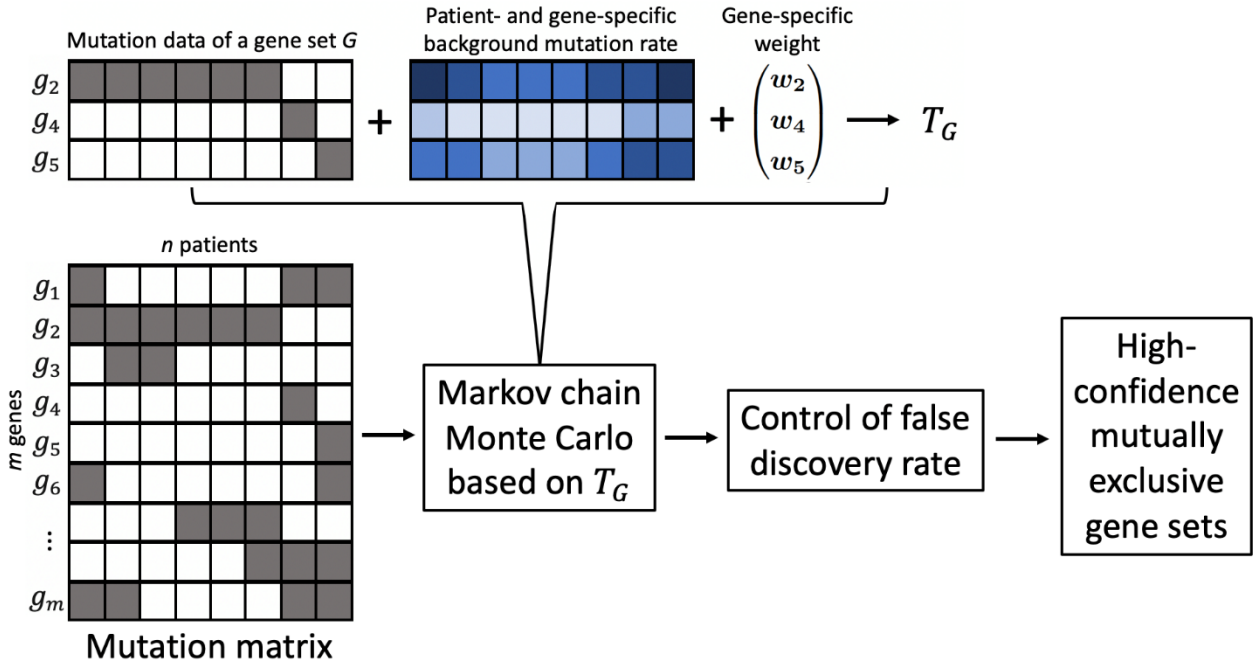


Figure 9. Overview of the MEScan framework.A key component of MEScan is a fast and powerful statistical test, T_G, for assessing mutual exclusivity of a candidate gene set. This test accounts for a patient- and gene-specific background mutation rate (for illustration, darker blue indicates higher and lighter blue indicates lower background mutation rate). By using a gene-specific weight, the test also balances the impact of each gene on the overall significance of the gene set. Based on this test, our genome-scale screening follows a multi-step procedure. Starting from the observed mutation data matrix, an MCMC algorithm is used to screen across candidate gene sets, where the probability of a gene set being sampled is proportional to the T_G score of that set. Next, significant gene sets are identified with the control of the FDR. Finally, high-confidence gene sets are selected based on the criteria that all subsets of them are also significant and they do not have substantial overlaps.

Building upon this test, we use a multi-step procedure for genome-scale screening of mutually exclusive gene sets. Firstly, we use an MCMC algorithm to efficiently identify potential mutually exclusive gene sets at the genomic scale. According to the COSMIC database (Forbes et al., 2015), somatic mutations have been identified in over 20,000 genes. Given the vast number of candidate genes and numerous combinations of genes to form gene sets, examining all the possible gene sets is impractical. Therefore, for each size of candidate gene set, we construct a Markov chain such that the probability of sampling each

gene set is proportional to the T_G score of that gene set. This method allows for more efficient screening and puts more focus on the gene sets that are more likely to be mutually exclusive. Secondly, we identify significant mutually exclusive gene sets by implementing an FDR control method. Finally, we select high-confidence mutually exclusive gene sets by investigating significant gene sets across different sizes. Roughly speaking, a high-confidence mutually exclusive gene set satisfies that 1) itself and all its subsets are significant gene sets; and 2) it does not have substantial overlap with another high-confidence mutually exclusive gene set. We expect these high-confidence mutually exclusive gene sets, which are strongly supported by the data and distinct from each other, are of most interest for further biological interpretation and investigation.

3.3 Testing mutual exclusivity of a single gene set

We have developed a new statistical test to examine the presence or absence of a mutually exclusive pattern for a gene set G based on mutation data from a cohort of n patients, while adjusting for patient- and gene-specific background mutation rate as well as the impact of highly mutated genes. Our test statistic quantifies the observed potential of mutual exclusivity beyond what is expected due to random background for each gene and patient, and then takes a summation across genes and patients. To favor more balanced patterns, each gene's contribution to the overall test is weighted by a factor inversely correlated with its mutation rate.

Let U_{ig} take value 1 or 0 to indicate whether the ith patient satisfies the mutually exclusive mutational pattern and the mutation occurs in a gene $g \in G$, that is,

$$U_{ig} = \begin{cases} 1 & \text{if } Y_{ig} = 1 \text{ and } Y_{is} = 0 \text{ for } s \neq g, s \in G \\ 0 & \text{otherwise,} \end{cases}$$

where Y_{ig} takes value 1 or 0 to indicate if gene g is mutated in patient i. Under the null hypothesis of no mutually exclusive pattern, the expectation of U_{ig} is

$$E(U_{ig}) = P(U_{ig} = 1) = \eta_{ig} \prod_{s \neq g, s \in G} (1 - \eta_{is}) \equiv \theta_{ig}$$

where η_{ig} is the background mutation rate for gene g in patient i calculated based on the MADGiC (Korthauer & Kendziorski, 2015) method. MADGiC considers a multiplicative model that quantifies the patient- and gene-specific background mutation rate by a product of parameters representing a number of factors that are known to affect the mutation rate. Those factors include patient-specific mutation rate, mutation type and dinucleotide context (the specific nucleotide change of the mutation and whether the mutation occurs in CpG dinucleotides), replication timing of the region and expression level of the gene. The empirical Bayes method is used to estimate the patient-specific mutation rate parameter, and the method of moments is used to estimate other parameters.

We quantify the contribution of gene g in patient i to the mutually exclusive pattern by $Z_{ig} = U_{ig}(U_{ig} - \theta_{ig})$, which calculates the difference between the observed value of U_{ig} and its expected value under the null hypothesis. By standardizing Z_{ig} and taking a weighted sum across genes in G, we obtain the following statistic to quantify the evidence of mutual exclusivity in the ith patient

$$T_i = \sum_{g \in G} w_g \times \frac{Z_{ig}}{\sqrt{Var(Z_{ig}) + \lambda}}, \quad (1)$$

where $\mathrm{Var}(\mathrm{Z_{ig}}) = \theta_{ig}(1-\theta_{ig})^3$ is the variance of $\mathrm{Z_{ig}}$, and λ is a small constant to mitigate the impact of extremely small θ_{ig}values. Following the suggestion of (Tusher, Tibshirani, & Chu, 2001), we set λ to be the 5th percentile of all $\sqrt{\theta_{ig}}$'s, where $\sqrt{\theta_{ig}}$ is approximately the standard deviation of $\mathrm{Z_{ig}}$ because θ_{ig} is usually much smaller than 1 so that $\sqrt{\theta_{ig}(1-\theta_{ig})^3} \approx \sqrt{\theta_{ig}}$.

It is important to note that in Equation (1), we include a gene-specific weight, $\mathrm{w_g}$, to adjust for the difference in mutation rate of genes in G. Specifically,

$$\mathrm{w_g} = \frac{1/\sum_{i=1}^{n} \mathrm{U_{ig}}}{\sum_{s \in G}[1/\sum_{i=1}^{n} \mathrm{U_{is}}]}.$$

As $\mathrm{w_g}$ is inversely correlated with the mutation rate of gene g, it down-weighs the impact of highly mutated genes, such as TP53, to the overall statistic, and therefore makes the statistic favor balanced patterns. The $\mathrm{w_g}$ removes the confounding effect of the difference in genes' mutation rates by standardizing the statistic to a balanced pseudo-population, where the number of subjects having mutations in g but not other genes in =-G is the same for each $\mathrm{g} \in \mathrm{G}$. It is analogous to the inverse probability weighting in survey sampling (Little, 1991; Pfeffermann, 1996).

Finally, we take the sum of T_i over all patients and standardize it to obtain our test statistic, T_G, for mutual exclusivity of gene set G:

$$\mathrm{T_G} = \frac{\sum_{i=1}^{n} \mathrm{T_i} - \sum_{i=1}^{n} \mathrm{E(T_i)}}{\sqrt{\sum_{i=1}^{n} \mathrm{Var(T_i)}}} \quad (2)$$

where the expectation and variance of T_i are

$$E(T_i) = \sum_{g \in G} \frac{w_g \theta_{ig}(1-\theta_{ig})}{\sqrt{\theta_{ig}(1-\theta_{ig})^3 + \lambda}},$$

and

$$Var(T_i) = \sum_{g \in G} \frac{w_g^2 \theta_{ig}(1-\theta_{ig})^3}{\left[\sqrt{\theta_{ig}(1-\theta_{ig})^3}+\lambda\right]^2} - \sum_{g,s \in G, g \neq s} \frac{2 w_g w_s \theta_{ig}(1-\theta_{ig})\theta_{is}(1-\theta_{is})}{\left[\sqrt{\theta_{ig}(1-\theta_{ig})^3}+\lambda\right]\left[\sqrt{\theta_{is}(1-\theta_{is})^3}+\lambda\right]}$$

The T_G can be calculated very quickly because the formula only involves simple algebra. This high computational efficiency is key to enable screening over a vast number of candidate gene sets.

3.4 Genome-wide screening

The efficiency test T_G makes it possible to perform genomic scale screening for mutually exclusive gene sets from thousands of genes. However, due to the vast number of candidate gene sets, it is still impractical to perform a mutual exclusivity test exhaustively on each of those gene sets. Therefore, we consider an MCMC method to screen candidate gene sets more efficiently and prioritize gene sets that are more likely to pertain the mutually exclusive pattern. We define a probability distribution on candidate gene sets satisfying that the probability of a candidate gene set is proportional to its T_G score. A Markov chain is then constructed to have that probability distribution as its equilibrium distribution. Therefore, the MCMC algorithm favors sampling gene sets with large T_G scores, which are more likely to be mutually exclusive sets. A similar approach was used in (M. D. M. Leiserson et al., 2015).

In the implementation, we consider a separate MCMC for each size of gene sets. For each MCMC, we use the following Metropolis-Hastings algorithm to obtain Monte Carlo samples. For a gene set G, we define NB(G) as a collection of its neighborhood gene sets who contain the same number of genes as G and differ from G by only one gene. We require that a gene set can only transit to its neighborhood gene sets. Specifically, at each MCMC iteration, the proposed state G' given the current state G is a random sample from $NB(G)$. The Metropolis acceptance probability for G' is

$$\mathrm{r}(\mathrm{G},\mathrm{G}') = \min\left(1, \left(\frac{\mathrm{T}_{\mathrm{G}'}}{\mathrm{T}_{\mathrm{G}}}\right)^{\tau}\right),$$

where τ is a tuning parameter to control the acceptance rate to be around 30%.

3.5 Determining a cutoff value to control the FDR

We identify significantly mutually exclusive gene sets by controlling the FDR<0.05 based on the local fdr method from a previous publication (Efron, 2004b). The local fdr method considers the observed distribution of T_G as a mixture of null and non-null distributions. It empirically estimates the null and non-null distributions for possibly non-independent test statistics of large-scale simultaneous hypothesis testing. The FDR is then calculated based on the empirical null and non-null distributions. A cutoff value of T_G corresponding to FDR < 0.05 is determined so that gene sets with T_G scores larger than the cutoff value are considered as significantly mutually exclusive. The cutoff value is determined for each size of gene sets separately. Note that the original method in (Efron, 2004b) requires using the T_G's of all candidate gene sets to estimate the empirical null and non-null distributions and determine the T_G cutoff value, which is computationally intractable for our situation. As those gene sets are randomly selected, they are likely to

represent the distribution of T_G in all candidate gene sets. Therefore, instead of using all candidate gene sets, we randomly sample 10^7 gene sets to estimate the empirical null distribution and determine the T_G cutoff value for each size of gene sets. Based on real data analysis, as shown in Figure 14. T_G cutoff value estimation. For each size of candidate gene sets, the cutoff value of T_G for controling FDR < 0.05 was estimated by sampling 10^7 (red square) or 10^8 (blue diamond) candidate gene sets., sampling 10^7 gene sets are sufficient to obtain stable cutoff values.

3.6 Identifying high-confidence mutually exclusive gene sets

By applying the T_G cutoff value as described in the last subsection, we can identify a number of significant mutually exclusive gene sets with FDR < 0.05 for each size of gene sets. Let $\mathcal{M}$ be a collection of the significant gene sets across all sizes.

Based on our experience, there can be a large number of gene sets in $\mathcal{M}$ and many of those gene sets overlap with each other. To promote more robust and focused inferences, we further define high-confidence mutually exclusive gene sets, satisfying that 1) all subsets are also significantly mutually exclusive; and 2) different gene sets do not have substantial overlaps. A two-step procedure is used to select high-confidence mutually exclusive gene sets. The first step identifies all maximal cliques in $\mathcal{M}$. A clique is defined as a gene set (size ≥ 3) such that itself and all of its subsets (size ≥ 2) are all in $\mathcal{M}$. A maximal clique is a clique that cannot be expanded by including any additional gene. These maximal cliques are likely to be real mutually exclusive sets because they are validated by all their subsets. Note that we do not consider gene sets of size 2 as cliques because they do not have subsets to validate. The second step removes largely overlapped maximal cliques of the same size. For maximal cliques of size > 3, if the number of overlapped genes

between two maximal cliques of the same size is > 50% of the size, we remove one of them with a lower T_G score. For maximal cliques of size = 3, because the size is too small to define meaningful overlaps, we simply select the top 100 maximal cliques with the largest -T_G scores. After the two-step procedure, the remaining maximal cliques are considered as high-confidence mutually exclusive gene sets.

3.7 Results

3.7.1 Simulation studies

We performed simulation studies to evaluate the performance of MEScan and compared to the following five existing methods: MEGSA, Dendrix (version 0.3), TiMEx (version 0.99.0), WExT (weighted-row exclusivity test, version 1.3.0) and CoMEt. For CoMEt, we used the WExT row-exclusivity test implementation as suggested by the paper (M. D. Leiserson et al., 2016). To mimic a real-world situation, the simulated datasets were generated based on the TCGA ovarian cancer dataset described in the Real data analysis subsection.

3.7.1.1 Simulation studies to evaluate methods' performance in identifying subsets of a true mutually exclusive gene set without the presence of highly mutated genes

As the goal of the analysis is to identify truly mutually exclusive mutation patterns while avoiding spurious patterns, the following simulation studies were conducted to evaluate and compare each method's performance in ranking candidate gene sets. We randomly selected 200 patients and 3 genes from TCGA ovarian cancer dataset and artificially added a mutually exclusive mutational pattern on 10%, 20%, 30%, or 40% of patients, which was referred to as the coverage. We considered two different mutually exclusive mutational patterns, one with a 1:1:1 ratio of mutation frequencies for the three

genes (equal number of mutations in each gene) and the other with a 3:2:1 ratio of mutation frequencies. We additionally included 17 other genes, each has at least 5 mutations in the TCGA ovarian cancer dataset, as “noisy” genes without any mutually exclusive pattern. We considered two different approaches to select those genes. One approach was to randomly select 17 genes from the real data. The other was to intentionally include TP53, which had a high mutation frequency of 94.6%, and randomly select the other 16 genes. This second approach aimed to assess each method's performance in the presence of a highly mutated driver gene but not part of the mutually exclusive pattern, where such a gene could yield spurious unbalanced mutually exclusive patterns by random chance.

We first evaluated methods' performance in identifying the true 3-gene mutually exclusive gene set. Under each scenario, we applied each method (except for TiMEx) to all candidate gene sets of size 3 and identified the top-ranked gene set. Here, the candidate gene sets were ranked based on the T_G score for MEScan, the weight W for Dendrix, the p-value for CoMEt, WExT and TiMEx, and the likelihood for MEGSA. Note that because TiMEx is computationally intensive, we only applied it to a smaller subset of candidate gene sets, i.e., the union of top 10 gene sets ranked by each of the other methods and the gene set with the true mutual exclusive pattern, which may bias the result in favor of this method. Our simulations were replicated 100 times and the frequency that the top-ranked gene set was the gene set containing the true mutually exclusive mutation pattern we generated was calculated, which was referred to as the power.

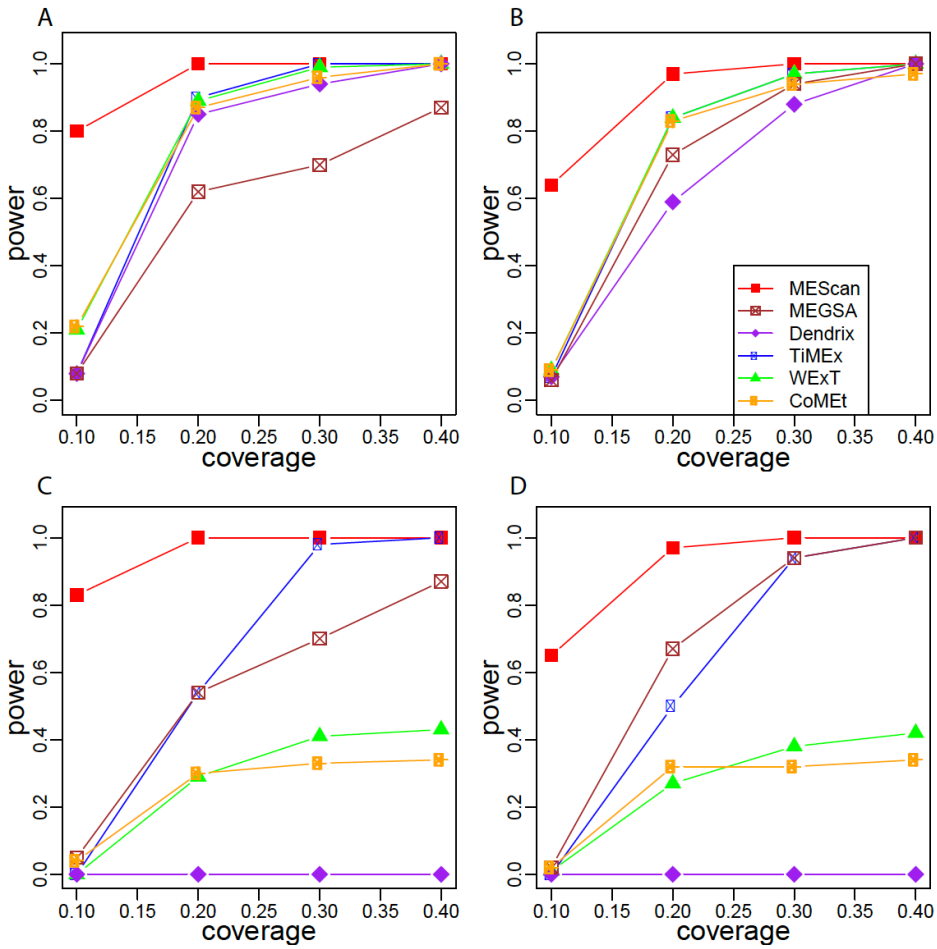


Figure 10. Comparison of power for identifying a true mutually exclusive gene set based on simulations. Each simulated dataset contained 20 genes, including 3 genes with a true mutually exclusive mutational pattern and the other 17 genes without any pattern. Simulations were replicated 100 times and the power was calculated as the frequency that the top-ranked gene set was the 3-gene set with the true mutually exclusive mutational pattern. Four scenarios were considered. A) The ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; B) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; C) the ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes included a highly mutated gene; and D) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes included a highly mutated gene.

