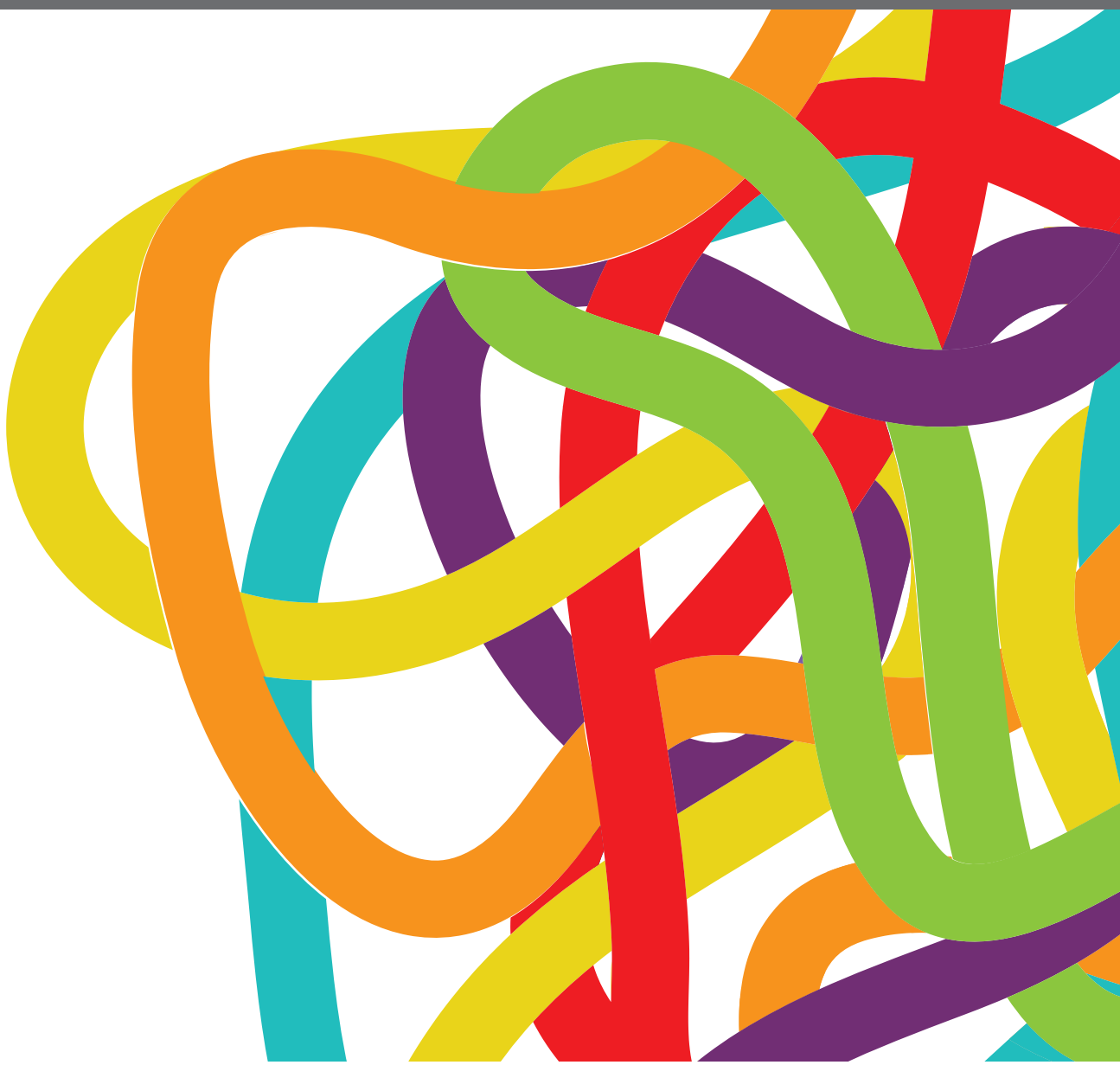
USING CANCER 'OMICS' TO UNDERSTAND CANCER

EDITED BY: Daoud Meerzaman and Barbara Karen Dunn
PUBLISHED IN: Frontiers in Oncology and Frontiers in Genetics



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-005-6
DOI 10.3389/978-2-88966-005-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

USING CANCER 'OMICS' TO UNDERSTAND CANCER

Topic Editors:
Daoud Meerzaman, George Washington University, United States
Barbara Karen Dunn, National Institutes of Health (NIH), United States

Citation: Meerzaman, D., Dunn, B. K., eds. (2020). Using Cancer 'Omics' to Understand Cancer. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-005-6

Table of Contents

frontiers in Oncology

EDITORIAL
published: 24 July 2020
doi: 10.3389/fonc.2020.01201

Editorial: Using Cancer 'Omics' to Understand Cancer

Barbara K. Dunn[1] *and Daoud Meerzaman*[2]*

[1] Division of Cancer Prevention (NCI), Bethesda, MD, United States, [2] Center for Biomedical Informatics and Information, Technology (NCI), Rockville, MD, United States

Keywords: big data, cancer genomics, cancer genomics data analysis, HPC (high performance computing), humanizing big data

Editorial on the Research Topic

Using Cancer 'Omics' to Understand Cancer

The notion of using the "big data" approach to study human disease is not new. Scientists have been tapping data from studies of genomics, proteomics, transcriptomics, metabolomics, and microbiomics since the initial mapping of the human genome (1). What has changed, however, is a fundamental shift in how we think about these technologies. The "omics" field is expanding in scope, blending biology, technology (radiomics), and clinical observations (electronic health records), as well as size. This amplification of content and quantity has required parallel development and application of novel informatic tools. The need to accommodate the ever-larger datasets critical to our understanding of cancer omics has instigated a movement toward development of high-performance computing, including both hardware and software to analyze the massive, generated big data. The manuscripts contained in this volume reflect this constantly evolving panel of bioinformatic programs and resources with capacity to carry out large-scale data analysis.

Most of the papers in this issue report findings that share the common feature that all distill a select number of biomarkers from a large spectrum of potential markers from an analysis of large datasets. This volume of Frontiers broadens its approach to include papers dealing directly with the attributes, management, and clinical application of big data. Focusing on some of the key databases, projects and methodologies developed to implement such analyses, emphasizing the ever-expanding scale of big data, exascale computing is discussed. At the initiation of marker discovery, the patients and other individuals who serve as the source of big data are highlighted, while encouraging big data researchers to keep in mind the humanity inherent in these data (Helzlsouer et al.).

To date much of our focus has been on comparing the omics information of cancer patients with that of "normal" controls, i.e., healthy individuals, and looking for genotypic or phenotypic differences that set the patients apart. This rudimentary approach has led to practical applications, including offering targets for early detection, prognosis, and treatment. Along these lines, in this special issue of Frontiers, several authors address genomic (and epigenomic) abnormalities that characterize specific cancers and may thus have practical applications at the clinical level.