Figure 10 shows the simulation results. In all scenarios, MEScan had the highest power compared to other methods, especially when the coverage was low. For example, when a true mutually exclusive pattern with equal number of mutations in each of the three genes was presented in 10% of patients, MEScan was able to achieve 80% power (Figure

10A). In contrast, all other methods had power less than 25%. This is likely due to the adjustment of the background mutation rate by MEScan, which provides a better detection of true patterns against spurious patterns coming from random noise.

3.7.1.2 Simulation studies to evaluate methods' performance in identifying subsets of a true mutually exclusive gene set with the presence of highly mutated genes

We next assessed the impact of a highly mutated noisy gene (TP53) that was not part of the true mutually exclusive pattern. Figure 11. Comparison of power for identifying subsets of a true mutually exclusive gene set based on simulations. Each simulated dataset contained 20 genes, including 3 genes with a true mutually exclusive mutational pattern and the other 17 genes without any pattern. Simulations were replicated 100 times and the power was calculated as the frequency that the top-ranked 2-gene set was a subset of the true 3-gene mutually exclusive mutational pattern. Four scenarios were considered. A) The ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; B) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; C) the ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes included a highly mutated gene; and D) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes included a highly mutated gene. compares the power of each method in the absence (top panels) vs. presence (bottom panels) of TP53. MEScan was able to maintain the power after the addition of the highly mutated gene, indicating that it was robust to such a gene that could cause spurious unbalanced patterns by random chance. In contrast, the power of all other methods decreased. Dendrix did not have any power even when the coverage increased, which is as expected, because it was sensitive to unbalanced spurious patterns.

In fact, the top-ranked gene set from Dendrix was always a set containing TP53. CoMEt, which used a conditional method to reduce the bias towards unbalanced patterns, also had substantial decrease in power. Therefore, the conditional method appeared not adequately address the issue of unbalanced spurious patterns.

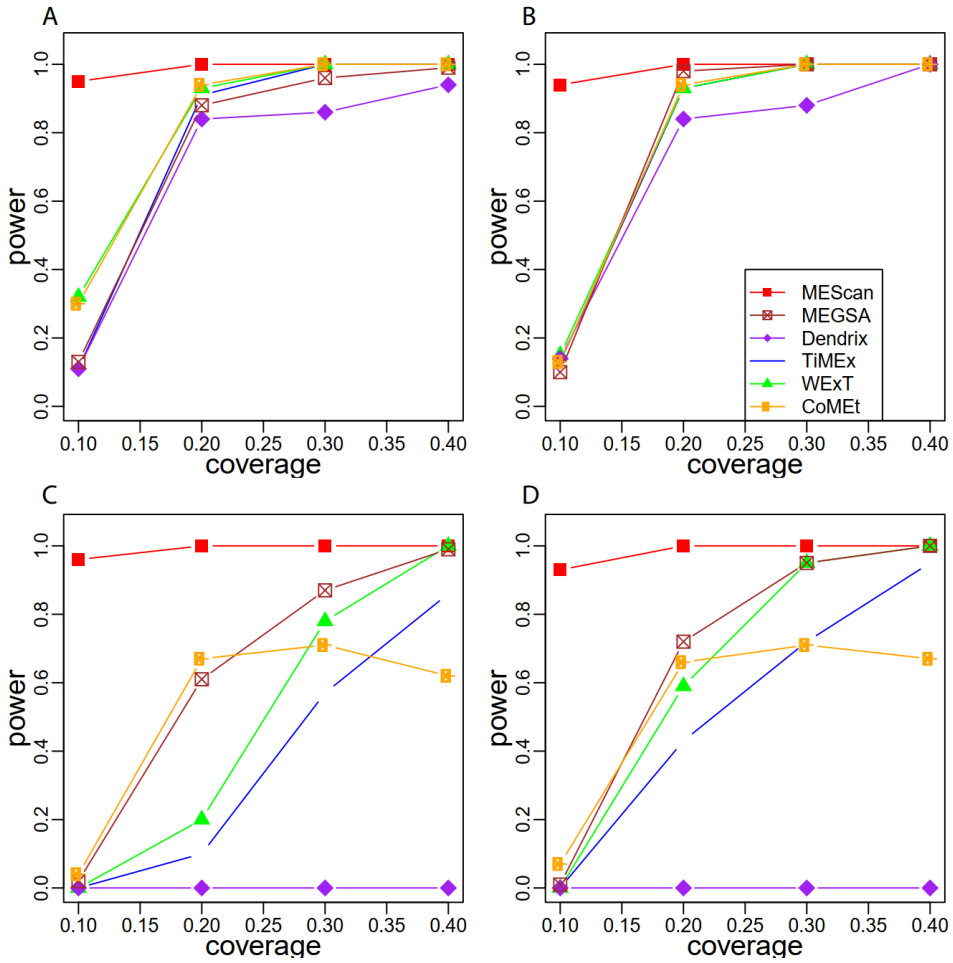


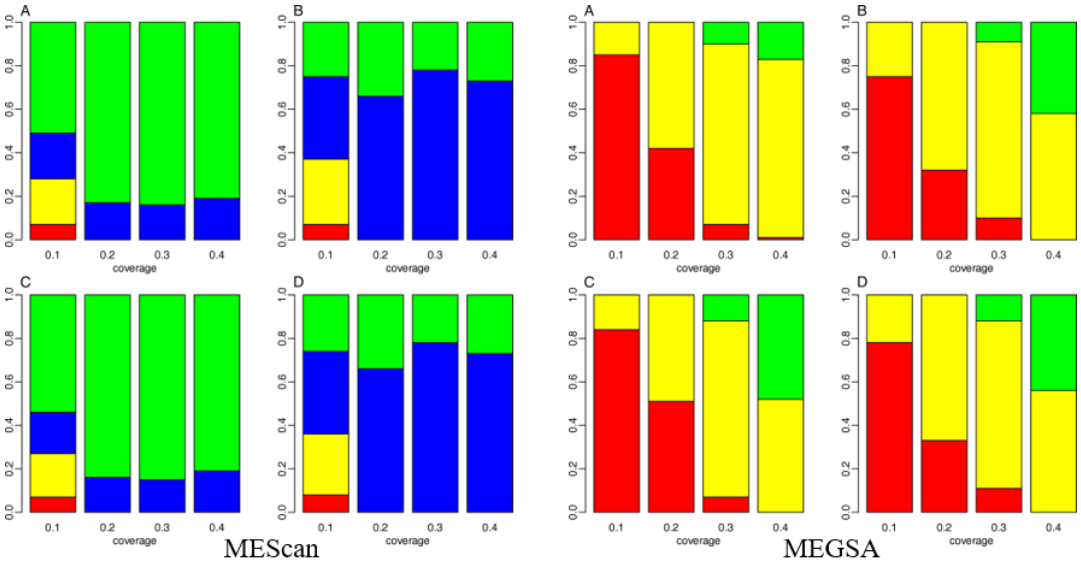
Figure 11. Comparison of power for identifying subsets of a true mutually exclusive gene set based on simulations. Each simulated dataset contained 20 genes, including 3 genes with a true mutually exclusive mutational pattern and the other 17 genes without any pattern. Simulations were replicated 100 times and the power was calculated as the frequency that the top-ranked 2-gene set was a subset of the true 3-gene mutually exclusive mutational pattern. Four scenarios were considered. A) The ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; B) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; C) the ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes included a highly mutated gene; and D) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes included a highly mutated gene.

3.7.1.3 Simulation studies to evaluate methods' performance in identifying the true mutually exclusive gene set across candidate gene sets of different sizes

We adapted the same simulation scenarios and applied each methods to all candidate gene sets of sizes from 2 to 6. We calculated the fraction of simulations that the

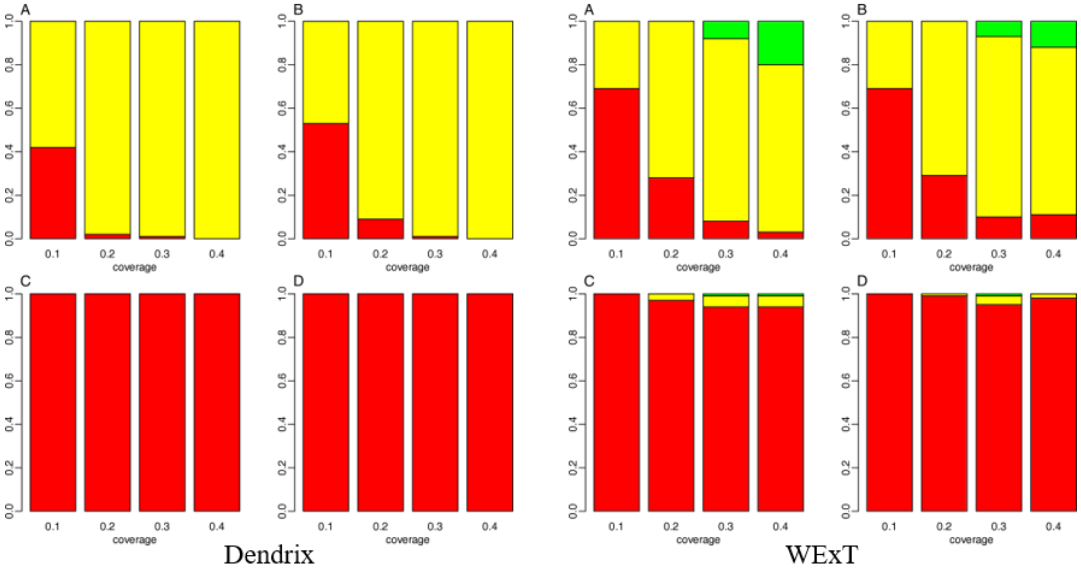
top-ranked gene set was exactly the true 3-gene mutually exclusive set, is a subset of the true set, contains the true set, or otherwise. Compared to other methods, MEScan yielded the highest fraction of simulations with the top-ranked gene set being the true mutually exclusive set when the coverage was low. It was also robust to the presence of a highly mutated noisy gene. Figure 12 Simulation results for applying MEScan, MEGSA, Dendrix, WExT and CoMEt across different sizes (2 to 6) of candidate gene sets. Bar graphs show the fractions of simulations that the top-ranked gene set is exactly the true 3-gene mutually exclusive set (green), is a subset of the true set (blue), contains the true set (yellow), or otherwise (red). Each simulated dataset contained 20 genes, including 3 genes with a true mutually exclusive mutational pattern and the other 17 genes without any pattern. Simulations were replicated 100 times. Four scenarios were considered. A) The ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; B) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; C) the ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes included a highly mutated gene; and D) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes included a highly mutated gene.shows the results for all the methods. For the scenarios that the ratio of mutation frequencies of the three genes in the true set was 1:1:1, MEScan had the largest fraction of simulations that ranked the true 3-gene set to the top among all methods when the coverage was 0.1 to 0.3, while CoMEt had the largest fraction when the coverage was 0.4. For the scenarios that the ratio of mutation frequencies of the three genes in the true set was 3:2:1, MEScan more frequently ranked a 2-gene subset of the true set to the top. This is as expected because one of the three genes had a low mutation

frequency, making it more dicult to be identied. For the purpose of identifying driver genes, MEScan was conservative since all genes in the top-ranked set were part of the true signal. In contrast, other methods tended to more frequently rank larger gene sets, containing some noisy genes in addition to genes in the true set, to the top, which was anti-conservative. In addition, MEScan was still able to identify the true 3-gene set in a fraction of simulations under low coverage situation (coverage = 0.1 or 0.2), where other methods were unable to identify the true set. Furthermore, MEScan's performance remained the same in the absence or presence of a highly mutated noisy gene, suggesting that MEScan was robust to the presence of such a gene. In contrast, the top-ranked gene sets based on Dendrix, WExT and CoMEt were almost always neither the true set nor a superset of the true set in the presence of a highly mutated noisy gene.



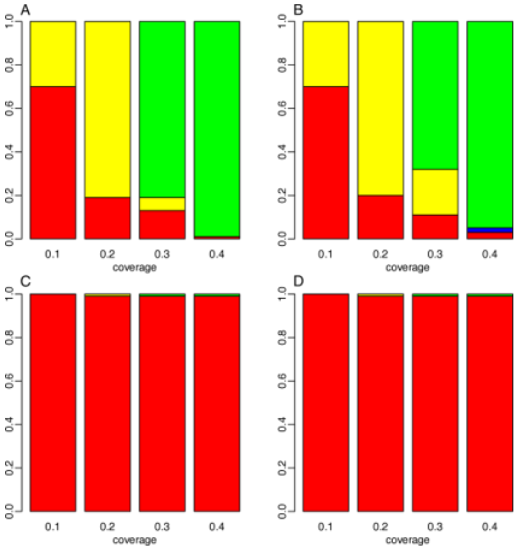
MEScan

MEGSA



Dendrix

WExT



CoMEt

Figure 12 Simulation results for applying MEScan, MEGSA, Dendrix, WExT and CoMEt across different sizes (2 to 6) of candidate gene sets. Bar graphs show the fractions of simulations that the top-ranked gene set is exactly the true 3-gene mutually exclusive set (green), is a subset of the true set (blue), contains the true set (yellow), or otherwise (red). Each simulated dataset contained 20 genes, including 3 genes with a true mutually exclusive mutational pattern and the other 17 genes without any pattern. Simulations were replicated 100 times. Four scenarios were considered. A) The ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; B) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; C) the ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes included a highly mutated gene; and D) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes included a highly mutated gene.

3.7.1.4 Simulation studies to evaluate methods' performance in controlling the false discovery rate (FDR)

We further evaluated the FDR control of our method. In the absence of a highly mutated noisy gene, the observed FDR was around the nominal FDR. In the presence of a highly mutated noisy gene, the observed FDR was smaller than thee nominal FDR. These results suggests that our method was able to control the FDR. We considered the same four simulation scenarios as described in the main text. We investigated all candidate gene sets of size 3, and calculated the observed FDR corresponding to the nominal FDR of 0.01, 0.05, 0.1, and 0.2. Note that the null hypothesis for a mutual exclusivity test is that the three genes do not have any mutually exclusive pattern. The alternative hypothesis is that there is a mutually exclusive pattern, which includes both the case of a full mutually exclusive pattern among all the three genes and the case of a partial mutually exclusive pattern in two of the three genes. Both cases are considered as true positives in our calculation. In our simulations, the full and partial patterns are overlapping, and thus correlated with each other. Figure 12 Simulation results for applying MEScan, MEGSA, Dendrix, WExT and CoMEt across different sizes (2 to 6) of candidate gene sets. compares the nominal FDR versus the observed FDR. In the absence of a highly mutated noisy gene, the observed FDR

was around the nominal FDR. In the presence of a highly mutated noisy gene, the observed FDR was smaller than thee nominal FDR. These results suggests that our method was able to control the FDR.

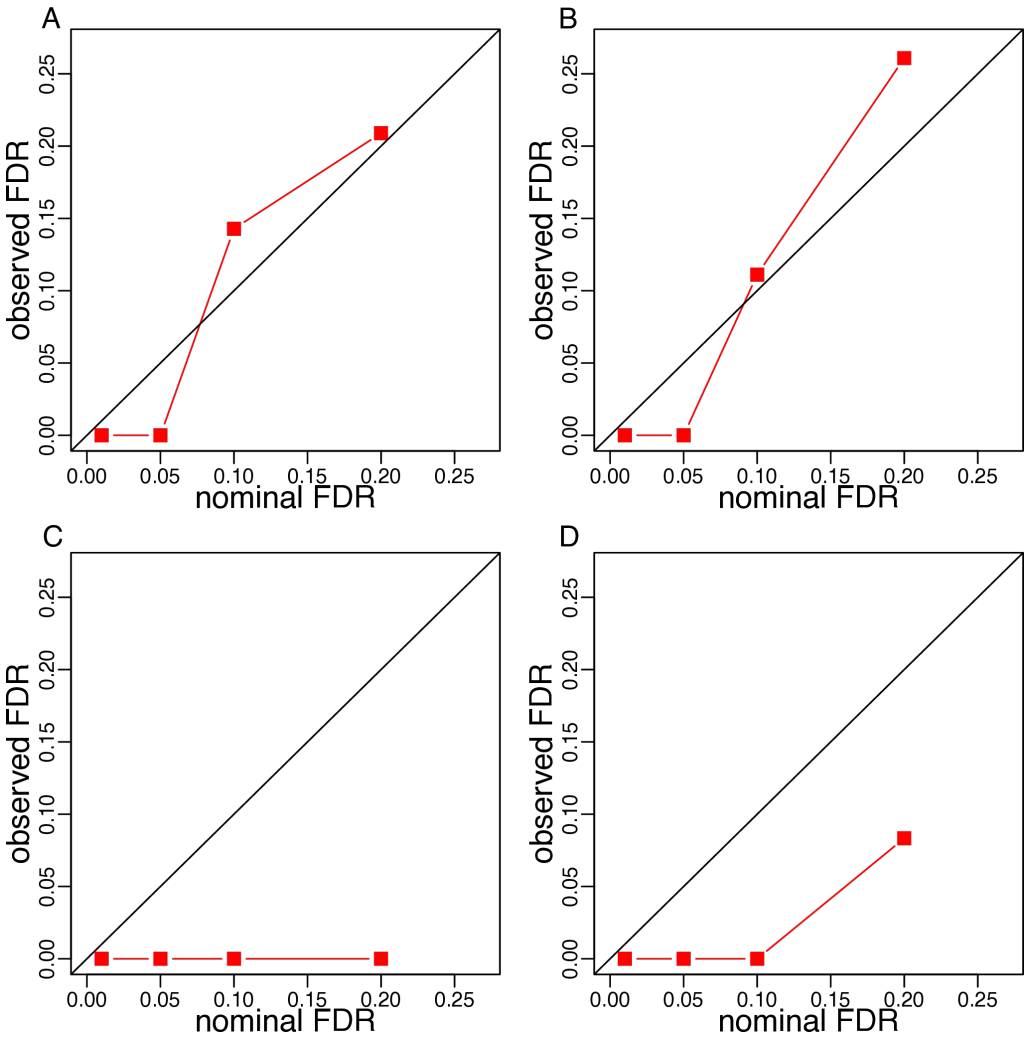


Figure 13 Evaluation of the FDR control based on simulations. Each simulated dataset contained 20 genes, including 3 genes with a true mutually exclusive mutational pattern having a coverage of 0.2 and the other 17 genes without any pattern. Simulations were replicated 100 times. Four scenarios were considered. A) The ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; B) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes did not include a highly mutated gene; C) the ratio of mutation frequencies was 1:1:1 for the 3 genes and the other 17 genes included a highly mutated gene; and D) the ratio of mutation frequencies was 3:2:1 for the 3 genes and the other 17 genes included a highly mutated gene.

Our proposed identification of high-confidence mutually exclusive gene sets provides a way to further select full patterns out of partial patterns. This is because a high-confidence mutually exclusive gene set requires that all its subsets are also significantly mutually exclusive. For a gene set containing a partial pattern, some of its subsets may not contain >1 genes from the mutually exclusive pattern, and thus are likely to be non-significant. Therefore, the identification of high-confidence mutually exclusive gene sets can potentially filter out those gene sets with partial patterns. To demonstrate this, we performed high-confidence mutually exclusive gene sets identification based on the simulated datasets. In all simulations, the true 3-gene set always remained in the high-confidence sets as long as the high-confidence sets was non-empty. For balanced pattern situations (scenarios A and C), in 44% to 50% of simulations, the high-confidence sets only contains a single gene set, which is the true set, suggesting that our method was able to filter out gene sets with partial pattern. For less balanced pattern situations (scenarios B and D), our method was able to identify the single true gene set in 23% to 24% of simulations. The reduced percentage compared to balanced pattern situations was due to the fact that there was a gene with low coverage in the true set under less balanced situations, which was harder to detect. It should also be pointed out that in 20% to 50% of simulations, the resulting high-confidence set was empty, suggesting that the selection of high-confidence set was very stringent so that the true set could sometimes be filtered out. To sum up, the identification of high-confidence mutually exclusive gene sets appears to be a conservative approach to filtering out gene sets with partial patterns.

3.7.2 Time cost comparison

Computational time is very critical for a mutual exclusivity test due to the vast number of candidate gene sets needing to be examined. We compared the computational time of MEScan, MEGSA, Dendrix, WExT and CoMEt for assessing 1000 candidate gene sets for each of the size 3 to 7 based on 200 patients randomly selected from the TCGA ovarian cancer dataset. Note that TiMEx was not included in the comparison, because it was substantially slower than other methods. Table 4 presents the running time of each method. MEScan was the fastest method. The only other method that was on the same scale as MEScan is Dendrix. However, as pointed out by Leiserson et al. (2015) (M. D. Leiserson et al., 2015) and also observed in our simulations, Dendrix did not adjust for the impact of highly mutated genes, and therefore could lead to spurious results. Apart from Dendrix, MEScan was at least two orders of magnitude faster than the rest three methods. For example, it took MEScan only 0.017 seconds to analyze 1000 gene sets of size 3, while the other three methods took more than 8 seconds. In addition, MEScan only had a less than 2-fold increase in computational time as the size of gene set increased from 3 to 7. In contrast, CoMEt and WExT had a 10-fold increase in computational time. Therefore, MEScan provides a very fast and robust test that is instrumental for genome-scale screening of mutually exclusive gene sets.

Table 4. Comparison of computational time. The reported computational time (in seconds) was for analyzing 1000 gene sets of a given size.

Size of gene set	MEScan	MEGSA	Dendrix	WExT	CoMEt
3	0.017	14.604	0.052	8.807	8.488
4	0.021	18.166	0.056	12.791	12.701
5	0.022	26.26	0.06	31.575	23.611
6	0.023	37.285	0.061	50.402	46.922
7	0.023	54.675	0.076	96.574	85.256

3.7.3 Choosing cutoff values of T_G to control FDR

In genome-scale screening, a cutoff value of T_G is determined based on the empirical distribution of T_G (Efron, 2004a) to control for the FDR for each size of gene sets. As it is impractical to obtain T_G for all candidate gene sets, we randomly selected a fraction of gene sets to estimate the empirical distribution of T_G for each gene set size and then determine the cutoff value. To determine how large the fraction is needed to obtain stable cutoff values, Figure 14 compared the cutoff value calculated from 10^7 or 10^8 randomly selected candidate gene sets for gene set sizes 3 to 7 based on the TCGA ovarian cancer dataset. The cutoff value determined using 10^7 gene sets was stable. Increasing the number to 10^8 did not lead to any notable change. As the computational time of calculating T_G's for 10^7 gene sets is acceptable in practice, we used this number in our real data analysis.

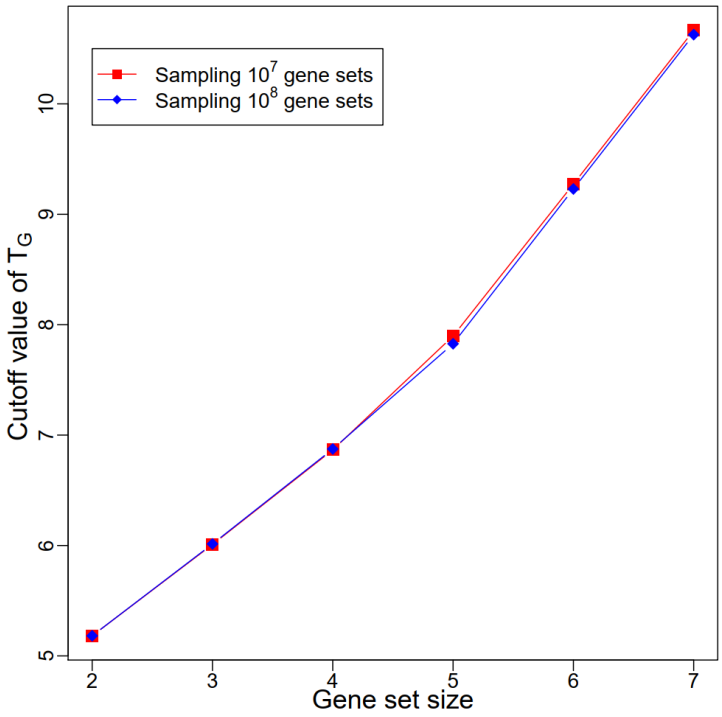


Figure 14. T_G cutoff value estimation. For each size of candidate gene sets, the cutoff value of T_G for controling FDR < 0.05 was estimated by sampling 10^7 (red square) or 10^8 (blue diamond) candidate gene sets.

3.7.4 Whole genome data analysis

We applied our method to TCGA glioblastoma (Brennan et al., 2013), squamous cell lung cancer (Cancer Genome Atlas Research, 2012), ovarian cancer (Cancer Genome Atlas Research, 2011), pan-cancer (Cancer Genome Atlas Research et al., 2013; Kandoth et al., 2013), and breast cancer (Cancer Genome Atlas, 2012) studies. All the data were downloaded from Synapse (syn1729383) (Kandoth et al., 2013). For each dataset, we limited our analysis to non-synonymous mutations and focused on non-synonymous mutations and filtered out genes with no more than one mutation. The filtered datasets contain 3193 to 16984 genes. We applied MEScan and searched for mutual exclusive gene sets of size between 2 and 7. For each gene set size, 4 independent MCMC chains, each having 10^8 iterations with 5×10^5 burn-in iterations, were generated using 4 different

random seeds and the results were pooled together. To control for FDR, we randomly selected 10^7 gene sets of a given size to estimate the empirical null distribution of T_G and FDR. We chose the cutoff value of T_G score such that $\text{FDR} < 0.05$ and called gene sets with T_G scores higher than the cutoff value as significant mutually exclusive gene sets. Finally, high-confidence mutually exclusive sets were determined by investigating the consensus of mutually exclusive gene sets across different sizes. Below, we focused on some of these high-confidence mutually exclusive sets and explored their biological interpretations. The selection of these interesting cases was based on the biological importance and relevance of these gene mutations as well as the clinically actionable mutations of interest.

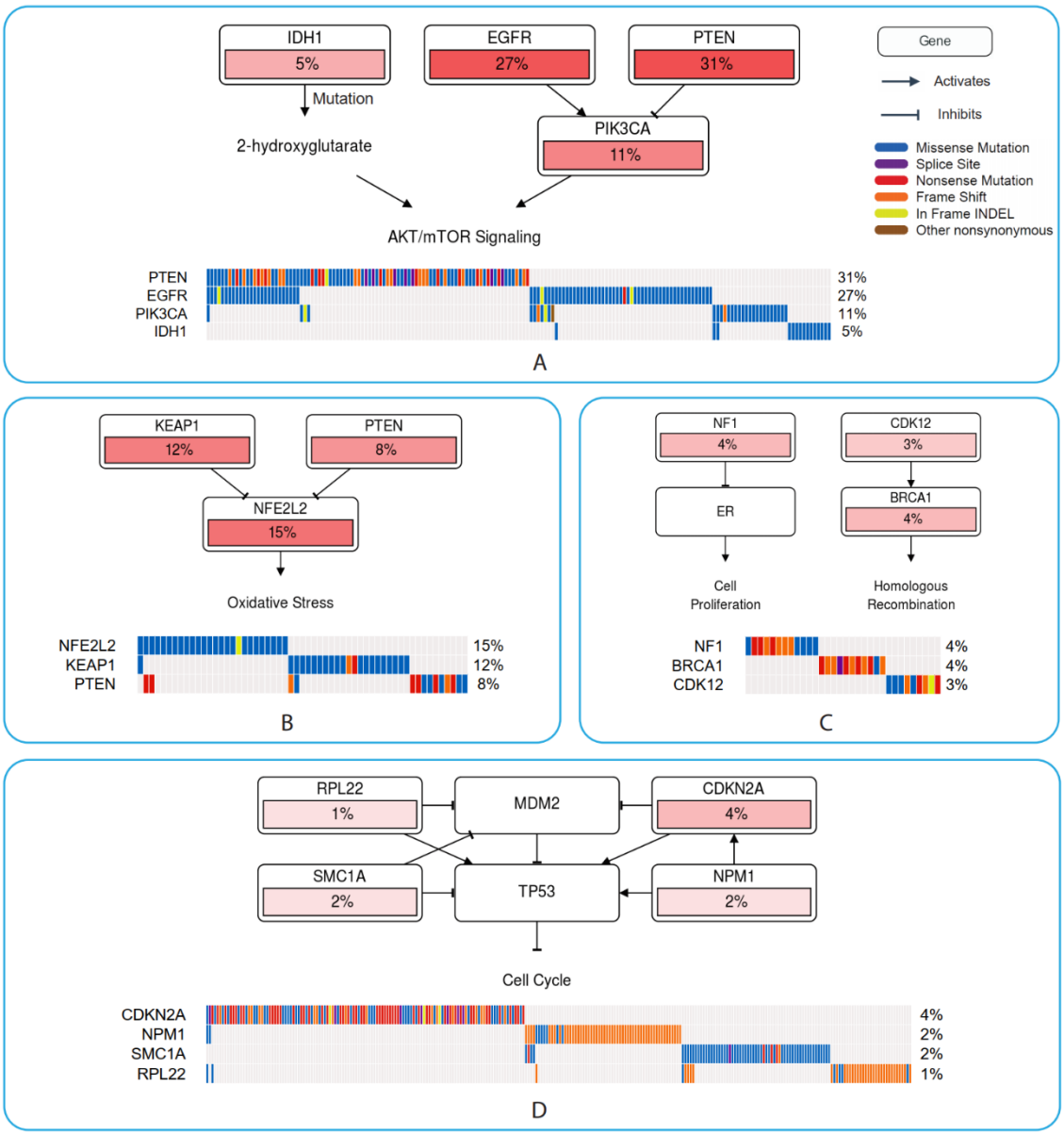


Figure 15. High-confidence mutually exclusive gene sets identified from real data analysis. MEScan was applied to identify high-confidence mutually exclusive gene sets based on A) TCGA glioblastoma (n=290); B) TCGA lung squamous cell carcinoma (n=174); C) TCGA ovarian cancer (n=314); and D) TCGA pan-cancer datasets (n=3205). One selected high-confidence mutually exclusive gene set from each dataset was presented in this figure.

3.7.4.1 Glioblastoma

We identified a high-confidence mutually exclusive set with IDH1, EGFR, PTEN and PIK3CA (Figure 15A). Aberrant PI3K/Akt signaling is frequently observed in cancers

including glioma, which often is achieved by loss of the PI3K inhibitor PTEN (phosphatase and tensin homolog) or gain-of-function mutations of EGFR (X. Li et al., 2016; H. Xu et al., 2017) or PI3KCA (Carracedo & Pandolfi, 2008). IDH1 is a NADP-dependent enzyme that catalyzes the oxidative decarboxylation of isocitrate to α-ketoglutarate (α-KG) in the TCA cycle. IDH mutations are frequently present in gliomas and result in a gain of enzyme function of NADPH-dependent reduction of α-ketoglutarate to 2-hydroxyglutarate that promotes tumorigenesis (Philip et al., 2018). Mutant IDH1 activates mTOR signaling downstream of the PI3K/AKT/TSC1/2 pathway by inhibiting KDM4A, an α-KG-dependent histone demethylase (Carbonneau et al., 2016). Thus, in glioblastoma, we have identified three regulators of PI3K/AKT/mTOR signaling.

3.7.4.2 Squamous cell lung cancer

A high-confidence mutually exclusive set we identified consists of three genes, kelch-like ECH-associated protein 1 (KEAP1), phosphatase and tensin homolog (PTEN) and nuclear factor erythroid-2-related factor 2 (NFE2L2/NRF2), see Figure 15B. NRF2 is a transcription factor and critical regulator of response to oxidative stress. KEAP1 is a negative regulator of NRF2, and in response to oxidative stress, NRF2 is released from KEAP1 where it travels to the nucleus and activates transcription of target genes, that are generally anti-oxidants. When KEAP1 is mutated, NRF2 accumulates (X. Chen, Zhang, Zhang, & Gao, 2019). Constitutive activation of NRF2, either through mutations in NRF2 itself or the regulatory partner KEAP1, is recognized to increase tumorigenesis as well as drive resistance to chemotherapies. In addition, many lung cancers have constitutive NRF2 activation in the absence of NRF2 and KEAP1 mutations (Kerins & Ooi, 2018). PTEN has recently emerged as a negative regulator of NRF2, and loss of PTEN is associated with

constitutive activation of NRF2 (Best et al., 2018). Therefore, our analysis has identified two regulators of a final common transcription factor, strongly implicated in tumorigenesis and resistance to chemotherapy.

3.7.4.3 Ovarian cancer

Breast cancer gene 1 (BRCA1), neurofibromatosis type (NF1) and cyclin-dependent kinase (CDK) 12 was identified as a high probability mutually exclusive gene set in ovarian cancer (Figure 15C). CDK12 is transcriptional regulator of DNA damage response (DDR) genes including those involved in the homologous recombination (HR) like BRCA1 (Joshi, Sutor, Huntoon, & Karnitz, 2014; Paculov\'a & Kohoutek, 2017), via phosphorylation of the RNA polymerase II C-terminal domain. Loss of function mutations of CDK12 result in compromised DDR and homologous recombination, which is observed in ovarian cancers(Joshi et al., 2014). Neurofibromatosis is a hereditary syndrome in which individuals typically develop benign neurofibromas because of neurofibromin 1 (NF1) mutations, but are also at increased risk of breast cancer(Jeon, Kim, Lim, Choi, & Suh, 2015). Recent work demonstrated an association between NF1 deletions and ESR1, the gene for the estrogen receptor (ER) expression and ER positivity (Dischinger et al., 2018). In breast cancer, NF1 binds to and represses ER and loss of function mutations of NF1 activate ER transcriptional pathways (Chang et al., 2018). Like breast cancer, ovarian cancer is a hormone-responsive cancer with ER present in about 60–100% of ovarian cancers (Modugno et al., 2012). It's likely that for ovarian cancers, there are two subtypes. One is driven by mutations in CDK12/BRCA signaling while the other is driven by mutations in NF1/ER signaling. Our gene set analysis has thus identified mechanisms of tumorigenesis of ovarian cancer.

3.7.4.4 Pan-cancer

CDKN2A, NPM1, RPL22, SMC1A was identified as a high-confidence mutually exclusive gene set from TCGA pan-cancer data (Figure 15D). P53 is a well-established tumor suppressor in human cancer. CDKN2A encodes p14ARF, which inhibits MDM2 and promotes p53 function such as cell cycle control, apoptosis and tumor suppression (Sherr, 2006). NPM1 (nucleophosmin) complexes with and stabilizes p14ARF (Sherr, 2006). Mutated NPM1 fails to protect p14ARF from degradation and attenuates the ability of p14ARF to promote p53 function (Colombo et al., 2006) NPM1 also directly interacts with p53 and positively regulates the stability and transcriptional activity of p53 (Colombo, Marine, Danovi, Falini, & Pelicci, 2002). RPL22 (Ribosomal protein L22) is highly mutated in various human cancers. Studies have shown that RPL22 binds with and inhibits MDM2 E3 ligase and thus functions as a p53 positive regulator (Cao et al., 2017). Finally, SMC1A is a component of the cohesin complex that plays a crucial role during mitosis in holding sister chromatids together from DNA replication in S phase to anaphase to ensure proper chromosome separation. SMC1A mutations would impact cohesin functions (Hirano, 2006) and would theoretically result in error-prone chromosome replication and segregation, which may induce p53-mediated cell cycle control, although it has not been experimentally confirmed. The cohesin complex has been shown to bind to the transcription start sites of p53 and mdm2, and the knockdown of Rad21 (a cohesin component) increased their transcription (Rhodes et al., 2010). It is possible that SMC1A mutations would enhance p53 and mdm2 transcription. Thus pan-cancer mutations in CDKN2A, NPM1, RPL22, and SMC1A can be functionally connected through the p14ARF-MDM2-p53 tumor suppressor pathway.

3.7.5 Real world validation and comparison

We used four real data sets to validate our method from different aspects. We also tried to compare our method to existing ones when possible. Note that we attempted to try existing methods on the whole-genome real data examples presented in the last subsection. However, all attempts with those methods failed to finish. Dendrix ran out of computer memory (i.e > 64 GB of RAM). MEGSA, CoMEt and WEXT did not finish after using over 6 days of CPU time. Therefore, we compared our methods to others using a smaller scale real data example presented in the first validation study of this subsection, where some of the existing methods were able to generate results.

Our first validation study considered the pan-cancer data on 299 driver genes from TCGA MC3 (Li Ding et al., 2018) to assess whether MEScan as well as other methods can identify the mutually exclusive patterns reported in the paper. Two sets of analyses were performed. The first set of analyses focused on examining all candidate gene sets of size 2. Ding et.al (Li Ding et al., 2018) reported 8 mutually exclusive gene sets based on the exact Mantel-Haenszel test. MEScan was able to identify all those 8 gene sets as significantly mutually exclusive. In contrast, WExT was able to identify 7, CoMEt was able to identify 1, Dendrix was unable to identify any of those gene sets, and MEGSA was unable to complete the analysis with 6 days of CPU time. The second set of analyses focused on examining candidate gene sets of size >2. Ding et.al (Li Ding et al., 2018) reported 4 such gene sets. MEScan was able to identify all of them as significantly mutually exclusive. In contrast, Dendrix was unable to identify any of those gene sets, and other methods were unable to complete the analysis with 6 days of CPU time. Our second validation study

considered the set of BRAF and NRAS, whose mutations are known to be mutually exclusive in melanoma (Akbani et al., 2015). We applied each method to TCGA melanoma data (Akbani et al., 2015) (n=253). We focused on all gene sets of size 2 and investigated whether a method was able to identify the gene set of BRAF and NRAS. MEScan gave a T_G score of 125.6 for the set of BRAF and NRAS, which is highly significant. In contrast, other methods were unable to finish after using over 6 days of CPU time. Our third validation study used an independent large-scale cohort, PCAWG (Consortium, 2020), to validate the our findings from TCGA pan-cancer analysis. The PCAWG cohort contains 1810 cancer patients after excluding overlapped patients between PCAWG and TCGA. After filtering out a few genes with no observed mutations in the PCAWG cohort, we examined a total of 149 high-confident mutually exclusive gene sets we identified from the TCGA pan-cancer cohort. 95% of those gene set were also significantly mutually exclusive in the PCAWG cohort. Our fourth validation study used an independent cohort of 2,433 primary breast tumors (Pereira et al., 2016) to validate our findings from TCGA breast cancer analysis. Because the validation cohort sequenced a panel of 173 genes, we focused our analysis on high-confidence gene sets consisting of those genes. 84% of those gene sets remained significant in the validation cohort. These four validation studies suggested that MEScan was able to identify known mutually exclusive patterns and provide reproducible results.

3.8 Discussion

We have introduced a statistical framework, MEScan, for accurate and efficient genome-wide *de novo* discovery of mutually exclusive gene sets. Our framework uses a simple yet powerful statistical test for identifying mutually exclusive gene sets. The test

allows adjustment of background mutation rate, mitigates the impact of highly mutated genes, and is very fast to calculate. Coupled with this test is an MCMC algorithm to efficiently screen candidate gene sets at the genomic scale. MEScan is able to search through thousands of candidate genes without restricting to known cancer drivers or genes with high mutation rates. To reduce false positives, we use an FDR control procedure to identify significant gene sets and a summarization method to further select high-confidence mutually exclusive gene sets. Although our method focuses on detecting mutual exclusive patterns, it could potentially be extended to detect other important mutational patterns, such as co-occurrence patterns (Avivar-Valderas et al., 2018; Li Ding et al., 2018; Thomas et al., 2007), but the formula needs to be tweaked towards quantifying those specific patterns. Another important extension of our method is to include somatic copy number variations in the analysis.