The manuscript by Yang et al. offers an example of the application of omics research to biomarker discovery. This paper describes potential diagnostic markers and therapeutic targets for leiomyosarcoma (LMS). This cancer is particularly aggressive, with invasive clinical characteristics and often a poor prognosis. Finding new biomarkers to assess malignancy and prognosis of LMS is critical. Yang et al. used Weighted Gene Co-expression Network Analysis (WGCNA), a systematic molecular clustering approach, to look for gene expression patterns that are associated with LMS and thereby should help to improve our understanding of the molecular mechanisms of this

OPEN ACCESS

Edited and reviewed by:
Heather Cunliffe,
University of Otago, New Zealand

***Correspondence:**
Daoud Meerzaman
meerzamd@mail.nih.gov

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *13 May 2020*
Accepted: *12 June 2020*
Published: *24 July 2020*

Citation:
Dunn BK and Meerzaman D (2020) Editorial: Using Cancer 'Omics' to Understand Cancer. Front. Oncol. 10:1201. doi: 10.3389/fonc.2020.01201

cancer. Their results showed that the expression of CDK4, CCT2, and MGAT1 in LMS tissues was significantly higher than that in adjacent tissues, suggesting that these genes may be part of the cancer signaling pathway. Such findings could pave the way for new strategies for diagnosing and treating LMS.

Another cancer, glioblastoma multiforme (GBM), is the focus of two articles in this special volume, by Cheng et al. and by Stajkovska et al. Cheng et al. employed a data mining approach by tapping into The Cancer Genome Atlas (TCGA). TCGA is managed by the Genomic Data Commons (GDC) (2) funded by the National Cancer Institute (NCI) which provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine. They then applied various bioinformatic tools aimed at discovery of relevant genes and pathways. They examined the gene expression patterns of transcription factors associated with GBM and identified four potential candidates based on their differential expression between tumor and adjacent tissue: *LHX2*, *MEOX2*, *SNAI2*, and *ZNF22*. By clustering transcription factors that are differentially expressed in GBM and screening these clusters using appropriate bioinformatic programs, they identified cancer pathways primarily associated with cell migration, cell adhesion, epithelial-mesenchymal transition (EMT), cell cycle, as well as other signaling pathways. Combining these results with patient characteristics, such as risk score, age, gender, type of treatment, and treatment response, these authors showed that their model was able to precisely predict the outcome of patients with GBM. GBM was further explored in the study by Stajkovska et al., in their description of a case report of a pediatric patient. Using targeted gene panel testing in blood and tumor tissue, these researchers identified a heterozygous frameshift mutation (c.333_334delTC; p.His112CysfsTer9) in the *MLH1* gene in addition to a known heterozygous missense variant of unknown significance/VUS (c.847C > T; p.Arg283Cys) in the *TP53* gene. Screening of the patient's parents revealed the presence of the *MLH1* abnormality in the father and the *TP53* variant in the mother. They report for the first time the co-occurrence of a genetic mutation in the *MLH1* gene of the mismatch repair pathway, often associated with Lynch syndrome, accompanied by a rare variant in the *TP53* gene. The authors stress that co-occurrence of multiple gene abnormalities should be considered as a possible contributory cause of a cancer. However, caution must be exercised in interpreting a VUS as contributing to the cancer phenotype, as these variants are of unproven pathogenicity, a subject addressed in Helzlsouer et al. in this volume.

Biomarkers also are the focus of the study by Wang Y. et al., who looked at new ways of predicting the progression and prognosis of bladder cancer (BC) using a big data approach. Through a series of screenings and WGCNA they identified "hub" genes (i.e., a hub gene serves as the focal point of interaction with other genes; in general, the genes connected to the hub are critical to gene regulation and other biological processes). Gene-set enrichment analysis (GSEA) revealed that the sets of highly expressed hub genes were mainly enriched in "bladder cancer," "cell cycle," and "ubiquitin-mediated proteolysis" related pathways. They further honed their results to two genes (*ANLN*, *HMMR*), which had prognostic value for different stages and grades of BC. These genes not only could accurately predict the overall survival of patients with BC, but also the progression-free survival, a common outcome measure in clinical trials.

In another biomarker study included in this volume, Wang X. et al. showed how a set of small nucleolar RNAs (snoRNAs), which guide the modification of other RNAs and which have been implicated in alternative splicing, can predict overall survival of gastric cancer patients. An eight-snoRNA risk signature serves as a prognostic factor in gastric cancer. The authors validated the expression patterns of these eight snoRNAs, both in cell lines and patients' tissues. The authors point out that seven of these snoRNAs correlate with survival, suggesting relevance of these markers to the clinical behavior of the bladder cancer. One snoRNA, U66, was linked to cell proliferation. These findings provide potential prognostic and therapeutic clues into gastric cancer.

Nersisyan et al. addressed the mechanistic basis of tumorigenesis by examining the component that involves telomere status. Unlike normal cells where telomeres are shortened with each cell division, telomere maintenance mechanisms (TMMs) are found in most cancers. Of the two types of TMMs found in cancer, most cancers exhibit a TMM that is activated via the classical "telomerase" pathway (TEL), using the telomerase ribonucleoprotein, which contains an RNA template that guides the synthesis of the telomere DNA. In contrast, the alternative TMM, which operates in a smaller proportion of tumors, is the "alternative lengthening of telomeres" (ALT) pathway. The ALT pathway, which relies on complex molecular mechanisms including homologous recombination events between telomeric sister chromatid strands, occurs in the context of an altered chromatin environment at the telomere region. Nersisyan and colleagues compared the TMM pathways in colorectal cancers (CRC) with microsatellite instability/MSI (both CRCs in Lynch syndrome/LS-CRC and sporadic MSI CRCs/MSI s-CRC) to a subset of sporadic microsatellite stable (MSS) CRCs as well as benign mucosa. In their study of alterations of telomere length, sequence composition, and transcriptional regulation in relation to the two types of TMMs (TEL, ALT) in CRCs, they applied bioinformatic analysis to big data from whole genome DNA and RNA sequencing together with a pathway model. They observed transcriptomic signatures that distinguish the two TMM subtypes in CRC, with ALT-TMM being slightly more prominent in hypermutated MSI s-CRC and LS-CRC.

Chen et al. show how DNA methylation, an important regulator of gene expression, can be used, along with other tumor and patient characteristics, to identify glioma subgroups that exhibit specific prognostic features. DNA methylation patterns were examined in 653 gliomas from the TCGA database of NCI. The authors used consensus clustering to narrow their findings of methylation levels at each CpG site known to influence survival into five subgroups. DNA methylation patterns were then correlated with age, tumor stage, and prognosis. WGCNA of the CpG sites identified 11 clusters that could be used to differentiate

between high- and low-methylation groups and which could be further used to determine prognostic information about the glioma patients. When applied to *in vitro* experiments, an inverse relationship was shown between methylation level of glioma cells and their ability to migrate or their inability to respond to standard glioma therapies, temozolomide or radiotherapy. Thus, epigenetic (methylation) subtypes could potentially serve as markers for prognosis as well as guides to glioma therapies.

Lee's article offers context to this study of methylation in carcinogenesis by providing an overview of epigenetics, highlighting how abnormal epigenetic modifications contribute to the development of cancer. Beyond reviewing the basic molecular mechanisms of epigenetic regulation of gene expression (methylation, histone modification, and non-coding RNAs), Lee discusses the role of epigenetics in regulating differentiation during development while simultaneously maintaining epigenetic memory during mitotic cell division. As an example, abnormal methylation of tumor suppressor genes downregulates expression, which when coupled with a mutation in the other allele contributes to carcinogenesis, according to the two-hit theory of Knudson. This concept is broadened to allele-specific gene expression (ASE) in general and its epigenetic regulation by allele-specific methylation (ASM). Starting from these descriptions of individual epigenetic abnormalities leading to cancer, the article extends into the epigenomic realm. Lee points out how the use of big datasets such as TCGA serve as a source not only of genomic information for analytic exploration but also for comparable investigations into large-scale epigenomic data. A prototypic example is the investigation of the TCGA dataset that identified a subset of GBMs with high CpG island methylation, subsequently labeled as a "glioma CpG island methylator phenotype" (G-CIMP). Clinical correlation of G-CIMP-positive tumors included higher prevalence among lower-grade gliomas and increased association with isocitrate dehydrogenase 1 (IDH1) somatic mutations. G-CIMP serves as merely one illustration of the extension of big data applications into the epigenetic, now the epigenomic, domain. This paper concludes by bringing the fruits of epigenetic/epigenomic research into the clinical realm, enumerating examples of approved cancer therapies that target cancer-inducing epigenetic abnormalities.

High throughput studies addressing big data also are helping us to identify subgroups of patients to better understand how disease affects certain populations. This approach has potential to predict which populations have patients who are more likely to respond to certain medications. Using lower throughput platforms, Kénémé and Sémbène studied genetic determinants of uterine fibroids (UF), benign tumors that are more frequent and are associated with more severe symptoms in African-American women. Focusing on 55 Senegalese women, their examination of genetic abnormalities in UFs in this population disclosed high genetic variability in repetition number of a GT dinucleotide microsatellite in the first intron of the *COL1A2* gene. In addition to microsatellite instability, two GT sites had distinct mutations in the UFs in subsets of women. Furthermore, beyond confirming the involvement of the *COL1A2* dinucleotide length polymorphism, $GT_{n,}$ in the occurrence of uterine fibroids in Senegalese women, these UF-associated genetic variants were additionally analyzed in relation to ethnicity, marital status, contraception use, diet, and physical activity. For the first time, these epidemiologic factors were shown to exhibit associations with the genetic underpinnings of UFs in this population. The authors consider that these results may create avenues for understanding the mechanisms involved in the racial variation in the prevalence and symptomatic severity of UFs as well as the predisposing factors.

The contents of this volume to this point have addressed the use of big data in investigations of various types of molecular mechanisms that underlie carcinogenesis in general and in specific cancers (and benign tumors). In contrast, Bhattacharya et al. delve directly into the nature and operation of data science, enumerating those attributes that enable its application to the discovery of carcinogenic mechanisms that are potentially targetable for prevention and treatment. The authors demonstrate how progress in the quantity and diversity of biomedical data, together with advances in artificial intelligence (AI) and machine learning (ML) algorithms, as well as computer architectures, enable advances in big data with a goal of accelerating cancer research. The authors take AI and ML to exascale levels, which are orders of magnitude higher than those of current high-end machines, in order to gain a deeper understanding of cancer. They describe a collaboration between the Department of Energy (DOE) and the NCI, the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C), which has three pilot projects intended to push the frontiers of computing technologies in cancer research at the cellular, molecular and population levels. An example of the first pilot involves the application of exascale computing technology to a precision medicine initiative to develop predictive capabilities of drug response in pre-clinical models, ultimately leading to targeted cancer therapies in the clinic. The evolving needs of population databases, such as the Surveillance, Epidemiology, and End Results (SEER) registry of U.S. cancer incidence, as they increase the breadth of information collected, are being addressed by the high-performance computing and AI, as seen in the third pilot. The potential scope of applications of exascale computing is vast and multimodal, with potential for improving our understanding and management of cancer.

As evidenced by this special issue of *Frontiers in Oncology*, the omics field and the big data tools designed to support cancer research already are yielding results that are being translated into clinical practice. Helzlsouer et al. remind us, however, that the source of every piece of data is a human being. This connection must not get lost as we delve into the technical processes of sample collection, preparation, and analysis, both in the laboratory and at the informatic levels. In essence, we must take special care to "humanize" these big data. Helzlsouer et al. show that it is also critical to examine the challenges of genetic/genomic testing at the individual level, i.e., the human level. The limitations to clinical implications derived from analyses of big data, including the probabilistic nature inherent in genetic findings, need to be made clear to patients, but also to all health care providers. Maintaining the human aspect of these

data sources is vital as we look to translate and apply findings to the cancer research field.

Today's big data require centralized, well-curated, and readily accessible databases that accommodate large-scale datasets. To this end, the National Institutes of Health and the NCI are actively contributing by establishing a number of data repositories within a larger Cancer Research Data Commons (CRDC) (3). These storehouses of data, coupled with large-scale, high-throughput sequencing technologies (genome, transcriptome, proteome), and deep machine learning, are resulting in exponential growth in data-driven solutions.

This special volume provides only a snapshot of articles featuring applications and approaches to omics data. Yet, this is an area that is just beginning to see its full potential. Big data are expanding our understanding of disease at its most fundamental level. The manuscripts in this special issue, given their diversity, reflect the multidisciplinary nature of the field. They further underscore the importance of collaboration using a fully integrated approach, from basic scientists to data/computational/modeling analysts (4).

We've come a long way since first mapping the genome. As we further unlock individual genomes we need to take care that we can protect personal information and avoid the potential for bias, highlighting the ethical aspect of data derived from humans Helzlsouer et al.. The use and reuse of data need to be carefully managed so that the interest and welfare of patients and others who share their data are maintained. In another decade we are sure to realize even greater advances in how we prevent, diagnose, and treat not only cancer, but a broad range of diseases, relying on the availability of robust big data.

AUTHOR CONTRIBUTIONS

BD and DM contributed to the conceptualization and writing the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank Barbara Vann with assistance in preparation of this manuscript.

REFERENCES

1. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol.* (2017)18:83. doi: 10.1186/s13059-017-1215-1
2. Genomic Data Commons (GDC). Available online at: https://gdc.cancer.gov.
3. Cancer Research Data Commons/CRDC. Available online at: https://datascience.cancer.gov/data-commons.
4. Meerzaman D, Dunn BK. Value of collaboration among multi-domain experts in analysis of high-throughput genomics data. *Cancer Res.* (2019) 79:5140–5. doi: 10.1158/0008-5472

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

REVIEW
published: 21 August 2019
doi: 10.3389/fonc.2019.00794

Understanding Cancer Through the Lens of Epigenetic Inheritance, Allele-Specific Gene Expression, and High-Throughput Technology

*Maxwell P. Lee**

High Dimension Data Analysis Group, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States

Epigenetic information is characterized by its stable transmission during mitotic cell divisions and plasticity during development and differentiation. This duality is in contrast to genetic information, which is stable and identical in all cells in an organism with exception of immunoglobulin gene rearrangements in lymphocytes and somatic mutations in cancer cells. Allele-specific analysis of gene expression and epigenetic modifications provides a unique approach to studying epigenetic regulation in normal and cancer cells. Extension of Knudson's two-hits theory to include epigenetic alteration as a means to inactivate tumor suppressor genes provides better understanding of how genetic mutations and epigenetic alterations jointly contribute to cancer development. High-throughput technology has greatly accelerated cancer discovery. Large initiatives such as TCGA have shown that epigenetic components are frequent targets of mutations in cancer and these discoveries provide new insights into understanding cancer etiology and generate new opportunities for cancer therapeutics.