We noticed that mutual exclusivity could originate from different mechanisms. The focus of most current research is on mutations of genes from the same biological pathway. However, gene mutations specific to different cancer subtypes could also form a mutually exclusive pattern. For example, from TCGA ovarian cancer data, we identified a high-confidence mutually exclusive gene set of BRCA1, NF1 and CDK12, which is likely to contain two different subtypes of ovarian cancer driven by CDK12/BRCA1 signaling and NF1/ER signaling, respectively. Therefore, a potential new use of mutual exclusivity analysis might be to identify cancer subtypes and subtype-specific gene mutations. Further research in this area will be of great interest. Furthermore, mutually incompatible mutations (e.g. synthetically lethal mutations) would also produce mutual exclusivity in mutation as possible.

One limitation of MEScan is that it does not account for intratumoral heterogeneity. Mutations identified using whole-exome bulk sequencing usually come from a mixture of multiple subclones within a tumor(McGranahan & Swanton, 2017). Delineating the intratumor heterogeneity (Schwartz & Sch\"a, 2017) could provide a cleaner signal to improve mutual exclusivity analysis. Furthermore, recent advances in single-cell sequencing technologies (Zhang et al., 2019) hold the promise of revealing intratumor heterogeneity at a much higher resolution. With the accumulation of such data, performing mutual exclusivity analysis at the single cell level will be an interesting topic for future research.

CHAPTER 4. FastCount: A Fast Gene Count Software for Single Cell and Bulk RNA-seq Data

4.1 Introduction

4.1.1 Bulk RNA-seq gene quantification

RNA sequencing (RNA-seq) has become one of the most commonly used techniques for transcriptome profiling in a wide spectrum of biomedical and biological research. Analyzing RNA-seq reads to quantify expression at each gene locus is the first step towards any downstream biological interpretation.

There are two popular gene expression estimation methods: gene count and transcript abundance. Gene count is essentially the total number of reads sequenced within a gene. Many popular statistical differential expression methods such as DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) take gene count as input. They model it as negative binomial distribution to deal with biological variability and overdispersion and determines differential expression using exact tests (Seyednasrollah et al., 2015). Several tools such as featureCounts (Liao et al., 2014) and HTSeq (Anders et al., 2015) are used to obtain the gene counts. These softwares require several preprocessing steps on the raw RNA-seq reads before performing read counts: 1) generally, a read trimming step is necessary to remove adapter sequences and low-quality bases from the FASTQ files (Bolger et al., 2014; Martin, 2011a). This improves the mappability of the reads during the downstream alignment step. The quality trimming criteria, such as minimum base quality score or a number of bases to be trimmed on start or end of each read, are selected empirically by the users. The processing time ranges from 10 ~ 60 minutes depending on the different algorithms. 2) next, trimmed reads are aligned to either the reference genome or the reference transcriptome using RNA-seq mappers, such as Bowtie2 (Langdon) and

STAR (Bohnert, Vivas, & Jansen), to generate BAM files. 3) Aligned reads in the BAM files are assigned to genes based on the genomic locations provided in the Gene Annotation File (GTF) for gene-level read counts. Although there are some efficient algorithms available, such as featureCounts to summarize read counts from the BAM file, read alignment is computationally intensive, requiring large memory and CPU time. Alternatively, read counts can be derived from transcript abundance using tools specially designed for transcript-level abundance estimation. Transcript expression is first quantified using tools like RSEM (B. Li & Dewey, 2011), Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016) and Salmon (Patro, Duggal, Love, Irizarry, & Kingsford, 2017), followed by additional gene-level expression estimation. RSEM, first performs read mapping using the read aligner mentioned above and uses the Expectation Maximization (EM) algorithm to estimate abundances at the isoform and gene levels. However, RSEM does not scale well due to the high computational requirements. Kallisto and Salmon on the other hand, are alignment-free algorithms where reads are not directly aligned but rather assigned to the most likely transcript that generated them using k-mers. Those methods avoid the time-consuming read alignment steps and report expression abundance (Teng et al., 2016) on transcript levels. Gene-level read counts are estimated using the transcript-level expression by customized scripts or third-party tools to correct gene length changes from differential isoform usage {Soneson, 2015 #162}. Therefore, current methods suffer from the following problems: the accuracy of alignment-based methods depends heavily on the performance of the aligners (Baruzzo et al., 2017) and the scalability is poor in large scale study; assigning reads to transcripts is more challenging than to gene due to the repetitive

sequences among the transcripts (Teng et al., 2016); and no efficient tools that are specially designed to correctly derive gene-level abundance (Soneson et al., 2015).

4.1.2 Single-cell RNA-seq gene quantification

Bulk RNA-seq quantifies the overall transcriptome changes in a collection of cells with the assumption that cells are homogenous within the sample. However, more evidences have shown that (Michael S. Lawrence et al., 2013) (Burrell et al., 2013) tumor cells have highly distinct cell types with each types of cells at different cell states. Bulk RNA-seq averages out the gene expression profile leading to the cell-to-cell variability information unusable (Suva & Tirosh, 2019). The advent of single cell RNA-seq (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) enables scientists to characterize the transcriptomic response of cancer cells under different treatment, understand intratumor expression heterogeneity and infer pseudo-time ordering in cancer development. The increasingly usage of scRNA-seq in the cancer research community necessitates the development of efficient and accurate algorithm to handle the large amount of scRNA-seq data.

The goal of scRNA-seq is to generate abundance × cell expression matrix that can be used for the downstream analyses. Similar to the bulk RNA-seq, the first step in analyzing scRNA-seq data is to assign reads to the reference transcriptome for quantifying gene expression level in each cell. In scRNA-seq , gene counts-based quantification which is the popular approach in bulk RNA-seq analysis (Conesa et al., 2016; Soneson et al., 2015), is largely biased due to cDNA amplification step in the library preparation (Wang & Navin, 2015) leading to distorted estimation of single cells expression level. Recent scRNA-seq protocols (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) have

employed the unique molecular identifiers (UMIs) technique to label the original RNAs before the amplification step to mitigate the bias (W. Chen et al., 2018). Instead of using gene counts, many published statistical methods are focusing on UMI-based count matrix, which is the total number of UMIs associated with each gene, for a more accurate single cell characterization (Butler, Hoffman, Smibert, Papalexi, & Satija, 2018; W. Chen et al., 2018). Several tools have been developed by reusing the bulk RNA-seq quantification methods and taking into account of the UMIs and cell barcodes information incorporated in the scRNA-seq protocols. Similar to bulk RNA-seq, scRNA-seq involves mapping read to the reference genome and assign mapped reads to gene features.

The first challenge in scRNA-seq analysis is the amount of data that need to be processed. A single cell experiment generates $10^6 \sim 10^{10}$ reads for $10^3 \sim 10^6$ cells (Svensson, Vento-Tormo, & Teichmann, 2018). Current methods for scRNA-seq analysis are mainly based on existing bulk RNA-seq tools for read mapping and assignment, with extended functions for UMI and cell barcode processing. Cell Ranger (Zheng et al., 2017), a toolkit developed by the commercialized scRNA-seq company 10X Genomics, is the gold standard to analyzed data generated by 10X Genomics Chromium library. It takes the paired-end FASTQs as input, extracts UMIs and cell barcodes for Read1 and aligns Read2 to the reference genome using STAR (Bohnert et al.) aligner. Customized python scripts are provided for UMI/cell barcode correction, UMI deduplication, and UMI counting. Cell Ranger only counts reads that are confidently mapped to the exonic regions with valid UMIs and cell barcodes. zUMIs (Parekh, Ziegenhain, Vieth, Enard, & Hellmann, 2018) first filters reads with low-quality cell barcodes and UMIs and then mapped the rest of the reads to the genome using STAR (Bohnert et al.) (default setting) or other user-defined

mappers. It uses featureCounts (Liao et al., 2014) to summarize reads mapped to both exon and introns and then output the count table with UMI and cell barcode information to R to summarize read count. UMI-tools (Smith, Heger, & Sudbery, 2017) uses BWA to map reads to the reference transcriptome and only counts exonic reads. However, as read mapping is computationally heavy, the alignment-based methods do not scale well with the large number of cells that may be predicted. For example, Cell Ranger takes 31 hours to process 784M reads in the 8K PMBCs including around 8,000 cells.

Another challenge in scRNA-seq data analysis is the scRNA-seq specific bias including 5'- or 3'- end read bias and low sequencing coverage in many of the droplet-based or well-based protocols. During the library preparation, full-length mRNA sequences are processed for enzymatic fragmentation. It produces transcript fragments with different sizes. However, in the process of PCR amplification, the primers only recognize fragments that contains the oligo sequences added with primer sequences for sequencing. So only the 5' or 3' portion of the transcript is retained after mRNA fragmentation (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017). The alignment-free methods rely heavily on transcript coverage information to infer the likelihood of the transcripts of origin for each read (Bray et al., 2016) (Patro et al., 2017). The resulting library will only produce reads that are mapped to the first or the last exon of the gene, leading to biased gene body coverage (Ma et al., 2019). For transcripts differ primarily on the 5' or 3' end, accurately determination the transcripts of origin is difficult. Additionally, average read coverage from existing scRNA-seq protocols is low, around 50, 000 reads per cell. The low sequencing depth per cell results in reduced sensitivity for resolving transcript-level conflicts. Therefore, the 5'/3'-end read bias as well as the low read coverage in scRNA-

seq make transcript-level quantification even harder due to the reduced information to infer the transcripts of origin in the alignment-free algorithms. Recently benchmarking study has shown that alignment-free algorithms report method-specific artifacts and variances in the number of cells and genes detected. For example, Kallisto detects highly expressed genes in the Vmn and Olfr families in several of the 10X Genomics datasets, where such genes are known to be not expressed in the tissues.

The third challenge in scRNA-seq analysis is to appropriately handle scRNA-seq specific information in the UMI and cell barcode sequences. Due to the small amount of RNA within each cell, a PCR amplification step is necessary to produce enough cDNA for the sequencing step. The amplification bias can be corrected computationally in the downstream analysis using the UMI sequences. If reads with the same UMI sequences from the same cell are mapped to the same transcript sequence, we can conclude that they are technical duplicates from PCR amplification and should be collapsed. However, sequencing errors in the UMIs result in artefactual UMIs inflating the UMI counts (Smith et al., 2017). Existing methods include UMI correction step to identify multiple similar UMI sequences and treat them as instances of the same UMI. Cell Ranger, zUMIs and scPipe (Tian et al., 2018) uses a greedy algorithm comparing each UMI sequence to identify UMIs within certain hamming distance and collapse them to the higher count UMI. UMI-tools links UMIs by a single edit distance and aims to reduce the UMI network into a representative UMI. Salmon constructs a UMI graph to find a minimal set of transcripts for UMI deduplication (Srivastava, Malik, Smith, Sudbery, & Patro, 2019). Kallisto, on the other hand, does not perform UMI correction, and uses a naïve method to collapse reads that contain the same UMI. The sequencing process also introduces errors in cell barcodes.

Therefore, the correction of errors in cell barcodes is important in cell identification. Cell Ranger, zUMIs and Kallisto first compares the sequenced barcodes to a whitelist which contains the pre-defined barcodes in the library preparation kit. If a cell barcode is not in the whitelist but is 1-hamming distance away from a barcode in the whitelist, it will be corrected to the corresponding barcode from the whitelist. Salmon generates a putative whitelist by analyzing the cumulative distribution of barcode frequencies. UMI-tools only selects cell barcodes in the given whitelist. The various combinations of different read mappers, criteria of read assignments and UMI+cell barcode handling have big impact on the gene expression quantification rendering inconsistency in expression differences detection, cell clustering and trajectory analysis (Simonsen et al., 2018; Tian et al., 2019).

4.2 FastCount algorithm

Despite the existence of large number of tools for bulk and scRNA-seq gene quantification, these methods face the following challenges: existing methods do not scale well for large data set especially in the application for the single cell experiment where millions of reads are sequenced for a single sample; the alignment-free methods are sensitive to the sequencing read depth and gene body coverage; and the alignment-free methods provide at least 4 times speed improvement but with the tradeoff for accuracy.

Therefore, we present FastCount, an alignment-free approach to directly assign read to gene-level features. FastCount skips the computationally intensive read mapping step and assign reads directly to genes of origin based on the gene-specific k-mers information. We hypothesize that using gene-specific k-mers information simplifies the gene-level read assignment problem than transcript-specific information and improves the performance in terms of speed and accuracy.

FastCount is capable of alignment-free supports alignment-free UMI count summarization for scRNA-seq data and gene count quantification for bulk RNA-seq data leveraging gene-specific *k*-mers. Different from the alignment-based algorithms that include a computationally intensive read mapping step, FastCount assigns an RNA-seq read directly to the potential gene based on its *k*-mer origin information. FastCount uses a novel data structure, GeneOthello, to assign reads to the corresponding gene identifiers sharing approximately the same set of *k*-mers. It conducts read count in a gene without necessitating detailed read alignment information. FastCount is implemented to handle UMI and cell barcode information specific to scRNA-seq data, allowing a speedy assessment of scRNA-seq cell distribution using raw sequencing reads. We demonstrate through experiment that FastCount scRNA-seq application using 10X Genomics data and compare the performance to 10X Genomics' toolkit, Cell Ranger. FastCount is over an order of magnitude faster than Cell Ranger with very competitive accuracy. We also demonstrate that FastCount is about two orders of magnitude faster than the gold standard bulk RNA-seq tool, RSEM while achieves competitive accuracy.

4.2.1 Gene *k*-mers signatures

The term "gene count" typically refers to the number of reads sequenced in each gene within a given RNA-seq sample. Calculating gene count requires assigning each read to the gene it is sequenced from. Unlike read alignment, read assignment does not require high resolution base-by-base continuous matching between the read sequence and genomic reference. Instead, read assignment can be simplified to the identification of the gene that best matches the set of *k*-mers present in the read. When $k = 21$, the majority of *k*-mers (93.9% in GRCh38) are unique to the genes carrying them (Figure 16). Therefore, the gene-

specific *k*-mers may serve as signatures for read classification problem. The rest of the *k*-mers (6.1% in GRCh38) appear in more than one gene. FastCount leverages this property and turns gene count into a read assignment problem.

The first step of FastCount method is to establish a mapping between *k*-mers and genes. The *k*-mers in RNA-seq are expected to be from mRNA transcriptome although these can be expanded to include other RNA species as well. Therefore, we extract all the *k*-mers present in the transcript sequences of each gene. These *k*-mers S are divided into two categories: gene-specific *k*-mers and gene-clique *k*-mers. Each of the gene-specific *k*-mers can uniquely identify one gene. Each of *k*-mers in a gene clique set is not unique to a gene and instead, is shared by a set of genes, which we refer to as a gene clique. Those *k*-mers are often a result of repetitive sequences but provide useful information on handling reads that map ambiguously to multiple genes. Thus, we would like to establish a mapping between the set of transcriptome *k*-mers and their associated gene features. a). Formally, let G be genes corresponding to gene-specific *k*-mers and C be the set of gene cliques. Let F be a feature set containing genes in G and gene cliques in C, where $F = G \cup C$ and $G \cap C = 0$ Thus, there exists a many-to-one mapping between $S \longrightarrow F$, such that $m(s) = F$.

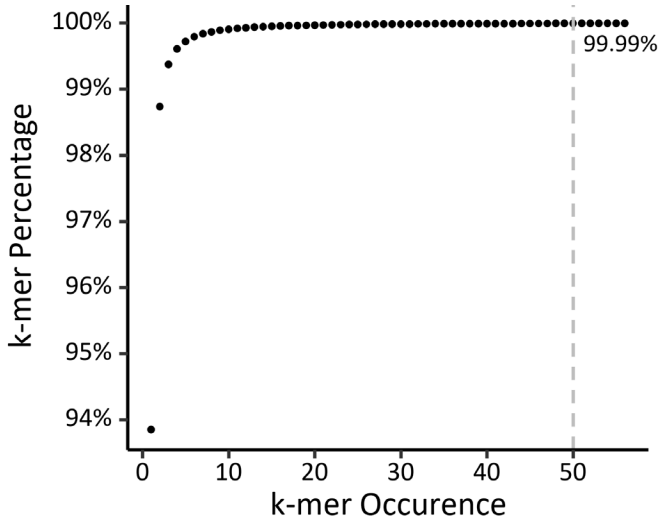


Figure 16. Cumulative k-mer percentage at different k-mer occurrence from Human reference genome GRCh38.

4.2.2 GeneOthello k-mers index

The efficiency of gene count procedure relies on how fast one can quickly map a read to a gene. In our case, it is translated to how fast we can map a *k*-mer to a gene. To this end, we have adopted a hashing classifier, called Othello (Yu, Belazzougui, Qian, & Zhang, 2018), to facilitate the mapping between *k*-mers and the features S. Othello is a minimal perfect hashing algorithm (MPH) that provides key-to-value query in constant time. The othello algorithm has demonstrated great scalability in both memory and query speed in several Bioinformatics applications (Liu et al., 2018; Yu, Liu, et al., 2018).

We build an index, named GeneOthello, in order to store the many-to-one mapping between the aforementioned transcriptome *k*-mers and gene features. We encode gene features F as a set of l-bit integers $V = \{0, 1, 2, \dots, |\mathrm{G}|, |\mathrm{G}|+1, |\mathrm{G}|+2, \dots, |\mathrm{G}|+|\mathrm{C}|\}$, where 0 is specially allocated for *k*-mers with occurrence > n and $l = \lceil log\_2(|F|+1) \rceil$. GeneOthello $\mathrm{O}(\mathrm{S}, \mathrm{V})$ maps the predefined set of *k*-mers S to the gene features in V. Let $\mathrm{T}: \mathrm{S} \rightarrow \mathrm{V}$ be the function that maps *k*-mers to the gene features in V, where T(k) indicates the

feature of a *k*-mer $s \in S$. GeneOthello maintains a query function $\tau: U \rightarrow I$ that maps the universal set of *k*-mers, U (i.e., $U = \theta^k, \theta = \{A, G, C, T\}$) to the set of all *l*-bit integers, $I = \{0,1, \ldots, 2^l - 1\}$. Therefore, $S \subset U$ and $V \subset I$. GeneOthello has the following properties: 1) given a *k*-mer $s \in S$, a GeneOthello querying $\tau(s)$ guarantees returning the correct gene feature id v; 2) for any alien *k*-mers, $s' \notin S$, GeneOthello has a higher probability to assign s' to a dummy feature $\tau(s') \in I - V$ than a false positive feature $\tau(s') \in V$ (Yu, Liu, et al., 2018).

The GeneOthello structure includes a pair of hash function $\langle h_a, h_b \rangle$ and two arrays of *l*-bit integers A and B. The hash functions and the content of the integer arrays are precomputed during GeneOthello construction. The hash functions provide the mapping between the universal *k*-mer set and the corresponding index location in A and B, that is $h_a: U \rightarrow \{0,1, \ldots, m_a - 1\}$ and $h_b: U \rightarrow \{0,1, \ldots, m_b - 1\}$. A query of a *k*-mer s on GeneOthello will first yield indices $h_a(s)$ and $h_b(s)$. The feature $\tau(s)$ of *k*-mer s is computed by the integer values at $A[h_a(s)]$ and $B[h_b(s)]$ as $\tau(s) = A[h_a(s)] \oplus B[h_b(s)]$. The procedure only accesses two memory locations in A and B and a bitwise XOR operation, making the query execution extremely fast.

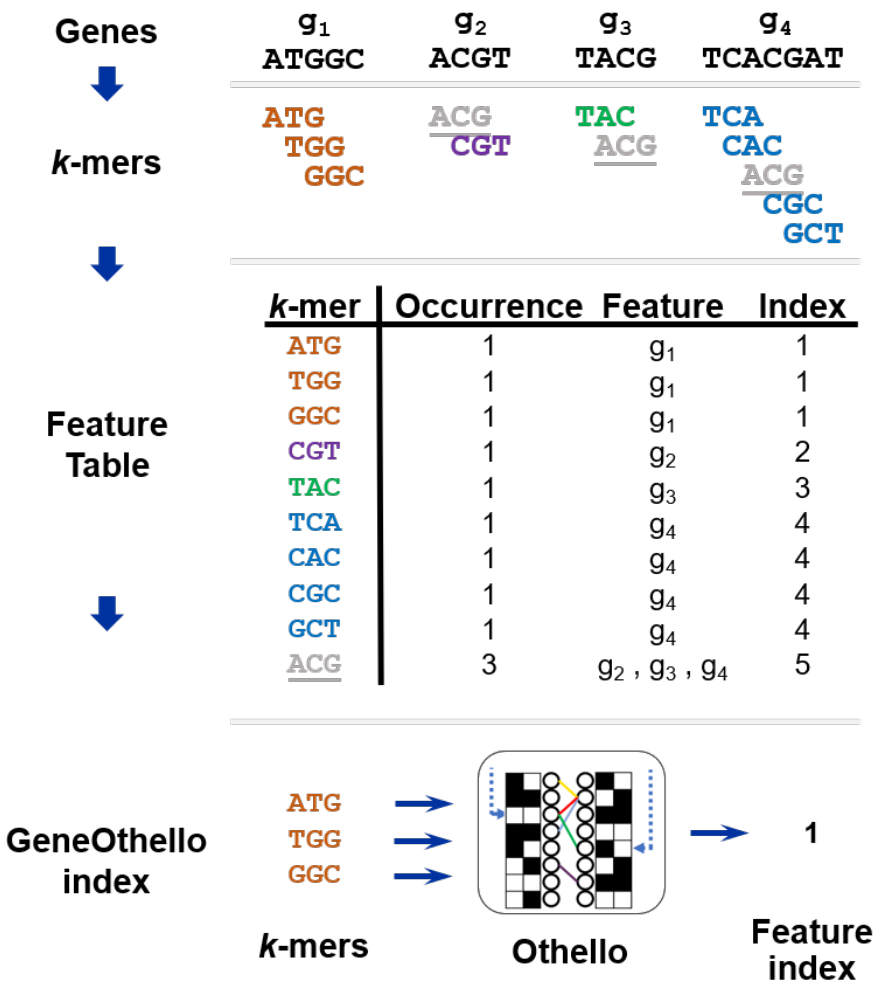


Figure 17 A toy example illustrating FastCount algorithm. The construction of GeneOthello index from the reference gene sequences. In this example, *k*-mers ($k = 3$) in the reference set are extracted from each reference transcript. Their occurrences within genes are tabulated and are categorized into gene signatures and cliques (Top). A GeneOthello index is constructed to store the many to one mapping between the *k*-mers and the gene feature indices (Bottom).

4.2.3 Read assignment to genes

One of the key steps in bulk and single cell RNA-seq analysis is to accurately assign read to the correct gene feature. To do this, FastCount first decomposes read into consecutive *k*-mers (Figure 18 Read assignment procedure. *k-mers* in a read are extracted and queried against GeneOthello to obtain their gene features. Continuous *k-mers* with the same gene feature are clustered into feature windows. The gene assignment of a read is

determined by the gene that dominates the longest feature window.) and each *k*-mer is queried against the GeneOthello index to retrieve its gene feature. Ideally, the read assignment is straightforward when all *k*-mers pointing to the same gene. However, this process is often complicated by the presence of alien *k*-mers that are absent from the reference transcriptome. Alien *k*-mers can be a result of system artifacts during the RNA-seq library preparation and sequencing, contamination or novel transcript isoforms (Levy et al., 2007; Taub, Corrada Bravo, & Irizarry, 2010). When querying on alien *k*-mers, majority of them can be easily detected using the dummy features returned by GeneOthello. But in the cases that an alien *k*-mer is falsely allocated to a reference gene feature, failure of detecting it may lead to a false positive read assignment.

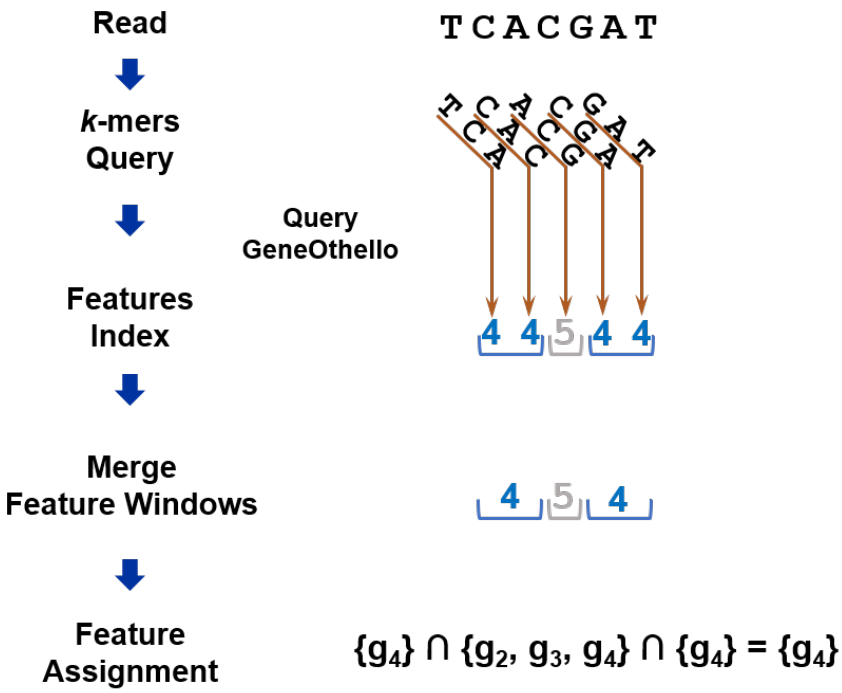


Figure 18 Read assignment procedure. *k*-mers in a read are extracted and queried against GeneOthello to obtain their gene features. Continuous *k*-mers with the same gene feature are clustered into feature windows. The gene assignment of a read is determined by the gene that dominates the longest feature window.

FastCount tackles the challenge of alien *k*-mers using the following strategies: 1) when constructing GeneOthello index, we intentionally expand the size of I such that many alien *k*-mers are categorized into a group that is not overlapping with any valid transcriptome features I - V. Given the predefined V, we can increase the probability of an alien *k*-mer falls in I - V by increasing *l*, that is

$$\frac{2^l - |V|}{2^l}$$

(Liu et al., 2018). 2) in the case that an alien *k*-mer is falsely assigned to the existing gene features, we require the presence of at least two consecutive *k*-mers returning the same gene feature to be considered as a feature segment for gene assignment. As the indices

queried by alien *k*-mers are random, the chance of seeing the same gene assignment from the two different alien *k*-mers is very low (Liu et al., 2018).

FastCount performs read assignment with the following procedure: 1) it iterates through each *k*-mer along the read to retrieve its gene feature by querying against the GeneOthello index. With the presence of alien *k*-mers, gene features along the reads may be divided into section. We merge continuous gene features into feature segments and reject the dummy features from alien *k*-mers. Each feature segment is weighted by the $w_j = d_j^2$, where d_j is the number of *k*-mers in the feature segment j, to take into account of the *k*-mer spatial information on the read. 2) We filter feature segments with length < 2 from feature assignment to remove the possible alien *k*-mers query falsely return to the known features. 3) Gene segments are then ranked by the weights to identify the most dominant feature. 4) The rest of the feature segments are compared to the dominant feature. If there are no consensus gene feature among the segments, the read is removed from the assignment; if they all indicate a single consensus gene, the read is assigned to the gene feature; if they indicate to a set of genes, the case of multi-mapping read, we will not include this read for feature assignment.

4.2.4 FastCount scRNA-seq implementation

In scRNA-seq analysis, another key step is to appropriately handle the single-cell specific information in the cell barcode and UMI sequence. FastCount accepts sequencing FASTQ files that are expected to contain cell barcode and UMI information in one of the read files and the gene identity file in another. For example, a typical read 10X Genomics Chromium data is in the form of read pairs, read1 and read2 (Figure 19 Individual barcode identification and gene assignment for a paired-end read.). Read1 is the cell barcode

sequence that is used to tag the cell origin of the cDNA followed by the UMI sequence which labels the cDNA molecule. Read2 is the sequenced nucleotide bases from the cDNA segment used to determine the gene identity. FastCount algorithm integrates read pre-filtering, read assignment, UMI correction and UMI counting steps.

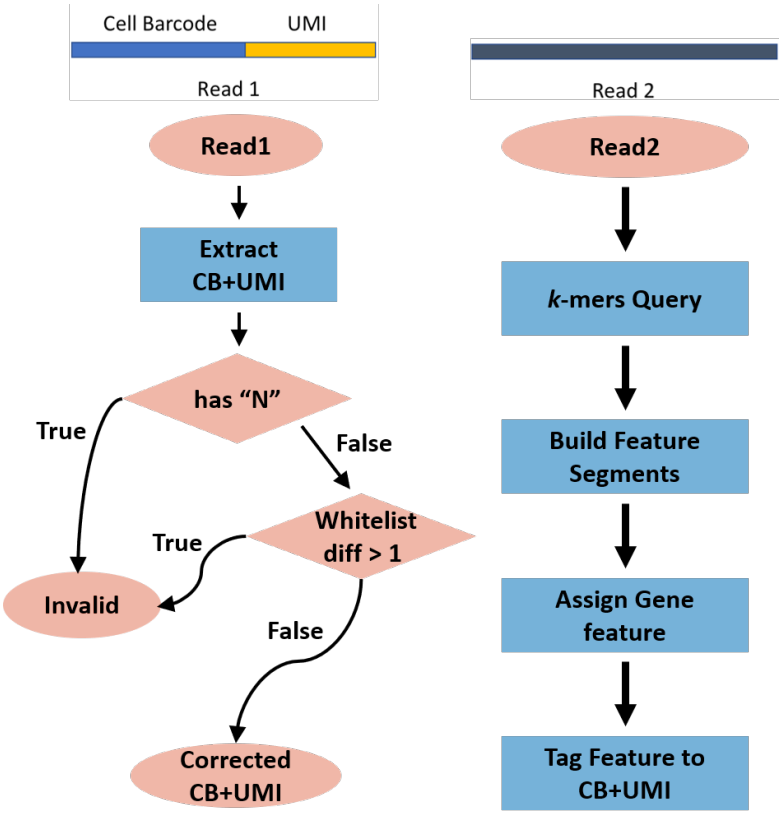


Figure 19 Individual barcode identification and gene assignment for a paired-end read. FastCount first removes read pairs with invalid CELL BARCODE or UMI sequences using read1. If read1 corresponds to valid CELL BARCODE and UMI barcodes, it proceeds to assign read2 to gene features using FastCount read assignment algorithm.

FastCount first filters the low-quality cell barcode and UMI sequences. Due to potential sequencing and amplification errors in the cell barcode and UMI sequences, the number of detected cells and UMI counts are usually inflated. FastCount parses the cell barcode and UMI from the read1 and looking for any "N" bases. Read pairs with 'N's in either cell barcode or UMI will be skipped for the read assignment step. Cell

barcode+UMIs pass this filter will be further compared against the cell barcode whitelist which is the full list of all known cell barcode sequences that are available for cell tagging during the library preparation. Cell barcodes that more than 1-hamming distance apart from the whitelist are considered invalid. Otherwise, the reads with valid cell barcode+UMI sequences are assigned to its potential gene features based on its *k*-mers information using the FastCount read assignment algorithm.

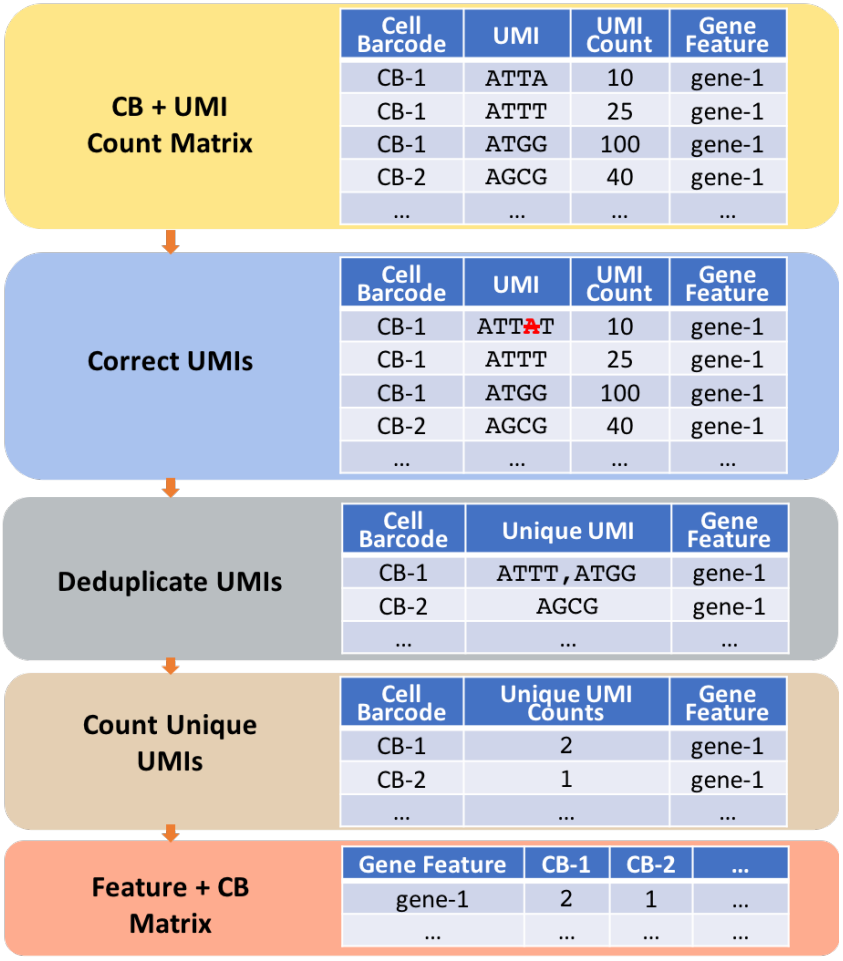


Figure 20 Cell-level feature summarization. Our method iterates through all the read pairs for gene-level assignment and keeps track of the number of reads assigned to each gene feature for individual cells using a CELL BARCODE+UMI counts matrix. It then corrects UMIs based on the UMI counts and UMI sequence differences for each gene feature. The final feature-barcode matrix summarizes the number of unique UMIs for each gene feature in each cell.

FastCount keeps tracking the association between the UMIs and gene features for each individual cell and stores the UMI counts in a cell barcode+UMI counts matrix. After iterating through all the reads for the feature assignment, FastCount will try to further mitigate errors in UMIs using the cell barcode+UMI counts information (Figure 20 Cell-level feature summarization. Our method iterates through all the read pairs for gene-level assignment and keeps track of the number of reads assigned to each gene feature for

individual cells using a CELL BARCODE+UMI counts matrix. It then corrects UMIs based on the UMI counts and UMI sequence differences for each gene feature. The final feature-barcode matrix summarizes the number of unique UMIs for each gene feature in each cell.). The UMI correction is performed per cell per gene feature. FastCount compares one UMI against the rest of UMIs under the same gene feature within each cell looking for the ones that are 1-hamming distance apart. An UMI with lower count is corrected towards the higher count UMI. The UMI corrected counts matrix then undergoes UMI deduplication to group duplicate UMIs and collapse them into a single consensus one. Finally, FastCount counts the number of unique UMIs for each gene-level feature in each individual cell and reports the feature-barcode matrix in the Market Exchange Format for downstream analysis.

4.3 Experimental results

We assess the performance of FastCount on bulk RNA-seq and scRNA-seq data analysis separately. In order to benchmark its performance on bulk RNA-seq data, we generate a set of simulated data and compare the gene count reported by FastCount to the ground truth. For scRNA-seq, datasets published by 10X Genomics are selected as reference sets. The feature-barcode matrix quantified by the standard 10X data analysis pipeline, CellRanger, is used as the ground truth to evaluate the accuracy measurement.

4.3.1 Bulk RNA-seq simulation datasets

We benchmark the performance of FastCount on bulk RNA-seq using simulated data generated by *rsem-simulate-reads* program from RSEM (B. Li & Dewey, 2011). The parameterization of the dataset is learned from real data sets following the procedure used in Kallisto(Bray et al., 2016). Specifically, *rsem-calculate-expression* was run on

NA12716_7 from the GEUVADIS RNA-seq data to learn the model parameters from the real data. We simulated 20 different data sets with 30 million 75 bp paired-ends reads using *rsem-simulate-reads*. The gene-level expression estimations reported by this program for each simulation set are used as the true expression levels. We calculated the pearsons and spearmans correlation values between the estimated abundance and the ground truth. We further measure the accuracy using median relative difference (MRD) (Bray et al., 2016) and 5% error fraction (EF) (B. Li & Dewey, 2011) statistics.

4.3.2 scRNA-seq 10X Genomics datasets

Due to the lack of ground truth scRNA-seq dataset, our evaluation focuses on the scRNA-seq data produced using the widely accepted 10X Genomics Chromium platform. The filtered feature-barcode matrix quantified by *count* function in the Cell Ranger pipeline is used as the reference dataset. We compare how well each tested method correlates with Cell Ranger's UMI counting results. We include six datasets to exam the performance of the different tools in terms of different species, tissue types, sequencing depth and library chemistry. Datasets pbmc_1k_v3 and pbmc_10k_v3 were derived from human peripheral blood mononuclear cells (PBMCs) and were prepared using the 10X Chromium v3 chemistry. We also include heart_1k_v3 and heart_10k_v3 which are cells from whole heart of an E18 mouse. Additionally, pbmc4k and pbmc8k are chosen to evaluate the performance on data generated by 10X Chromium v2 chemistry. The number of cells in these datasets ranges from ~ 1,000 to ~ 8,000 per sample. They were processed by 10X Genomics using the standard Cell Ranger pipeline. The raw sequencing FASTQ files are analyzed by each tool using the recommended settings. The resulting feature-barcode matrices are compared with the reference datasets quantified by Cell Ranger *count* to

evaluate the concordance in gene- and cell-level expression as well as in the downstream biological inference.

4.3.3 Comparison with other bulk RNA-seq tools

We evaluate FastCount's performance in terms of accuracy and scalability on analyzing traditional bulk RNA-seq data. We compare its performance with three popularly used gene count pipelines: STAR (Dobin et al., 2013) in quantMode, featureCounts (Liao et al., 2014) (STAR aligner) and the gold standard RSEM (B. Li & Dewey, 2011) (STAR aligner).

STAR was run in 1-pass mode with "quantMode" option enable to summarize the number of reads assigned per gene during the mapping procedure. featureCounts was run on the BAM file generated from STAR output and the General Transfer Format (GTF) file to summarize the read counts on the gene level features. RSEM was run with STAR aligner and the gene level expression estimates was used for the comparison. All the tools are ran using their default settings.

For each of the simulated data, we quantify the gene abundances with the four tested methods and measure the accuracy of gene count estimates. Table 5 Accuracy of gene count quantification in terms of Pearson and Spearman correlation, MRD and 5% EF using simulated data. shows the median values of the Pearson, Spearman, MRD and 5% EF for each of the tested methods using the 20 simulated bulk RNA-seq data. In general, FastCount achieves a competitive and sometimes slightly better performance over the set of gene count tools in comparison. FastCount outperforms STAR 1-PASS and featureCount+STAR pipelines. One major reason for the lower performance in the two STAR related methods is likely due to that STAR and fetureCount do not handle read

mapped to multiple gene features. Such reads are dropped from gene counting. RSEM, on the other hand, uses the Expectation Maximization (EM) algorithm to estimate the abundance based on the uniquely aligned reads and multi-mapped reads. In FastCount, multi-mapped reads are allocated to genes in proportion to the gene count quantified using uniquely-mapped reads.