Keywords: epigenetics, cancer, allele, inheritance, therapy

OPEN ACCESS

Edited by:
Daoud Meerzaman, George Washington University, United States

Reviewed by:
Anna Maria Pinto, University of Siena, Italy
Jun Zhong, National Cancer Institute (NCI), United States

***Correspondence:**
Maxwell P. Lee
leemax@mail.nih.gov

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *05 June 2019*
Accepted: *06 August 2019*
Published: *21 August 2019*

Citation:
Lee MP (2019) Understanding Cancer Through the Lens of Epigenetic Inheritance, Allele-Specific Gene Expression, and High-Throughput Technology. Front. Oncol. 9:794. doi: 10.3389/fonc.2019.00794

INTRODUCTION

Epigenetics, first coined by Conrad Waddington in 1940s, was a conceptual model that describes the development process of forming a multicellular organism from a fertilized zygote (1). The concept had its root in the earlier studies in embryology and developmental biology. This epigenetic concept provided mechanisms that can bring about cellular changes in development and physiology but not involving changes of genetic materials. Although Mendel's work on genetic inheritance was well-recognized but the exact biochemical nature of genetic material was not known until a decade later when the double helix model of DNA was proposed in 1953 (2, 3). Following the discovery of the double helix structure of DNA, there was an explosion of studies to understand how DNA sequences were replicated and used as templates to synthesize mRNAs, and how mRNA sequences were translated to produce proteins, resulting in different cellular phenotypes and ultimately organism phenotypes (4). This was culminated as the central dogma of molecular biology in 1958 (5). A major focus of the biological research since that time was to elucidate the molecular mechanisms that underlie the differential gene expression programing in cellular differentiation in development, physiological response in daily activities, and pathological changes in diseases. The details emerging from these studies led to a general understanding of association among DNA methylation, gene expression, and physiological changes at the levels of

organisms and cells. The contemporary definition of epigenetics proposed by Holliday stated that epigenetics is the study of gene expression changes during cellular differentiation and mitotic inheritance of cellular gene expression pattern, which doesn't involve changes in DNA sequence (6). The expanded view of epigenetics includes many phenomena that can't be explained by Mendelian inheritance. Some prominent examples are X-chromosome inactivation and genomic imprinting in mammal and position effect variegation in *Drosophila* (7–11). Indeed, it was the study of these non-mendelian phenomena that largely initiated the identification and characterization of the biochemical components of epigenetic machinery. The current view of epigenetic system consists of DNA methylation, histone acetylation and methylation and other posttranslational modifications, chromatin remodeling complexes, and non-coding RNAs (12–15). Together, these epigenetic components control gene expression and form the basis of epigenetic memory that can be transmitted through mitotic cell division without DNA sequence changes.

There are numerous excellent reviews on epigenetics and cancer epigenetics. A few are cited here (12–14, 16, 17). In this short review, I will focus on a few selected topics that capture some aspects of epigenetics and epigenetic regulation in cancer from the perspective of epigenetic stability vs. plasticity and from the perspective of the allele-specific gene expression.

EPIGENETIC INHERITANCE AND PLASTICITY

Epigenetic information is characterized by its stable transmission during mitotic cell divisions and plasticity during development and differentiation. This duality differs from genetic information, which remains the same in every cell in an organism with the exception of a few cases such as immunoglobulin gene rearrangements in lymphocytes and somatic mutations in cancer. This duality is depicted in **Figure 1**. The two-states model provides a useful conceptual framework to think about epigenetic stability vs. plasticity.

Two classical examples of epigenetic phenomena are mammalian X chromosome inactivation and genomic imprinting (18, 19). Both are characterized by establishing active and inactive chromatin states in the two chromosomes in early embryogenesis, which are maintained during the life time of an organism. X chromosome inactivation involves gene expression silencing in one of the two X chromosomes, which ensures similar level of gene expression in both female and male cells. Which of the two X chromosomes to be inactivated is chosen randomly in the early embryogenesis. However, in the case of genomic imprinting, the inactivation occurs in specific genomic loci and the choice of which chromosome to be silenced is determined by the parental origin. Hundreds of imprinted genes have been identified. Some show gene silencing in all expressed tissues and during the entire life of the organism while others may display genomic imprinting only in selected tissues and affected by developmental stages and environmental exposure.

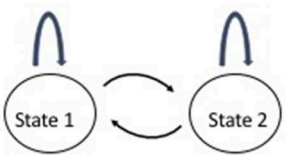


FIGURE 1 | Epigenetic stability vs. plasticity. The two states may represent any two conceptual epigenetic states such as an active chromatin vs. an inactive chromatin state or normal cell vs. cancer cell. The arc above the state represents maintenance of the state through events such as mitotic division whereas the arrows between the two states represent interconversion between the two states such as changes in chromatin structure during cellular differentiation, physiological response, or disease.

Studies of X chromosome inactivation and genomic imprinting played an instrumental role in establish DNA methylation and histone protein post-translational modifications and chromatin remodeling as the primary determinants of epigenetic state. There are many reasons why X chromosome and genomic imprinting are the excellent models to study epigenetics. The presence of a pair of active and inactive chromatin provides an ideal system to identify epigenetic marks that are specific for each epigenetic state but absent in the other epigenetic state. The DNA modification is relatively simple, involving methylation of the C5 in a cytosine (20). In mammalian genomes, the CpG dinucleotide occurs at much lower frequency than the other dinucleotides. This is because of the selective loss of CpG resulting from the conversion of 5-methylcytosine to thymine. However, there are genomic regions, where cluster of CpG dinucleotides are not methylated and consequently protected from the conversion, leading to the formation of CpG islands (CGIs) (21). About half of the mammalian genes contain CGIs, which are located near their transcription start sites.

Modifications of chromatin proteins are much more complex (12, 16). Both H3 and H4 histones undergo extensive post-translational modifications in their tails. These modifications include methylation, acetylation, phosphorylation, ubiquitination, etc. The combination of these modifications is referred to as the "histone code" (22), which carries the epigenetic information responsible for the maintenance of epigenetic state and dynamic change of epigenetic state.

From the perspective of epigenetic inheritance, DNA methylation state is maintained through DNA replication because semi-methylated DNA, the product of DNA replication, can be converted to fully methylated DNA by the action of DNMT1, which catalyze DNA methylation using semi-methylated DNA as substrate. DNA methyl transferase, DNMT3A, and DNMT3B, catalyze *de novo* methylation on DNA, thus providing a mechanism to acquire new DNA methylation marks to change chromatin state. However,

the effort of searching for enzymes that can catalyze DNA demethylation was unsuccessful until about 10 years ago. It led to the thinking in the past that perhaps DNA demethylation could be mediated only by passively losing half of the methylation during each cycle of DNA replication in the absence of DNMT1 activity. This was changed recently, when it was discovered that Ten-eleven translocation (TET) enzymes can catalyze demethylation of 5-methycytosine through sequential conversion of 5-methycytosine to 5-hydroxymethyl cytosine, to 5-formylcytosine, then to 5-carboxylcytosine, which can be converted to unmodified cytosine by terminal deoxynucleotidyl transferase (TDT) (23, 24).