Table 5 Accuracy of gene count quantification in terms of Pearson and Spearman correlation, MRD and 5% EF using simulated data.

Pipeline	Pearson	Spearman	MRD	5% EF
FastCount	1.000	0.980	0.002	12.7
RSEM+STAR	1.000	0.986	0.039	13.4
featureCounts+STAR	0.998	0.956	0.060	83.2
STAR 1-PASS	0.996	0.939	0.091	89.5

We then measure the speed and memory usage of FastCount on bulk RNA-seq. All the pipelines are run on a 16-core Intel Xeon E5-2670 @ 2.60 GHZ Linux server with 64 GB of RAM using 10 threads. FastCount is approximately an order of magnitude faster than STAR 1-PASS and featureCounts+STAR and is about 2 orders of magnitude faster than RSEM+STAR (Figure 21 Speed and memory usage of FastCount and 3 other pipelines : RSEM+STAR, featureCounts+STAR and STAR 1-PASS mode on the 20 simulated data sets. (a) Total runtime for processing all the 20 simulation datasets. (b) Median peak memory requirement.). Additionally, FastCount requires a maximum of 1.3 GB of memory while any other pipelines requiring about 30GB memory due to the dependency of STAR alignment.

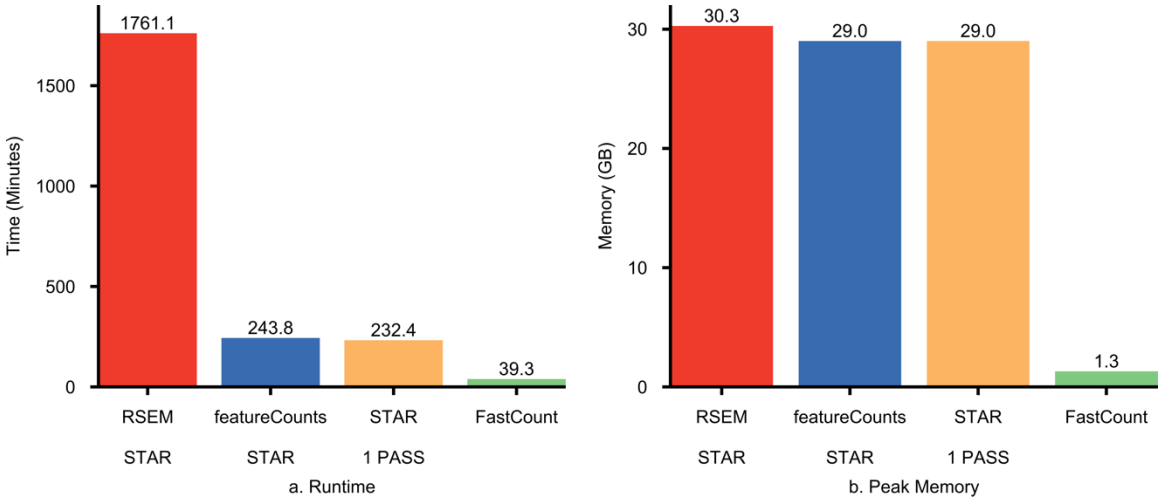


Figure 21 Speed and memory usage of FastCount and 3 other pipelines : RSEM+STAR, featureCounts+STAR and STAR 1-PASS mode on the 20 simulated data sets. (a) Total runtime for processing all the 20 simulation datasets. (b) Median peak memory requirement.

4.3.4 Comparison with other scRNA-seq pipelines

We assess the accuracy of FastCount in scRNA-seq quantification and compare its performance to Cell Ranger, Salmon-Alevin (Srivastava et al., 2019) and Kallisto (Bray et al., 2016). To make the evaluation comparable to Cell Ranger, we download the reference packages provided by 10X Genomics for human and mouse samples. We then subset the GENCODE (Frankish et al., 2019) transcript sequence files for either species (human GRCh38.p12 release 30, mouse GRCm38.p6 release M21) with the same set of transcripts used by CellRanger. The common transcripts are used to construct the reference indices for FastCount, Kallisto and Salmon-Alevin.

Table 6 UMI count concordance between different methods and Cell Ranger in terms of median Pearson, Spearman and MRD for the 6 datasets.

Datasets	FastCount			Alevin			Kallisto		
	Pearson	***Spearman***	***MRD***	***Pearson***	***Spearman***	***MRD***	***Pearson***	***Spearman***	***MRD***
pbmc_1k_v3	0.999	0.975	0.029	0.996	0.951	0.057	0.998	0.924	0.096
pbmc_10k_v3	0.999	0.973	0.033	0.996	0.948	0.060	0.998	0.920	0.100
heart_1k_v3	0.998	0.980	0.023	0.989	0.952	0.066	0.992	0.936	0.080
heart_10k_v3	0.996	0.977	0.031	0.988	0.954	0.062	0.992	0.943	0.067
pbmc4k	0.999	0.989	0.022	0.993	0.972	0.050	0.993	0.935	0.113
pbmc8k	0.999	0.987	0.024	0.991	0.970	0.051	0.994	0.940	0.100

Each of the six datasets is processed using default settings by the tested algorithms to generate the raw feature-barcode matrix. We then subset this matrix to keep the same set of features and cells reported by Cell Ranger. The filtered matrix is then compared against the reference Cell Ranger results.

We first calculate the gene-level correlation of the UMI counts for each gene across all cells between the tested algorithms and Cell Ranger *count* using the pearson's correlation, spearman's correlation and median relative difference (MRD) statistics per cell (Table 6 UMI count concordance between different methods and Cell Ranger in terms of median Pearson, Spearman and MRD for the 6 datasets.). The results indicate that FastCount shows the highest degree of agreement to the reference results in all datasets (median pearson's correlation = 0.998, spearman's correlation = 0.973 and MRD = 0.033). In comparison, Alevin and Kallisto show a less degree of concordance with cell ranger especially with scRNA-seq from heart tissues.

We next examine the correlation on cell-level expression in total UMI counts between the tested algorithms and Cell Ranger. We calculate the total mRNA abundance within each cell by the sum of the UMI counts for each expressed gene in a cell. The correspondence in the total UMI counts per cell between the tested methods and Cell

Ranger is presented in Figure 22 The scatter plots of the total number of UMI counts per cell. We see high agreements between the three alignment-free methods and Cell Ranger with points centered around the diagonal line in the scatter plots. FastCount has the highest degree of concordance in all the six tested datasets, while the total UMI counts by Alevin and Kallisto show larger variations from the Cell Ranger reference results. Kallisto consistently overestimates the total gene expression level in the cells. We also observe that many low-expressed cells in the two mouse heart tissue datasets (heart_1k_v3 and heart_10k_v3) are missing in Alevin while they are all identified by FastCount, Kallisto and Cell Ranger *count*. These findings agree with the benchmark paper reported by (Brüning, Tombor, Schulz, Dimmeler, & John, 2021).

We further assess the cell-level concordance by calculating the pearson's correlation in the UMI count of the same cell between the tested methods and Cell Ranger. We plot the density of the correlation coefficient as a function of total UMI counts per cell in Figure 23 Pearson correlation of UMI counts within each cell as a function of total UMI counts per cell. Each point represents a cell and the points are colored based on the number of neighboring points indicating the overall distribution of the correlation. The pearson's correlation in the cell-level UMI counts between FastCount and Cell Ranger clustered into a much narrower range and most of the cells have correlation coefficiency close to 1. In contrast, many cells show relatively lower correlation in Alevin and Kallisto indicated by the much more spread-out points on the graphs.

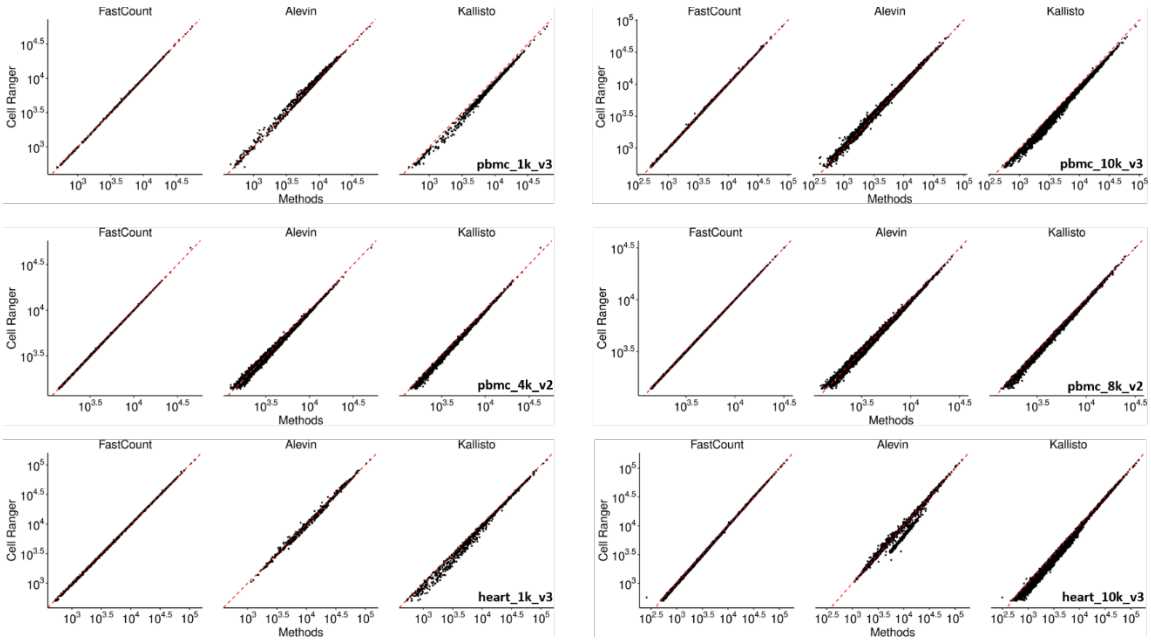


Figure 22 The scatter plots of the total number of UMI counts per cell between Cell Ranger and each of the tested algorithms (FastCount, Salmon and Kallisto) from the six 10X Genomics scRNA-seq datasets.

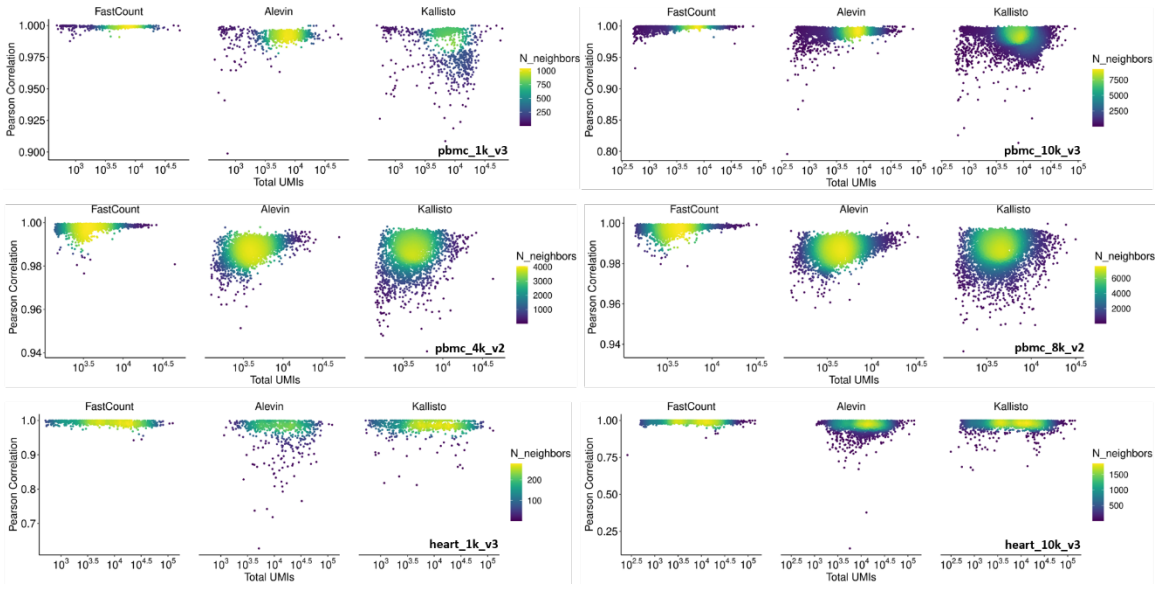


Figure 23 Pearson correlation of UMI counts within each cell as a function of total UMI counts per cell using the six 10X Genomics scRNA-seq datasets.

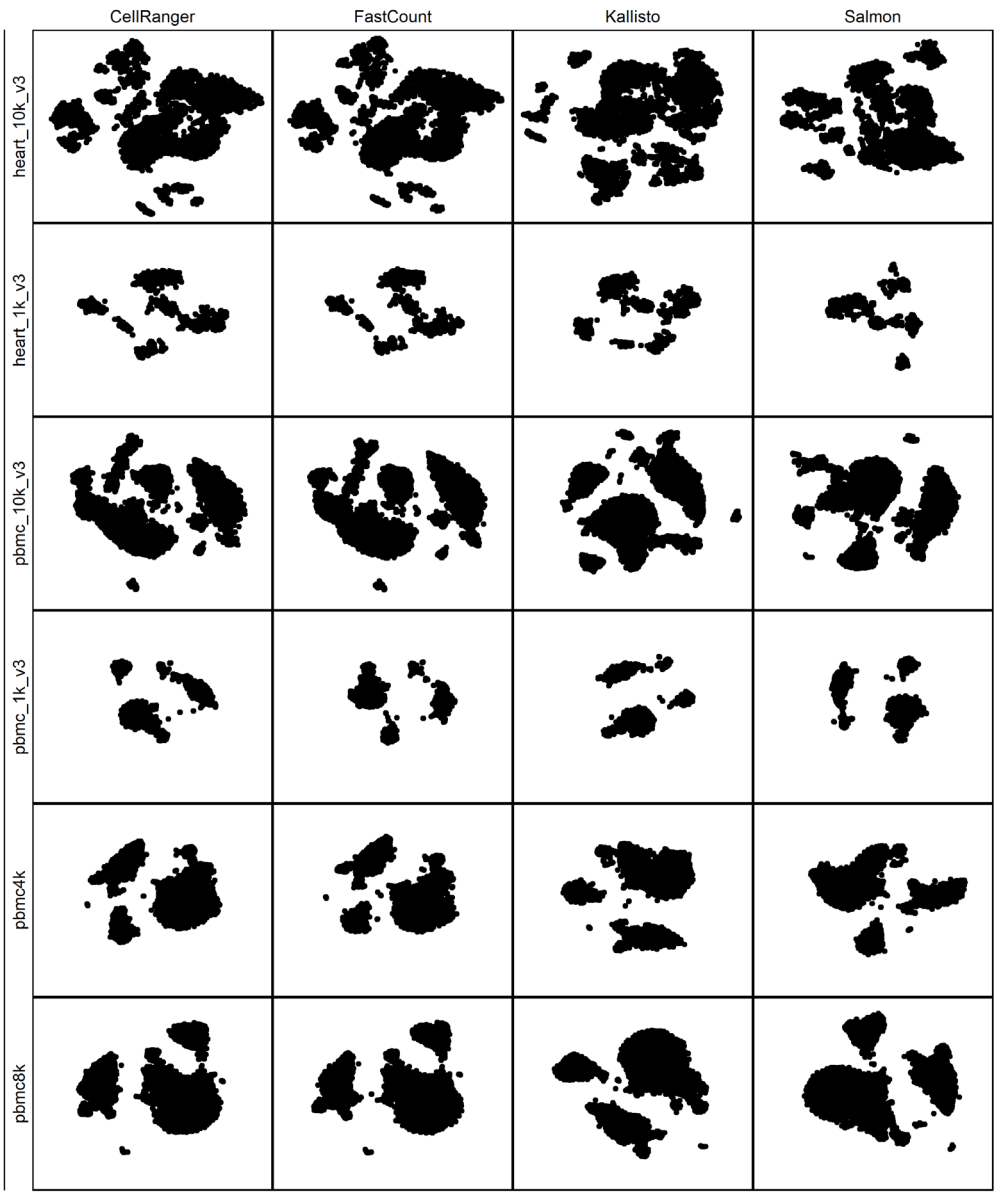


Figure 24 Compatible t-SNE plots using the feature-barcode matrices generated by Cell Ranger, FastCount, Kallisto and Salmon for the six 10X Genomics scRNA-seq datasets.

Lastly, to evaluate the the impact of the different cell gene count tools on downstream cell clustering analysis. We perform dimensionality reduction using t-Distributed Stochastic Neighbor Embedding (t-SNE) for each dataset using scater package {McCarthy, 2017 #752}. Figure 24 Compatible t-SNE plots using the feature-barcode matrices shows that the t-SNE plots from Cell Ranger and FastCount are very similar

across all six data sets, retaining all the clustering structures and embedding shapes. In contrast, the distribution of clusters varies significantly using Alevins and Kallisto cell count results. The difference becomes more apparent when the composition of clusters is complex.

4.3.5 Runtime comparison

We benchmark the runtime used for each algorithm using the six testing datasets. All methods are evaluated on a 16-core Intel Xeon E5-2670 @ 2.60 GHZ Linux server with 64 GB of RAM using 10 threads. The result is shown in Table 7 Runtime comparison on 10X Genomics single cell gene expression datasets in different tools. In general, FastCount is at least 1.5 orders of magnitude faster than Cell Ranger *count*. For example, FastCount used only 4.1 minutes to process 1K PBMCs data set which contains around 1000 cells and 66 million of reads, while Cell Ranger took 154.2 minutes. For 8K PBMCs which contains around 8000 cells and 784 million of reads, FastCount used only 46.5 minutes, while Cell Ranger needed 31 hours. FastCount is around 3 times faster than Alevin except for the 2 mouse heart samples. This is because Alevin pre-filtered many (25.6% in 1K Heart, 26.8% in 10K Heart) cells before the read assignment and UMI quantification steps which largely reduced the number of reads required for processing. This pre-filtering step resulted in many cells missing from the final expression matrix. Kallisto is the fastest algorithm but the runtime between FastCount and Kallisto is quite close.

Table 7 Runtime comparison on 10X Genomics single cell gene expression datasets in different tools

Dataset	Chemistry	Species	Total Reads	Runtime (Minutes)			
				FastCount	Cell Ranger	Alevin	Kallisto
1K PBMCs	V3	Human	66,601,887	4.1	154.2	11.9	3.0
10K PBMCs	V3	Human	638,901,019	38.9	1563.5	121.7	26.4
4K PBMCs	V2	Human	379,462,522	22.2	914.0	66.7	16.9
8K PBMCs	V2	Human	784,064,148	46.5	1865.3	132.6	34.4
1K Heart	V3	Mouse	84,512,390	5.1	186.2	7.4	3.6
10K Heart	V3	Mouse	290,439,571	21.6	674.5	26.4	12.1

4.4 Conclusion

In this chapter, we present FastCount, a novel alignment-free approach to quantify gene count in individual cells with scRNA-seq data. Comparing to the other alignment-free algorithms, FastCount shows a higher degree of concordance with Cell Ranger's *cellranger count* in both cell-level and gene-level expression, in the meantime provides 1.5 orders of magnitude speed improvement over Cell Ranger.

FastCount assigns sequencing reads directly to genes using gene-specific and gene-clique *k*-mers. It fully utilizes the cell barcode and UMI information in the sequencing reads to computationally allocate reads to the cells of origin while identifying technical artifacts from PCR amplification. We evaluate the accuracy of FastCount on scRNA-seq quantification using the results reported by Cell Ranger as the reference data sets. We observe that different quantification pipeline implementation influences the single cell experiment qualities in terms of genes/cells expression level and the downstream analysis.

The read processing step of scRNA-seq data is the first and most fundamental step during the scRNA-seq analysis pipeline. However, the commonly used Cell Ranger *count* often requires many hours of CPU times and extensive resources that are only available to computational servers. Being capable of processing 1,000 cells under 5 minutes on a regular laptop, Fastcount provides a lightweight but fast alternative, making it possible to combine cell count together with downstream analysis such as Seruat or Scanpy in one pipeline.

CHAPTER 5. Conclusion

Cancer research is in the genomics era. Large-scale cancer genomics research has been focusing on comprehensive characterization of the cancer landscape using massive production of genomic, transcriptomic, epigenomic, and proteomic data. Bioinformatics methods have become one of the key components in cancer study that transform the sequencing reads generated by various Next Generation Sequencing assays into information that is interpretable by the biologist. Accurate analysis of NGS data is a critical step upon which virtually all downstream interpretation process relies. Despite an active research field in the past decade, existing methods are still facing the challenge of high complexity and weakness in performance. This dissertation presents three novel computational methods for developing robust and reproducible NGS pipeline platform, efficient genome-wide driver mutation identification algorithm and alignment-free quantification algorithm for bulk RNA-seq and single cell RNA-seq .

In the second chapter of the dissertation, I developed a robust, reproducible and scalable Bioinformatics pipeline framework that streamlines the DNA sequencing analysis workflow for cancer genomic mutation identification. It automates the best practice mutation calling pipelines to detect somatic single nucleotide polymorphisms, indels and copy number variation from DNA sequencing data and perform various downstream analyses. It integrates mutation annotation, clinically actionable therapy prediction and data visualization that simplifies the sequence-to-report data transformation. It has been applied to the real-world data that characterizes the genomic landscape of squamous cell lung cancer from Appalachian Kentucky using whole exome sequencing data. It is the first sequencing effort that provides an overview of the somatic alterations and copy-number

variations, explores unique mutational patterns comparing to the TCGA cohorts, and indicates clinically actionable assessment of mutations in this population.

Large-scale sequencing efforts identify thousands of somatic mutations from cancer samples. Differentiating the driver mutations among a vast pool of passenger mutations is an important but challenging task in cancer research. In the third chapter, I developed MEScan, which is one of the first method that enables genome-scale driver mutations identification based on mutual exclusivity test using cancer mutation data. MEScan implements an efficient statistical framework to *de novo* screen mutual exclusive patterns and in the meantime taking into patient-specific and gene-specific background mutation rate and adjusting the heterogenous mutation frequency. It outperforms several existing methods based on simulation studies and is at least 2-fold of magnitudes faster than most of the tools. MEScan implements a Markov chain Monte Carlo (MCMC) algorithm to efficiently scan for mutually exclusive gene sets at the genomic scale, a false discovery rate (FDR) adjustment procedure to control false positives, and a summarization procedure to select high-confidence findings. Genome-wide screening using existing TCGA somatic mutation data discovers novel cancer-specific and pan-cancer mutually exclusive patterns.

Recent advances in scRNA-seq technology allows cancer biologists to appreciate heterogeneous gene expression changes on the cellular level. scRNA-seq data processing using current methods is computationally challenging due to the volume of the data generated in a single cell experiment as well as the single cell specific information carried in the reads. In the fourth chapter, I designed and implemented a light-weight RNA-seq read classification algorithm based on gene-specific *k*-mers in the transcriptome. It

provides comparable accuracy to the current gold standard algorithm, but achieves around two orders of magnitude speed improvement. I further extended this read classification algorithm for single cell RNA-seq quantification by incorporating the cell barcode and UMI information. It quantifies 800 Million reads of 8,000 cells in less than 50 minutes which is over an order-of-magnitude faster than the classic 10X Genomics Cell Ranger (31 hours) workflow while providing competitive accuracy in terms of UMI counts, cell clustering and differentially expressed genes.

Cancer research has become increasingly data centric and relies heavily on the big data generated from various sequencing assays to interrogate genomic changes using multi-Omics data. Cancer Bioinformatics research has become a key component in supporting the cutting-edge cancer research. Therefore, novel method development in Cancer Bioinformatics continues to be in high demand driven by novel biomedical applications. The works presented in this dissertation are expected to resolve some of the fundamental challenges faced by cancer researchers in analyzing large-scale cancer genomics data.

REFERENCES

Adebali, O., Reznik, A. O., Ory, D. S., & Zhulin, I. B. (2016). Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. *Genet Med, 18*(10), 1029-1036. doi:10.1038/gim.2015.208

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., . . . others. (2015). Genomic classification of cutaneous melanoma. *Cell, 161*(7), 1681-1696.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166-169. doi:10.1093/bioinformatics/btu638

Arita, A., Shamy, M. Y., Chervona, Y., Clancy, H. A., Sun, H., Hall, M. N., . . . Costa, M. (2012). The effect of exposure to carcinogenic metals on histone tail modifications and gene expression in human subjects. *J Trace Elem Med Biol, 26*(2-3), 174-178. doi:10.1016/j.jtemb.2012.03.012

Avivar-Valderas, A., McEwen, R., Taheri-Ghahfarokhi, A., Carnevalli, L. S., Hardaker, E. L., Maresca, M., . . . Cruzalegui, F. (2018). Functional significance of co-occurring mutations in PIK3CA and MAP3K1 in breast cancer. *Oncotarget, 9*(30), 21444.

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods, 14*(2), 135-139. doi:10.1038/nmeth.4106

Best, S. A., De Souza, D. P., Kersbergen, A., Policheni, A. N., Dayalan, S., Tull, D., . . . others. (2018). Synergy between the KEAP1/NRF2 and PI3K pathways drives non-small-cell lung cancer with an altered immune microenvironment. *Cell metabolism, 27*(4), 935-943.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., . . . Canaider, S. (2013). An estimation of the number of cells in the human body. *Ann Hum Biol, 40*(6), 463-471. doi:10.3109/03014460.2013.807878

Biterge, B., Richter, F., Mittler, G., & Schneider, R. (2014). Methylation of histone H4 at aspartate 24 by protein L-isoaspartate O-methyltransferase (PCMT1) links histone modifications with protein homeostasis. *Sci Rep, 4*, 6674. doi:10.1038/srep06674

Bohnert, R., Vivas, S., & Jansen, G. (2017). Comprehensive benchmarking of SNV callers for highly admixed tumor data. *PLoS One, 12*(10), e0186175. doi:10.1371/journal.pone.0186175

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol, 34*(5), 525-527. doi:10.1038/nbt.3519

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., . . . Network, T. R. (2013). The somatic genomic landscape of glioblastoma. *Cell, 155*(2), 462-477. doi:10.1016/j.cell.2013.09.034

Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., & John, D. (2021). Comparative Analysis of common alignment tools for single cell RNA sequencing. *bioRxiv*, 2021.2002.2015.430948. doi:10.1101/2021.02.15.430948

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature, 501*(7467), 338-345. doi:10.1038/nature12625

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology, 36*(5), 411-420. doi:10.1038/nbt.4096

Cairns, R. A., & Mak, T. W. (2013). Oncogenic isocitrate dehydrogenase mutations: mechanisms, models, and clinical opportunities. *Cancer Discov, 3*(7), 730-741. doi:10.1158/2159-8290.CD-13-0083

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., . . . Meyerson, M. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet, 48*(6), 607-616. doi:10.1038/ng.3564

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70. doi:10.1038/nature11412

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*(7353), 609-615. doi:10.1038/nature10166

Cancer Genome Atlas Research, N. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature, 489*(7417), 519-525. doi:10.1038/nature11404

Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., . . . Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet, 45*(10), 1113-1120. doi:10.1038/ng.2764

Cao, B., Fang, Z., Liao, P., Zhou, X., Xiong, J., Zeng, S., & Lu, H. (2017). Cancer-mutated ribosome protein L22 (RPL22/eL22) suppresses cancer cell survival by blocking p53-MDM2 circuit. *Oncotarget, 8*(53), 90651.

Carbonneau, M. e., lissa, Gagn\'e, Laurence M., Lalonde, M.-E., Germain, M.-A., Motorina, A., Guiot, M.-C., . . . others. (2016). The oncometabolite 2-hydroxyglutarate activates the mTOR signalling pathway. *Nat Commun, 7*, 12700.

Carracedo, A., & Pandolfi, P. P. (2008). The PTEN--PI3K pathway: of feedbacks and cross-talks. *Oncogene, 27*(41), 5527.

Chakravarty, D., Gao, J., Phillips, S. M., Kundra, R., Zhang, H., Wang, J., . . . Schultz, N. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol, 2017*. doi:10.1200/PO.17.00011

Chang, E. C., Anurag, M., Gao, J., Cakar, B., Du, X., Li, J., . . . others. (2018). NF1 as an estrogen receptor-$$ co-repressor in breast cancer. In.

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., & Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol, 19*(1), 70. doi:10.1186/s13059-018-1438-9

Chen, X., Zhang, X.-l., Zhang, G.-h., & Gao, Y.-f. (2019). Artesunate promotes Th1 differentiation from CD4+ T cells to enhance cell apoptosis in ovarian cancer via miR-142. *Brazilian Journal of Medical and Biological Research, 52*(5).

Chervona, Y., Arita, A., & Costa, M. (2012). Carcinogenic metals and the epigenome: understanding the effect of nickel, arsenic, and chromium. *Metallomics, 4*(7), 619-627. doi:10.1039/c2mt20033c

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., . . . Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotech, 31*(3), 213-219. doi:10.1038/nbt.2514 http://www.nature.com/nbt/journal/v31/n3/abs/nbt.2514.html#supplementary-information

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., . . . Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol, 31*(3), 213-219. doi:10.1038/nbt.2514

Colombo, E., Marine, J.-C., Danovi, D., Falini, B., & Pelicci, P. G. (2002). Nucleophosmin regulates the stability and transcriptional activity of p53. *Nature cell biology, 4*(7), 529.

Colombo, E., Martinelli, P., Zamponi, R., Shing, D. C., Bonetti, P., Luzi, L., . . . others. (2006). Delocalization and destabilization of the Arf tumor suppressor by the leukemia-associated NPM mutant. *Cancer research, 66*(6), 3044-3050.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol, 17*, 13. doi:10.1186/s13059-016-0881-8

Consortium, I. T. P.-C. A. o. W. G. (2020). Pan-cancer analysis of whole genomes. *Nature, 578*(7793), 82.

Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenfuhrer, J., & Beerenwinkel, N. (2016). TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics, 32*(7), 968-975. doi:10.1093/bioinformatics/btv400

Dang, L., White, D. W., Gross, S., Bennett, B. D., Bittinger, M. A., Driggers, E. M., . . . Su, S. M. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature, 462*(7274), 739-744. doi:10.1038/nature08617

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., . . . others. (2018). Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell, 173*(2), 305-320.

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., . . . Cancer Genome Atlas Research, N. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell, 173*(2), 305-320 e310. doi:10.1016/j.cell.2018.03.033

Dischinger, P. S., Tovar, E. A., Essenburg, C. J., Madaj, Z. B., Gardner, E. E., Callaghan, M. E., . . . others. (2018). NF1 deficiency correlates with estrogen receptor signaling and diminished survival in breast cancer. *NPJ breast cancer, 4*(1), 29.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21.

Efron, B. (2004a). Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association, 99*(465), 96-104. doi:10.1198/016214504000000089

Efron, B. (2004b). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association, 99*(465), 96-104.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., . . . Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res, 43*(Database issue), D805-811. doi:10.1093/nar/gku1075

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., . . . Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res, 47*(D1), D766-d773. doi:10.1093/nar/gky955

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., . . . Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer, 4*(3), 177-183. doi:10.1038/nrc1299

Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol, 11*(8), R86. doi:10.1186/gb-2010-11-8-r86

Guo, Y., Li, J., Li, C. I., Long, J., Samuels, D. C., & Shyr, Y. (2012). The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics, 13*, 666. doi:10.1186/1471-2164-13-666

Halvorsen, A. R., Silwal-Pandit, L., Meza-Zepeda, L. A., Vodak, D., Vu, P., Sagerup, C., . . . Helland, A. (2016). TP53 Mutation Spectrum in Smokers and Never Smoking Lung Cancer Patients. *Front Genet, 7*, 85. doi:10.3389/fgene.2016.00085

Hamamoto, R., Saloura, V., & Nakamura, Y. (2015). Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nat Rev Cancer, 15*(2), 110-124. doi:10.1038/nrc3884

Hirano, T. (2006). At the heart of the chromosome: SMC proteins in action. *Nature reviews Molecular cell biology, 7*(5), 311.

Hua, X., Hyland, P. L., Huang, J., Song, L., Zhu, B., Caporaso, N. E., . . . Shi, J. (2016). MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. *The American Journal of Human Genetics, 98*(3), 442-455.

Hua, X., Hyland, P. L., Huang, J., Song, L., Zhu, B., Caporaso, N. E., . . . Shi, J. (2016). MEGSA: A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations. *Am J Hum Genet, 98*(3), 442-455. doi:10.1016/j.ajhg.2015.12.021

International Cancer Genome, C., Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., . . . Yang, H. (2010). International network of cancer genome projects. *Nature, 464*(7291), 993-998. doi:10.1038/nature08987

Jeon, Y. W., Kim, R. M., Lim, S. T., Choi, H. J., & Suh, Y. J. (2015). Early-onset breast cancer in a family with neurofibromatosis type 1 associated with a germline mutation in BRCA1. *Journal of breast cancer, 18*(1), 97-100.

Joberty, G., Boesche, M., Brown, J. A., Eberhard, D., Garton, N. S., Humphreys, P. G., . . . Drewes, G. (2016). Interrogating the Druggability of the 2-Oxoglutarate-Dependent Dioxygenase Target Class by Chemical Proteomics. *ACS Chem Biol, 11*(7), 2002-2010. doi:10.1021/acschembio.6b00080

Johnson, N., Shelton, B. J., Hopenhayn, C., Tucker, T. T., Unrine, J. M., Huang, B., . . . Li, L. (2011). Concentrations of arsenic, chromium, and nickel in toenail samples from Appalachian Kentucky residents. *J Environ Pathol Toxicol Oncol, 30*(3), 213-223. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22126614

Joshi, P. M., Sutor, S. L., Huntoon, C. J., & Karnitz, L. M. (2014). Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly (ADP-ribose) polymerase inhibitors. *Journal of Biological Chemistry, 289*(13), 9247-9253.

Kamps, R., Brandao, R. D., Bosch, B. J., Paulussen, A. D., Xanthoulea, S., Blok, M. J., & Romano, A. (2017). Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *Int J Mol Sci, 18*(2). doi:10.3390/ijms18020308

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., . . . Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature, 502*(7471), 333-339. doi:10.1038/nature12634

Kerins, M. J., & Ooi, A. (2018). A catalogue of somatic NRF2 gain-of-function mutations in cancer. *Sci Rep, 8*(1), 12846.

Kim, Y.-A., Madan, S., & Przytycka, T. M. (2017). WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics, 33*(6), 814-821.

Kim, Y., Hammerman, P. S., Kim, J., Yoon, J.-a., Lee, Y., Sun, J.-M., . . . Park, K. (2014). Integrative and Comparative Genomic Analysis of Lung Squamous Cell Carcinomas in East Asian Patients. *Journal of Clinical Oncology, 32*(2), 121-128. doi:10.1200/JCO.2013.50.8556

Kim, Y. A., Madan, S., & Przytycka, T. M. (2017). WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics, 33*(6), 814-821. doi:10.1093/bioinformatics/btw242

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., . . . Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell, 161*(5), 1187-1201. doi:10.1016/j.cell.2015.04.044

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., . . . Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res, 22*(3), 568-576. doi:10.1101/gr.129684.111

Korthauer, K. D., & Kendziorski, C. (2015). MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics, 31*(10), 1526-1535. doi:10.1093/bioinformatics/btu858

Koster, J., & Rahmann, S. (2018). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics, 34*(20), 3600. doi:10.1093/bioinformatics/bty350

Labbe, R. M., Holowatyj, A., & Yang, Z. Q. (2013). Histone lysine demethylase (KDM) subfamily 4: structures, functions and therapeutic potential. *Am J Transl Res, 6*(1), 1-15. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24349617

Langdon, W. B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min, 8*(1), 1. doi:10.1186/s13040-014-0034-0

Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., . . . Seven Bridges, C. G. C. T. (2017). The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Res, 77*(21), e3-e6. doi:10.1158/0008-5472.CAN-17-0387

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., . . . Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature, 499*(7457), 214-218. doi:10.1038/nature12213

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., . . . others. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature, 499*(7457), 214.

Lee, J. C., Kang, S. U., Jeon, Y., Park, J. W., You, J. S., Ha, S. W., . . . Han, J. W. (2012). Protein L-isoaspartyl methyltransferase regulates p53 activity. *Nat Commun, 3*, 927. doi:10.1038/ncomms1933

Leiserson, M. D., Reyna, M. A., & Raphael, B. J. (2016). A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics, 32*(17), i736-i745. doi:10.1093/bioinformatics/btw462

Leiserson, M. D., Wu, H. T., Vandin, F., & Raphael, B. J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol, 16*, 160. doi:10.1186/s13059-015-0700-7

Leiserson, M. D. M., Blokh, D., Sharan, R., & Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol, 9*(5), e1003054.

Leiserson, M. D. M., Reyna, M. A., & Raphael, B. J. (2016). A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics, 32*(17), i736-i745.

Leiserson, M. D. M., Wu, H.-T., Vandin, F., & Raphael, B. J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology, 16*(1), 160.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., . . . Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biol, 5*(10), e254. doi:10.1371/journal.pbio.0050254

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Li, C., Gao, Z., Li, F., Li, X., Sun, Y., Wang, M., . . . Wei, Q. (2015). Whole Exome Sequencing Identifies Frequent Somatic Mutations in Cell-Cell Adhesion Genes in Chinese Patients with Lung Squamous Cell Carcinoma. *Sci Rep, 5*, 14237. doi:10.1038/srep14237

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi:10.1093/bioinformatics/btp698

Li, X., Wu, C., Chen, N., Gu, H., Yen, A., Cao, L., . . . Wang, L. (2016). PI3K/Akt/mTOR signaling pathway and targeted therapy for glioblastoma. *Oncotarget, 7*(22), 33440.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923-930. doi:10.1093/bioinformatics/btt656

Little, R. J. A. (1991). Inference with survey weights. *Journal of Official Statistics, 7*(4), 405.

Liu, X., Yu, Y., Liu, J., Elliott, C. F., Qian, C., & Liu, J. (2018). A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with k-mer signatures. *Bioinformatics, 34*(1), 171-178. doi:10.1093/bioinformatics/btx432

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Lu, C., Ward, P. S., Kapoor, G. S., Rohle, D., Turcan, S., Abdel-Wahab, O., . . . Thompson, C. B. (2012). IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature, 483*(7390), 474-478. doi:10.1038/nature10860

Ma, F., Fuqua, B. K., Hasin, Y., Yukhtman, C., Vulpe, C. D., Lusis, A. J., & Pellegrini, M. (2019). A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC Genomics, 20*(1), 9. doi:10.1186/s12864-018-5393-3

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., . . . McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell, 161*(5), 1202-1214. doi:10.1016/j.cell.2015.05.002

Martin, M. (2011a). Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011, 17*(1). doi:10.14806/ej.17.1.200

pp. 10-12

Martin, M. (2011b). Cutadapt removes adapter sequences from high-throughput sequencing reads. *17*(1). doi:http://dx.doi.org/10.14806/ej.17.1.200

McFadden, P. N., & Clarke, S. (1982). Methylation at D-aspartyl residues in erythrocytes: possible step in the repair of aged membrane proteins. *Proc Natl Acad Sci U S A, 79*(8), 2460-2464. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6123997

McGranahan, N., & Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell, 168*(4), 613-628.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol, 12*(4), R41. doi:10.1186/gb-2011-12-4-r41

Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet, 11*(10), 685-696. doi:10.1038/nrg2841

Mina, M., Raynaud, F., Tavernari, D., Battistello, E., Sungalee, S., Saghafinia, S., . . . others. (2017). Conditional selection of genomic alterations dictates cancer evolution and oncogenic dependencies. *Cancer cell, 32*(2), 155-168.