Histone modifications are far more complicated than DNA methylation. But the general strategy is similar. There exist a pair of enzyme systems, histone post-translational modification "writers" and "erasers." For examples, histone acetyltransferase (HAT) serves as a writer whereas histone deacetylase (HDAC) serves as an eraser. Likewise, there are histone lysine methyltransferase (KMT) and histone lysine demethylase (KDM) to serve as writer and eraser, respectively. There are also protein arginine methyltransferases (PRMT), which act on arginine. Their opposing enzymes are peptidyl arginine deiminase (PADI). Each family also contains a large number of enzymes that can recognize specific substrate sequences. There is a third class of proteins called "readers" that can specifically bind to these post-translational modifications. For examples, bromodomain binds to acetylated lysine residue and chromodomain recognizes lysine methylation. Interestingly, histone acetyltransferase often contains bromodomain in addition to its activity to add acetylation to lysine. The multi-function structure of HAT enables it to catalyze acetylation in a processive manner to spread this post-translation mark (PTM) to the nearby nucleosome. This provides a potential mechanism for maintaining the PTM through mitotic division. Unlike DNA methylation, which produces semi-methylated DNA after DNA replication, the nucleosomes are randomly distributed into each of the two daughter cells after cell division. Half of the nucleosomes are derived from the parental cell and half are from newly deposited nucleosomes, which don't have PTMs. The ability of HAT to bind acetylated lysine and then catalyze addition of acetyl group to the nearby nucleosome allows the maintaining of this PTM through mitotic cell division.

The hallmark of epigenetics is the transmission of epigenetic marks through mitotic cell divisions. The duration of maintaining an epigenetic state varies. In the case of X chromosome inactivation and genomic imprinting, the active or inactive chromatin states are maintained throughout the lifetime. However, in most of cellular response to physiologic needs, the new epigenetic state is established and reversed back to normal state and the duration varies depending on particular physiology. DNA methylation mark is more stable while histone post-translational marks and other chromatin remodeling complexes display wide range of response time, serving different physiologic purposes.

ALLELE-SPECIFIC GENE EXPRESSION

X chromosome inactivation and genomic imprinting are characterized by the mono-allelic gene expression and epigenetic modifications. Mono-allelic gene expression also occurs in a number of other biological systems. In B lymphocytes, once an immunoglobulin gene rearrangement takes place on one chromosome, the rearrangement of the same gene from the other chromosome would be prevented. This phenomenum is termed as allelic exclusion, which ensures that an individual lymphocyte expresses a unique amino acid sequence of an immunoglobulin protein (25). Similar mechanism also operates in T lymphocytes for activating TCR genes (26). Another example of mono-allelic expression is the expression of human olfactory receptor genes. There are about 1,000 olfactory receptor genes, each of which is expressed from only one chromosome in a sensory neuron (27).

In addition, quantitative differences in the degree of gene expression between two alleles, marked with SNPs, are a widespread phenomenon, hereinafter referred to as allele-specific expression (ASE). We initially studied allele-specific gene expression using the Affymetrix SNP arrays and found extensive allelic variation in expression in the human genome (28). ASE differs from mono-allelic expression described in the previous sections. ASE showed differential gene expression between the two alleles in the range of 2–4 fold, which is in contrast to mono-allelic expression observed in imprinting and X chromosome inactivation. ASE is commonly affected by genetic polymorphisms near the gene and these polymorphisms play a regulatory role affecting gene expression (29, 30). This is particularly relevant since most of the GWAS identified SNPs are located in intragenic or intergenic regions. These SNPs impact phenotypes through gene expression regulation at the epigenetic level or post-transcriptional level.

UNDERSTANDING CANCER FROM THE PERSPECTIVE OF EPIGENETIC REGULATION AND ALLELE-SPECIFIC GENE EXPRESSION

It is well-established that cancer is caused by mutations that are acquired either from parents through germline inheritance or generated in somatic cells. Inactivating tumor suppressor genes and activating oncogenes both contribute to cancer development. In the case of inactivation of tumor suppressor genes, both alleles have to be inactivated in the cancer cell. This is best illustrated by Knudson's two-hits theory (31). The two-hits theory was postulated to explain why familial retinoblastoma develops earlier and bilateral while the sporadic retinoblastoma develops later and often unilateral. Based on the epidemiologic observation, Knudson hypothesized that retinoblastoma was caused by inactivation of both alleles of a tumor suppressor gene and in the case of familial syndrome one allele was inactivated in germline and the 2nd allele was inactivated in somatic tumor whereas in the case of sporadic cancer both alleles were inactivated in the somatic tumor. This paradigm can extend to include epigenetic alteration as a means to inactivate

tumor suppressor genes. Many tumor suppressor genes, such as BRCA1/2 and CDKN2A/B, are frequently silenced by DNA methylation and inactive chromatin marks (32–34).

Epigenetic alteration through germline inheritance can occur in familial cancer syndrome. Lynch syndrome is caused by mutations in the genes involved in DNA mismatch repair. Germline mutations in MLH1 and MSH2 causes the majority of Lynch syndrome. However, in several Lynch syndrome families, no mutations in mismatch repairs genes were found despite extensive effort of searching for causative mutations. Instead, heritable DNA methylation in the promoter regions of MLH1 or MSH2 was identified that silenced gene expression. Chan et al. analyzed a three-generation family using allele-specific methylation (ASM) and reported that methylation of MSH2 gene in the germ line cells correlated with the loss of the MSH2 protein in the colorectal adenocarcinomas (35). Besides silencing gene expression by DNA methylation, a somatic frameshift mutation was found in MSH2. The authors concluded that ASM in germline transmission was the first hit while the somatic mutation was the 2nd hit. In a separate study of a HNPCC family, the EPCAM gene had a germline deletion in the 3′ end and resulted transcription read-through into downstream MSH2 and an increase in DNA methylation in the promoter region of MSH2 (36). The deletion was co-segregated with ASM of the MSH2 promoter and the disease. Epimutation of the RB1 gene was recently found in a six generations retinoblastoma family (37). The germline methylation was inherited from the maternal chromosome. Interestingly, the detailed pedigree analysis also found a germline mutation that was transmitted through the paternal chromosome and showed incomplete penetrance. The authors concluded that both genetic mutation and epimutation contributed to the retinoblastoma in this family. A rare epimutation in the RB1 gene was also identified from another recent study (38). The authors showed that germline DNA methylation was associated with silencing of the RB1 gene expression.

Beckwith–Wiedemann syndrome (BWS), is another familial syndrome, which increases risk of developing multiple pediatric cancers. It has served as a model system for studying genomic imprinting and how abnormal genomic imprinting causes cancer. Multiple genetic and epigenetic mechanisms were identified that cause BWS, including mutation in CDKN1C (39), loss of imprinting in IGF2 (40), translocation involving KCNQ1 (41), and abnormal imprinting of a lincRNA, KCNQ1OT1 (42). All four genes are imprinted. IGF2 is normally expressed from the paternal chromosome but expressed from both chromosomes in tumors. KCNQ1OT1 is normally methylated on maternal chromosome but the methylation is frequently lost in BWS patient germline DNA. Allele-specific gene expression and allele-specific methylation analysis have played an instrumental role in elucidating various epigenetic mechanisms (43).