Modugno, F., Laskey, R., Smith, A. L., Andersen, C. L., Haluska, P., & Oesterreich, S. (2012). Hormone response in ovarian cancer: time to reconsider as a clinical target? *Endocrine-related cancer, 19*(6), R255-R279.

Molenaar, R. J., Thota, S., Nagata, Y., Patel, B., Clemente, M., Przychodzen, B., . . . Maciejewski, J. P. (2015). Clinical and biological implications of ancestral and non-ancestral IDH1 and IDH2 mutations in myeloid neoplasms. *Leukemia, 29*(11), 2134-2142. doi:10.1038/leu.2015.91

Mondesir, J., Willekens, C., Touat, M., & de Botton, S. (2016). IDH1 and IDH2 mutations as novel therapeutic targets: current perspectives. *J Blood Med, 7*, 171-180. doi:10.2147/JBM.S70716

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., . . . others. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature, 534*(7605), 47-54.

Okumura, F., Joo-Okumura, A., Nakatsukasa, K., & Kamura, T. (2016). The role of cullin 5-containing ubiquitin ligases. *Cell Div, 11*, 1. doi:10.1186/s13008-016-0016-3

Paculov\'a, Hana, & Kohoutek, J. v., ,\'\i,. (2017). The emerging roles of CDK12 in tumorigenesis. *Cell Div, 12*(1), 7.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2018). zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience, 7*(6). doi:10.1093/gigascience/giy059

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods, 14*(4), 417-419. doi:10.1038/nmeth.4197

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., . . . others. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun, 7*(1), 1-16.

Petitjean, A., Achatz, M. I., Borresen-Dale, A. L., Hainaut, P., & Olivier, M. (2007). TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene, 26*(15), 2157-2165. doi:10.1038/sj.onc.1210302

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical methods in medical research, 5*(3), 239-261.

Philip, B., Diana, X. Y., Silvis, M. R., Shin, C. H., Robinson, J. P., Robinson, G. L., . . . others. (2018). Mutant IDH1 promotes glioma formation in vivo. *Cell reports, 23*(5), 1553-1564.

Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., . . . Getz, G. (2015). Oncotator: cancer variant annotation tool. *Hum Mutat, 36*(4), E2423-2429. doi:10.1002/humu.22771

Rhodes, J. M., Bentley, F. K., Print, C. G., Dorsett, D., Misulovin, Z., Dickinson, E. J., . . . Horsfield, J. A. (2010). Positive regulation of c-Myc by cohesin is direct, and evolutionarily conserved. *Developmental biology, 344*(2), 637-649.

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. doi:10.1093/bioinformatics/btp616

Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., . . . Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics, 27*(19), 2648-2654. doi:10.1093/bioinformatics/btr462

Schoenberg, N. E., Huang, B., Seshadri, S., & Tucker, T. C. (2015). Trends in cigarette smoking and obesity in Appalachian Kentucky. *South Med J, 108*(3), 170-177. doi:10.14423/SMJ.0000000000000245

Schwartz, R., & Sch\"a, f., Alejandro A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics, 18*(4), 213.

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform, 16*(1), 59-70. doi:10.1093/bib/bbt086

Sherr, C. J. (2006). Autophagy by ARF: a short story. *Molecular cell, 22*(4), 436-437.

Simonsen, A. T., Hansen, M. C., Kjeldsen, E., Moller, P. L., Hindkjaer, J. J., Hokland, P., & Aggerholm, A. (2018). Systematic evaluation of signal-to-noise ratio in variant detection from single cell genome multiple displacement amplification and exome sequencing. *BMC Genomics, 19*(1), 681. doi:10.1186/s12864-018-5063-5

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res, 27*(3), 491-499. doi:10.1101/gr.209601.116

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res, 4*, 1521. doi:10.12688/f1000research.7563.2

Srivastava, A., Malik, L., Smith, T., Sudbery, I., & Patro, R. (2019). Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome biology, 20*(1), 65. doi:10.1186/s13059-019-1670-y

Stallman RM, M. R. (1991). GNU Make—A Program for Directing Recompilation. Retrieved from http://www.gnu.org/software/make/

Suva, M. L., & Tirosh, I. (2019). Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Mol Cell, 75*(1), 7-12. doi:10.1016/j.molcel.2019.05.003

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc, 13*(4), 599-604. doi:10.1038/nprot.2017.149

Szczurek, E., & Beerenwinkel, N. (2014). Modeling mutual exclusivity of cancer mutations. *PLoS Comput Biol, 10*(3), e1003503. doi:10.1371/journal.pcbi.1003503

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., . . . Lopez-Bigas, N. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep, 3*, 2650. doi:10.1038/srep02650

Taub, M. A., Corrada Bravo, H., & Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine, 2*(12), 87. Retrieved from https://doi.org/10.1186/gm208

Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., . . . Irizarry, R. A. (2016). A benchmark for RNA-seq quantification pipelines. *Genome Biol, 17*, 74. doi:10.1186/s13059-016-0940-1

The Cancer Genome Atlas Research Network. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature, 489*(7417), 519-525. doi:10.1038/nature11404

Thomas, R. K., Baker, A. C., DeBiasi, R. M., Winckler, W., LaFramboise, T., Lin, W. M., . . . others. (2007). High-throughput oncogene mutation profiling in human cancer. *Nature genetics, 39*(3), 347.

Tian, L., Dong, X., Freytag, S., Le Cao, K. A., Su, S., JalalAbadi, A., . . . Ritchie, M. E. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using

mixture control experiments. *Nat Methods, 16*(6), 479-487. doi:10.1038/s41592-019-0425-8

Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., . . . Ritchie, M. E. (2018). scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput Biol, 14*(8), e1006361. doi:10.1371/journal.pcbi.1006361

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A, 98*(9), 5116-5121. doi:10.1073/pnas.091062498

Vandin, F., Upfal, E., & Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Research, 22*(2), 375-385.

Venkatraman E. Seshan, A. O. DNAcopy: DNA copy number data analysis. R package version 1.44.0.

Walport, L. J., Hopkinson, R. J., Chowdhury, R., Schiller, R., Ge, W., Kawamura, A., & Schofield, C. J. (2016). Arginine demethylation is catalysed by a subset of JmjC histone lysine demethylases. *Nat Commun, 7*, 11974. doi:10.1038/ncomms11974

Wang, Y., & Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Mol Cell, 58*(4), 598-609. doi:10.1016/j.molcel.2015.05.005

Ward, P. S., Cross, J. R., Lu, C., Weigert, O., Abel-Wahab, O., Levine, R. L., . . . Thompson, C. B. (2012). Identification of additional IDH mutations associated with oncometabolite R(-)-2-hydroxyglutarate production. *Oncogene, 31*(19), 2491-2498. doi:10.1038/onc.2011.416

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., . . . Goble, C. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res, 41*(Web Server issue), W557-561. doi:10.1093/nar/gkt328

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J, 16*, 15-24. doi:10.1016/j.csbj.2018.01.003

Xu, H., Zong, H., Ma, C., Ming, X., Shang, M., Li, K., . . . Cao, L. (2017). Epidermal growth factor receptor in glioblastoma. *Oncology letters, 14*(1), 512-516.

Xu, K., Wu, Z. J., Groner, A. C., He, H. H., Cai, C., Lis, R. T., . . . Brown, M. (2012). EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science, 338*(6113), 1465-1469. doi:10.1126/science.1227604

Yan, H., Parsons, D. W., Jin, G., McLendon, R., Rasheed, B. A., Yuan, W., . . . Bigner, D. D. (2009). IDH1 and IDH2 mutations in gliomas. *N Engl J Med, 360*(8), 765-773. doi:10.1056/NEJMoa0808710

Youn, A., & Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics, 27*(2), 175-181. doi:10.1093/bioinformatics/btq630

Yu, Y., Belazzougui, D., Qian, C., & Zhang, Q. (2018). Memory-Efficient and Ultra-Fast Network Lookup and Forwarding Using Othello Hashing. *IEEE/ACM Transactions on Networking, 26*(3), 1151-1164.

Yu, Y., Liu, J., Liu, X., Zhang, Y., Magner, E., Lehnert, E., . . . Liu, J. (2018). SeqOthello: querying RNA-seq experiments at scale. *Genome Biol, 19*(1), 167. doi:10.1186/s13059-018-1535-9

Zhang, L., Dong, X., Lee, M., Maslov, A. Y., Wang, T., & Vijg, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proceedings of the National Academy of Sciences, 116*(18), 9014-9019. Retrieved from https://www.pnas.org/content/116/18/9014

Zhao, S., Liu, J., Nanga, P., Liu, Y., Cicek, A. E., Knoblauch, N., . . . He, X. (2019). Detailed modeling of positive selection improves detection of cancer driver genes. *Nat Commun, 10*(1), 1-13.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., . . . Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat Commun, 8*, 14049. doi:10.1038/ncomms14049

Zhu, J., Sammons, M. A., Donahue, G., Dou, Z., Vedadi, M., Getlik, M., . . . Berger, S. L. (2015). Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature, 525*(7568), 206-211. doi:10.1038/nature15251

VITA

Jinpeng Liu

- Education
  - M.S. in Computer Science Jan. 2013 - June. 2014

    University of Kentucky, Lexington, KY, USA
  - M.S. in Pharmaceutical Science Sep. 2006 - July. 2009

    Shanxi Medical University, Taiyuan, Shanxi, China
  - B.S. in Biological Engineering Sep. 2002 - July. 2006

    Beijing Institute of Technology, Beijing, China
- Employment History
  - Bioinformatics Analyst Senior, June 2014 - Present

    University of Kentucky, Lexington, KY, USA
  - Research Assistant, Jan. 2013 - June. 2014

    University of Kentucky, Lexington, KY, USA
  - Shanxi Institute for Food and Drug Control, Sep. 2006 - Jan. 2011

    Taiyuan, Shanxi, China
- Publications

1. Liu J, Liu X, Yu Y, Wang C, Liu JZ: FastCount: A Fast Gene Count Software for Single-cell RNA-seq Data. Manuscript in preparation 2021.
2. Chen F, Byrd A, Liu J, Robert M. Flight, Tanner J. DuCote, Mojtaba Bakhtiari, Xiulong Song, Abigail R. Edgin,Alexsandr Lukyanchuk, Danielle T. Dixon, Stuart H. Orkin5, Hunter N.B. Mosely, Chi Wang, Christine Fillmore Brainson: Polycomb deficiency drives a FOXP2-high aggressive state targetable by epigenetic inhibitors. Nature Communications. Manuscript in revision 2021.
3. Liu S*, Liu J*, Xie Y, Zhai T, Hinderer EW, Stromberg AJ, Vanderford NL,

Kolesar JM, Moseley HNB, Chen L, Liu C, Wang C: MEScan: a powerful statistical framework for Genome-Scale mutual exclusivity analysis of cancer mutations. Bioinformatics (Oxford, England) 2020.

4. Lin N*, Liu J*, Castle J, Wan J, Shendre A, Liu Y, Wang C, He C: Genome-wide DNA methylation profiling in human breast tissue by illumina TruSeq methyl capture EPIC sequencing and infinium methylationEPIC beadchip microarray. Epigenetics 2020:1-16.
5. Conroy L, Youn L, Stanback A, Austin L. G, Liu J, Liu JZ, Allison D, Sun C. R: Mass spectrometry imaging of N-glycans reveals racial discrepancies in low grade prostate tumors. The Journal of Clinical Investigation. Manuscript in review 2020.
6. Liu J*, Murali T*, Yu T*, Liu C, Sivakumaran TA, Moseley HNB, Zhulin IB, Weiss HL, Durbin EB, Ellingson SR, Liu J, Huang B, Hallahan BJ, Horbinski CM, Hodges K, Napier DL, Bocklage T, Mueller J, Vanderford NL, Fardo DW, Wang C, Arnold SM: Characterization of Squamous Cell Lung Cancers from Appalachian Kentucky. Cancer Epidemiol Biomarkers Prev 2019, 28(2):348-356.
7. Yu Y, Liu J, Liu X, Zhang Y, Magner E, Lehnert E, Qian C, Liu J: SeqOthello: querying RNA-seq experiments at scale. Genome biology 2018, 19(1):167.
8. Liu X, Yu Y, Liu J, Elliott CF, Qian C, Liu J: A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with k-mer signatures. Bioinformatics (Oxford, England) 2018, 34(1):171-178.
9. Wang M, Yu T, Liu J, Chen L, Stromberg AJ, Villano JL, Arnold SM, Liu C, Wang C: A probabilistic method for leveraging functional annotations to enhance estimation of the temporal order of pathway mutations during carcinogenesis. BMC bioinformatics 2019, 20(1):620
10. Zhong Y, Mohan K, Liu J, Al-Attar A, Lin P, Flight RM, Sun Q, Warmoes MO, Deshpande RR, Liu H: Loss of CLN3, the gene mutated in juvenile neuronal ceroid lipofuscinosis, leads to metabolic impairment and autophagy induction in retina pigment epithelium. BBA - Molecular Basis of Disease 2020.
11. Tripathi R, Liu Z, Jain A, Lyon A, Meeks C, Richards D, Liu J, He D, Wang C, Nespi M, Rymar A, Wang P, Wilson M, Plattner R: Combating acquired resistance to MAPK inhibitors in melanoma by targeting Abl1/2-mediated reactivation of

MEK/ERK/MYC signaling. Nature communications 2020, 11(1):5463.

12. Liu J, He D, Cheng L, Huang C, Zhang Y, Rao X, Kong Y, Li C, Zhang Z, Liu J, Jones K, Napier D, Lee EY, Wang C, Liu X: p300/CBP inhibition enhances the efficacy of programmed death-ligand 1 blockade treatment in prostate cancer. Oncogene 2020, 39(19):3939-3951.
13. Li J, Li X, Song J, Yan B, Rock SA, Jia J, Liu J, Wang C, Weiss T, Weiss HL, Gao T, Alam A, Evers BM: Absence of neurotensin attenuates intestinal dysbiosis and inflammation by maintaining Mmp7/alpha-defensin axis in diet-induced obese mice. FASEB J 2020, 34(6):8596-8610.
14. Liu L, Qi L, Knifley T, Piecoro DW, Rychahou P, Liu J, Mitov MI, Martin J, Wang C, Wu J, Weiss HL, Butterfield DA, Evers BM, O'Connor KL, Chen M: S100A4 alters metabolism and promotes invasion of lung cancer cells by up-regulating mitochondrial complex I protein NDUFS2. J Biol Chem 2019, 294(18):7516-7527.
15. Liu L, Qi L, Knifley T, Piecoro DW, Rychahou P, Liu J, Mitov MI, Martin J, Wang C, Wu J: S100A4 alters mitochondrial metabolism to promote invasion and metastasis of non-small cell lung cancer cells through upregulation of NDUFS2. In.: AACR; 2019.
16. Li H, Li J, Han R, Deng X, Shi J, Huang H, Hamad N, McCaughley A, Liu J, Wang C, Chen K, Wei D, Qiang J, Thatcher S, Wu Y, Liu C, Thibault O, Wei X, Chen S, Qian H, Zhou BP, Xu P, Yang XH: Deletion of tetraspanin CD151 alters the Wnt oncogene-induced mammary tumorigenesis: A cell type-linked function and signaling. Neoplasia 2019, 21(12):1151-1163.
17. Jafari N, Drury J, Morris AJ, Onono FO, Stevens PD, Gao T, Liu J, Wang C, Lee EY, Weiss HL, Evers BM, Zaytseva YY: De Novo Fatty Acid Synthesis-Driven Sphingolipid Metabolism Promotes Metastatic Potential of Colorectal Cancer. Mol Cancer Res 2019, 17(1):140-152.
18. Zaytseva YY, Rychahou PG, Le AT, Scott TL, Flight RM, Kim JT, Harris J, Liu J, Wang C, Morris AJ, Sivakumaran TA, Fan T, Moseley H, Gao T, Lee EY, Weiss HL, Heuer TS, Kemble G, Evers M: Preclinical evaluation of novel fatty acid synthase inhibitors in primary colorectal cancer cells and a patient-derived xenograft model of colorectal cancer. Oncotarget 2018, 9(37):24787-24800.

19. Wang Q, Kim JT, Zhou Y, Li C, Liu J, Wang C, Evers BM: 87–Mir-181A-5P Contributes to Intestinal Cell Differentiation. Gastroenterology 2019, 156(6):S-19-S-20.
20. Thampi P, Liu J, Zeng Z, MacLeod JN: Changes in the appendicular skeleton during metamorphosis in the axolotl salamander (Ambystoma mexicanum). Journal of anatomy 2018, 233(4):468-477.
21. Tripathi R, Fiore LS, Richards DL, Yang Y, Liu J, Wang C, Plattner R: Abl and Arg mediate cysteine cathepsin secretion to facilitate melanoma invasion and metastasis. Sci Signal 2018, 11(518).
22. Chaiswing L, Thorson J, Xu FF, Wang C, Liu J, Clair DS, Clair WS: RelB-BLNK axis: a novel determinant of cell fate. Free Radical Biology and Medicine 2018, 128:S64-S65.
23. Wang Q, Zhu J, Wang YW, Dai Y, Wang YL, Wang C, Liu J, Baker A, Colburn NH, Yang HS: Tumor suppressor Pdcd4 attenuates Sin1 translation to inhibit invasion in colon carcinoma. Oncogene 2017, 36(45):6225-6234
24. Oben KZ, Alhakeem SS, McKenna MK, Brandon JA, Mani R, Noothi SK, Liu J, Akunuru S, Dhar SK, Singh IP: Oxidative stress-induced JNK/AP-1 signaling is a major pathway involved in selective apoptosis of myelodysplastic syndrome cells by Withaferin-A. Oncotarget 2017, 8(44):77436.
25. Carpenter BL, Liu J, Qi L, Wang C, O'Connor KL: Integrin α6β4 Upregulates Amphiregulin and Epiregulin through Base Excision Repair-Mediated DNA Demethylation and Promotes Genome-wide DNA Hypomethylation. Scientific reports 2017, 7(1):1-14.
26. Yu T, Chen X, Zhang W, Liu J, Avdiushko R, Napier DL, Liu AX, Neltner JM, Wang C, Cohen D, Liu C: KLF4 regulates adult lung tumor-initiating cells and represses K-Ras-mediated lung cancer. Cell Death Differ 2016, 23(2):207-215.
27. Wang H, Horbinski C, Wu H, Liu Y, Sheng S, Liu J, Weiss H, Stromberg AJ, Wang C: NanoStringDiff: a novel statistical method for differential expression analysis based on NanoString nCounter data. Nucleic acids research 2016, 44(20):e151.
28. Wang C, Liu J, Fardo DW: Causal effect estimation in sequencing studies: a

Bayesian method to account for confounder adjustment uncertainty. In: BMC proceedings: 2016: BioMed Central; 2016: 411-415.

29. Rouchka EC, Jeoung M, Jang ER, Liu J, Wang C, Li X, Galperin E: Data set for transcriptional response to depletion of the Shoc2 scaffolding protein. Data Brief 2016, 7:770-778.
30. Oben KZ, Gachuki B, Akunuru S, Liu J, Brandon J, St Clair D, Wang C, Geiger H, Gupta R, Bondada S: Growth inhibitory effects of Withaferin A on MDS-L cells–A human 5q Myelodysplastic Syndrome (MDS) cell line. In.: Am Assoc Immnol; 2016.
31. Jeoung M, Jang ER, Liu J, Wang C, Rouchka EC, Li X, Galperin E: Shoc2-tranduced ERK1/2 motility signals—novel insights from functional genomics. Cellular signalling 2016, 28(5):448-459.
32. Gal TS, Ellingson SR, Wang C, Liu J, Jarrett SG, D'Orazio JA: Using large public data repositories to discover novel genetic mutations with prospective links to melanoma. BMC bioinformatics 2015, 16(15):1-3.