Two papers brought about wide appreciation of quantitative difference in gene expression between two alleles of APC in familial adenomatous polyposis (FAP) (44, 45). Yan et al. showed that 50% reduction in gene expression in APC was associated with predisposition to FAP. They studied six patients from two FAP families and didn't find any mutation in the APC gene. However, using ASE, they found 2-fold difference in gene expression between the two alleles in all 6 patients. Furthermore, tumors displayed loss of heterozygosity (LOH) and deleted specifically the high-expression allele.

An interesting question is whether DNA methylation causes inactive chromatin state or vice versa. An elegant study from Vogelstein's lab provided an important insight into answering this question (46). They generated double knockout of DNMT1 and DNMT3B, which eliminated most of DNA methylation in HCT116, a colon cancer cell line. The double knockout cells had slow growth rate in early passage cells but the late passage cells grow to comparable rate to the parental cells. This corresponded to a gradual increase in p16 methylation and consequently silencing of p16. The kinetics of chromatin marks changes was faster. Histone H4-acetylation increased as early as the passage 5 and H3K9-methylation appeared at the passage 22. But DNA methylation appeared at the passage 50. The p16 is heterozygous in HCT116, allowing tracking of both alleles for allele-specific analysis of gene expression, DNA methylation, and chromatin marks. The study revealed that only the wild type allele showed dynamic changes of epigenetic marks. These observations established that the order of epigenetic changes in this system was that it began with the gain of H4-acetylation and expression of p16, followed by H3K9 methylation and silencing of p16 expression and faster growth, and eventually cells fully re-gained DNA methylation and lose H4-acetylation mark and grew at comparable rate as the parental cells. The work also demonstrated the important role of silencing p16 in driving cellular proliferation.

ACCELERATING CANCER DISCOVERY WITH HIGH-THROUGHPUT TECHNOLOGY

The high-throughput analysis of gene mutations in human cancer was made possible after the human genome sequencing was completed in 2003. Some of the earliest studies that leveraged human genome sequence data to systematically identify mutated genes in human cancer were reported by the researchers from Johns Hopkins and Sanger Institute. These included the large scale analysis of coding sequences of human transcriptome in breast and colorectal cancer from Johns Hopkins in 2006 and 2007 (47, 48) and analysis of the coding exons of 518 protein kinase genes in multiple human cancers from Sanger Institute in 2007 (49). In 2005, NCI and NHGRI initiated The Cancer Genome Atlas (TCGA) initiative to comprehensively characterize genomic alterations in all major cancers. The pilot project was initially focused on glioblastoma multiforme (GBM) and ovary cancer, and it was extended to include more than 30 types of cancer in 2010. The first paper published from the TCGA initiative was the comprehensive analysis of mutation, DNA methylation, and gene expression of GBM in 2008 (50).

A very interesting study came from the comprehensive analysis of mutation in glioblastoma multiforme (GBM) by the Hopkins team in 2008, which led to the discovery of a recurrent mutation R132H in isocitrate dehydrogenase 1 (IDH1) and the mutation was shown to be associated with better survival (51).

The R132H mutation was always present as a heterozygous mutation, suggesting it functions as an oncogene. Detailed biochemical study showed that the IDH1 mutation generated 2-hydroxyglutarate (2HG) instead of alpha-ketoglutarate, which is the normal product of the wildtype enzyme (52). Excessive accumulation of 2HG contributed to the formation of gliomas, suggesting that 2HG acted as an onco-metabolite.

In 2010, TCGA team found that a subset of GBM has high CpG island methylation, which was termed as a glioma CpG island methylator phenotype (G-CIMP) (53). G-CIMP tumors were more prevalent among lower-grade gliomas and associated with IDH1 somatic mutations. The association between DNA methylation and IDH1 was not unique to GBM and it also occurred in acute myeloid leukemia (AML) (54). What was really intriguing was the finding that mutations in IDH1/2 and TET2 were mutually exclusive in AML. TET2 was known to catalyze the conversion of 5 methyl cytosine to 5 hydroxy methyl cytosine. This immediately suggested that IDH mutation inhibits TET2 activity. Indeed, expression of IDH1 mutant inhibited the production of 5 hydroxy methyl cytosine. The mechanism for the inhibition was because IDH mutants produced 2-hydroxyglutarate instead of 2-oxoglutarate. 2-oxoglutarate was co-factor for TET2 to catalyze hydroxy methylation whereas 2-oxoglutarate served as a competitive inhibitor to TET2. Therefore, either Tet2 mutation or IDH2 could cause accumulation of 5 methyl cytosine, generating the CpG island methylation phenotype. This explained why either IDH1/2 or TET2 mutation could block hematopoietic differentiation and cause proleukemogenic effect.

One of the emerging concepts from the high-throughput mutational analysis of human cancer genomes was the finding that chromatin components are the frequent targets of mutations in human cancer (55–57). Some examples are provided here. Recurrent mutations of the histone methyltransferase MLL2 were detected in 89% of follicular lymphoma (FL) and 32% of diffuse large B-cell lymphoma (DLBCL) (58). The histone H3K27 demethylase UTX was mutated in multiple human cancers (59). Mutations in EZH2, a histone H3K27 methyltransferase, was found in GCB subtype of DLBCL and follicular lymphoma (60). Mutations in DNA methyltransferase DNMT3A were identified in 25% of acute myeloid leukemia (AML) (61). Not only the histone modifiers were frequently mutated, but histone proteins were also the direct targets of mutations in human cancer. Both histone H3 variant H3.3 (H3F3A) and the histone H3.1 (HIST1H3B) were mutated in 30% of pediatric glioblastomas (62, 63). Interestingly, these mutations occurred at the specific sites, K27M and G34R/G34V, and were present in heterozygous. Detailed biochemical studies showed that H3K27M mutant acted in a dominant-negative manner to inhibit PRC2 activities and consequently reduced H3K27me3 level (64).

EPIGENETIC DYNAMICS IN CANCER TREATMENT

DNA methyltransferase inhibitors, 5-azacytidine (Vidaza) and 5-aza-2′-deoxycytidine (decitabine), are FDA approved drugs for myelodysplastic syndrome (MDS) and AML patients. Treatment with 5-azacitidine and decitabine increased overall survival in MDS patients than conventional care in phase III clinical trials (65, 66). The response rates are between 30 and 60%. The response in myeloid malignancies are better than lymphoid leukemia or solid tumors (67). This might be related to the observation that myeloid leukemia has a relatively low mutational burden but has mutations in the genes involved in controlling DNA methylation, such as TET2 or DNMT3A (68). Many studies were conducted to understand what the clinical factors and molecular alterations are associated with treatment response, only TET2 mutation was found to be weakly associated with clinical response to therapy (69, 70). DNA methyltransferase inhibitors have bi-modal activities. At low dose, they cause hypomethylation whereas at high dose, they are cytotoxic. Following treatment, there was global decrease in DNA methylation and hypomethylation was associated with better response (71). Hypomethylation of specific tumor suppressor genes such as CDKN2B was also observed, which was associated with reactivation of protein expression to a normal level (72). This is consistent with the mechanism of drug action.

Vorinostat (SAHA), belinostat (PXD101), and romidepsin are FDA approved histone deacetylase (HDAC) inhibitors for cutaneous T cell lymphoma (CTCL) patients (73). The response rates are between 30 and 40% of patients with CTCL (74–76). Panobinostat in combination with the proteasome inhibitor bortezomib is FDA approved for the treatment of drug-resistant multiple myeloma (77). However, the success of these drugs is limited to cutaneous T cell lymphoma and multiple myeloma, and they are not effective for solid tumors. Similar to DNA methyltransferase inhibitors, HDAC inhibitors also show bi-modal activities. Their efficacy is dose-dependent, and the drugs are cytotoxic at high dose (78). The drug targets are more complex since there are eleven HDACs and also many non-histone targets. The targets could be nuclear or cytoplasmic. The complexity makes it hard to predict what factors could determine how well patients respond to treatment.

Besides targeting DNA methyltransferases and histone deacetylase, recently identified mutations in histone modifiers and chromatin remodeling proteins offer new opportunities for targeted therapy (55, 79). These include development of JQ1 and I-BET that bind to acetyl lysine recognition motifs of bromodomain and extra-terminal (BET) of BRD4 (80, 81), which is involved in DNA translocation in several cancers and activation of MYC oncogene; development of an inhibitor of H3K79 *N*-methyltransferase (DOT1L), which is involved in leukemogenesis in mixed lineage leukemia (MLL) (82); development of a small molecule GSK2879552 that inhibits lysine demethylase 1 (LSD1) (83).

A major concern in cancer therapy, either chemotherapy or targeted therapy, is the development of resistance to cancer drugs. Many mechanisms contribute to drug resistance, including drug efflux and mutations in the targeted genes or related pathways. However, recent studies suggested that epigenetic alterations could provide another mechanism to acquire drug resistance, especially for slowly acquired resistance (84). The drug resistance involved activation of IGF1 signaling

pathway and chromatin alteration mediated by the histone demethylase KDM5A in a small population of drug-tolerant cells. Treatment with IGF1 receptor inhibitors or HDAC inhibitors can eliminate the drug-tolerant cells. The combination of chemo with HDAC inhibitors provide a potential new strategy to prevent development of drug resistance.

In conclusion, the studies of epigenetics and allele-specific gene expression and application of high-throughput technology provide powerful approaches to enhancing our understanding of cancer etiology and progression and also provide new opportunity for cancer therapeutics. There are some limitations when we try to understand cancer through the lens of epigenetic inheritance, allele-specific gene expression, and high-throughput technology. Germline epimutations are very rare events and some of which may be caused by yet unknown genetic variants. Although allele-specific gene expression can provide a unique perspective on the role of genetic variants on gene expression regulation, we are often more interested in the combined gene expression contributed from both alleles and how the gene expression is associated with other biological phenomena. High-throughput technology is powerful for the discovery phase of the research. However, new findings should be rigorously validated by additional experiments.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was supported by the Intramural Research Program of the NIH and the National Cancer Institute, Center for Cancer Research (CCR).

REFERENCES

1. Waddington CH. The epigenotype. *Endeavour.* (1942) 1:18–20.
2. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature.* (1953) 171:964–7. doi: 10.1038/171964b0
3. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* (1953) 171:737–8. doi: 10.1038/171737a0
4. Cobb M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol.* (2017) 15:e2003243. doi: 10.1371/journal.pbio.2003243
5. Crick FH. On protein synthesis. *Symp Soc Exp Biol.* (1958) 12:138–63.
6. Holliday R. DNA methylation and epigenetic inheritance. *Philos Trans R Soc Lond B Biol Sci.* (1990) 326:329–38. doi: 10.1098/rstb.1990.0015
7. Bartolomei MS, Zemel S, Tilghman SM. Parental imprinting of the mouse H19 gene. *Nature.* (1991) 351:153–5. doi: 10.1038/351153a0
8. DeChiara TM, Robertson EJ, Efstratiadis A. Parental imprinting of the mouse insulin-like growth factor II gene. *Cell.* (1991) 64:849–59. doi: 10.1016/0092-8674(91)90513-X
9. Ferguson-Smith AC, Cattanach BM, Barton SC, Beechey CV, Surani MA. Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature.* (1991) 351:667–70. doi: 10.1038/351667a0
10. Goldschmidt RB. Chromosomes and genes. *Cold Spring Harb Symp Quant Biol.* (1951) 16:1–11. doi: 10.1101/SQB.1951.016.01.003
11. Lyon MF. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature.* (1961) 190:372–3. doi: 10.1038/190372a0
12. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet.* (2016) 17:487–500. doi: 10.1038/nrg.2016.59
13. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer.* (2004) 4:143–53. doi: 10.1038/nrc1279
14. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* (2003) 33(Suppl. 1):245–54. doi: 10.1038/ng1089
15. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nat Rev Genet.* (2002) 3:415–28. doi: 10.1038/nrg816
16. Berger SL. Histone modifications in transcriptional regulation. *Curr Opin Genet Dev.* (2002) 12:142–8. doi: 10.1016/S0959-437X(02)00279-4
17. Feinberg AP. The epigenetics of cancer etiology. *Semin Cancer Biol.* (2004) 14:427–32. doi: 10.1016/j.semcancer.2004.06.005
18. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol.* (2014) 6:a018382. doi: 10.1101/cshperspect.a018382
19. Brockdorff N, Turner BM. Dosage compensation in mammals. *Cold Spring Harb Perspect Biol.* (2015) 7:a019406. doi: 10.1101/cshperspect.a019406
20. Bird AP. Functions for DNA methylation in vertebrates. *Cold Spring Harb Symp Quant Biol.* (1993) 58:281–5. doi: 10.1101/SQB.1993.058.01.033
21. Bird AP. CpG-rich islands and the function of DNA methylation. *Nature.* (1986) 321:209–13. doi: 10.1038/321209a0
22. Jenuwein T, Allis CD. Translating the histone code. *Science.* (2001) 293:1074–80. doi: 10.1126/science.1063127
23. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature.* (2010) 466:1129–33. doi: 10.1038/nature09303
24. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science.* (2011) 333:1300–3. doi: 10.1126/science.1210597
25. Storb U, Ritchie KA, O'Brien R, Arp B, Brinster R. Expression, allelic exclusion and somatic mutation of mouse immunoglobulin kappa genes. *Immunol Rev.* (1986) 89:85–102. doi: 10.1111/j.1600-065X.1986.tb01474.x
26. Khor B, Sleckman BP. Allelic exclusion at the TCRbeta locus. *Curr Opin Immunol.* (2002) 14:230–4. doi: 10.1016/S0952-7915(02)00326-6
27. Serizawa S, Miyamichi K, Sakano H. One neuron-one receptor rule in the mouse olfactory system. *Trends Genet.* (2004) 20:648–53. doi: 10.1016/j.tig.2004.09.006
28. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, et al. Allelic variation in gene expression is common in the human genome. *Genome Res.* (2003) 13:1855–62. doi: 10.1101/gr.1006603
29. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet.* (2003) 33:422–5. doi: 10.1038/ng1094
30. Knight JC, Keating BJ, Kwiatkowski DP. Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1. *Nat Genet.* (2004) 36:394–9. doi: 10.1038/ng1331
31. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA.* (1971) 68:820–3. doi: 10.1073/pnas.68.4.820
32. Belinsky SA, Nikula KJ, Palmisano WA, Michels R, Saccomanno G, Gabrielson E, et al. Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. *Proc Natl Acad Sci USA.* (1998) 95:11891–6. doi: 10.1073/pnas.95.20.11891
33. Dobrovic A, Simpfendorfer D. Methylation of the BRCA1 gene in sporadic breast cancer. *Cancer Res.* (1997) 57:3347–50.
34. Esteller M. Epigenetic lesions causing genetic lesions in human cancer: promoter hypermethylation of DNA repair genes. *Eur J Cancer.* (2000) 36:2294–300. doi: 10.1016/S0959-8049(00)00303-8
35. Chan TL, Yuen ST, Kong CK, Chan YW, Chan AS, Ng WF, et al. Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. *Nat Genet.* (2006) 38:1178–83. doi: 10.1038/ng1866
36. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, et al. Heritable somatic methylation and inactivation of MSH2 in families with

Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat Genet.* (2009) 41:112–7. doi: 10.1038/ng.283
37. Quiñonez-Silva G, Dávalos-Salas M, Recillas-Targa F, Ostrosky-Wegman P, Aranda DA, Benítez-Bribiesca L. Monoallelic germline methylation and sequence variant in the promoter of the RB1 gene: a possible constitutive epimutation in hereditary retinoblastoma. *Clin Epigenet.* (2016) 8:1. doi: 10.1186/s13148-015-0167-0
38. Gelli E, Pinto AM, Somma S, Imperatore V, Cannone MG, Hadjistilianou T, et al. Evidence of predisposing epimutation in retinoblastoma. *Hum Mutat.* (2019) 40:201–6. doi: 10.1002/humu.23684
39. Hatada I, Ohashi H, Fukushima Y, Kaneko Y, Inoue M, Komoto Y, et al. An imprinted gene p57KIP2 is mutated in Beckwith-Wiedemann syndrome. *Nat Genet.* (1996) 14:171–3. doi: 10.1038/ng1096-171
40. Rainier S, Johnson LA, Dobry CJ, Ping AJ, Grundy PE, Feinberg AP. Relaxation of imprinted genes in human cancer. *Nature.* (1993) 362:747–9. doi: 10.1038/362747a0
41. Lee MP, Hu RJ, Johnson LA, Feinberg AP. Human KVLQT1 gene shows tissue-specific imprinting and encompasses Beckwith-Wiedemann syndrome chromosomal rearrangements. *Nat Genet.* (1997) 15:181–5. doi: 10.1038/ng0297-181
42. Lee MP, DeBaun MR, Mitsuya K, Galonek HL, Brandenburg S, Oshimura M, et al. Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc Natl Acad Sci USA.* (1999) 96:5203–8. doi: 10.1073/pnas.96.9.5203
43. Lee MP. Allele-specific gene expression and epigenetic modifications and their application to understanding inheritance and cancer. *Biochim Biophys Acta.* (2012) 1819:739–42. doi: 10.1016/j.bbagrm.2012.02.007
44. Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM, et al. Small changes in expression affect predisposition to tumorigenesis. *Nat Genet.* (2002) 30:25–6. doi: 10.1038/ng799
45. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. *Science.* (2002) 297:1143. doi: 10.1126/science.1072545
46. Bachman KE, Park BH, Rhee I, Rajagopalan H, Herman JG, Baylin SB, et al. Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. *Cancer Cell.* (2003) 3:89–95. doi: 10.1016/S1535-6108(02)00234-9
47. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* (2006) 314:268–74. doi: 10.1126/science.1133427
48. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* (2007) 318:1108–13. doi: 10.1126/science.1145720
49. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature.* (2007) 446:153–8. doi: 10.1038/nature05610
50. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* (2008) 455:1061–8. doi: 10.1038/nature07385
51. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* (2008) 321:1807–12. doi: 10.1126/science.1164382
52. Dang L, White DW, Gross S, Bennett BD, Bittinger MA, Driggers EM, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature.* (2009) 462:739–44. doi: 10.1038/nature08617
53. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell.* (2010) 17:510–22. doi: 10.1016/j.ccr.2010.03.017
54. Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell.* (2010) 18:553–67. doi: 10.1016/j.ccr.2010.11.015
55. Audia JE, Campbell RM. Histone modifications and cancer. *Cold Spring Harb Perspect Biol.* (2016) 8:a019521. doi: 10.1101/cshperspect.a019521
56. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell.* (2012) 150:12–27. doi: 10.1016/j.cell.2012.06.013
57. Garraway LA, Lander ES. Lessons from the cancer genome. *Cell.* (2013) 153:17–37. doi: 10.1016/j.cell.2013.03.002
58. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature.* (2011) 476:298–303. doi: 10.1038/nature10351
59. van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, et al. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet.* (2009) 41:521–3. doi: 10.1038/ng.349
60. Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, Goya R, et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet.* (2010) 42:181–5. doi: 10.1038/ng.518
61. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med.* (2010) 363:2424–33. doi: 10.1056/NEJMoa1005143
62. Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, Jacob K, et al. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature.* (2012) 482:226–31. doi: 10.1038/nature10833
63. Wu G, Broniscer A, McEachron TA, Lu C, Paugh BS, Becksfort J, et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet.* (2012) 44:251–3. doi: 10.1038/ng.1102
64. Lewis PW, Müller MM, Koletsky MS, Cordero F, Lin S, Banaszynski LA, et al. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. *Science.* (2013) 340:857–61. doi: 10.1126/science.1232245
65. Fenaux P, Mufti GJ, Hellstrom-Lindberg E, Santini V, Finelli C, Giagounidis A, et al. Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study. *Lancet Oncol.* (2009) 10:223–32. doi: 10.1016/S1470-2045(09)70003-8
66. Lübbert M, Suciu S, Hagemeijer A, Rüter B, Platzbecker U, Giagounidis A, et al. Decitabine improves progression-free survival in older high-risk MDS patients with multiple autosomal monosomies: results of a subgroup analysis of the randomized phase III study 06011 of the EORTC Leukemia Cooperative Group and German MDS Study Group. *Ann Hematol.* (2016) 95:191–9. doi: 10.1007/s00277-015-2547-0
67. Issa JP, Kantarjian HM. Targeting DNA methylation. *Clin Cancer Res.* (2009) 15:3938–46. doi: 10.1158/1078-0432.CCR-08-2783
68. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med.* (2013) 368:2059–74. doi: 10.1056/NEJMoa1301689
69. Bejar R, Lord A, Stevenson K, Bar-Natan M, Pérez-Ladaga A, Zaneveld J, et al. TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. *Blood.* (2014) 124:2705–12. doi: 10.1182/blood-2014-06-582809
70. Itzykson R, Kosmider O, Cluzeau T, Mansat-De Mas V, Dreyfus F, Beyne-Rauzy O, et al. Impact of TET2 mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. *Leukemia.* (2011) 25:1147–52. doi: 10.1038/leu.2011.71
71. Yang AS, Doshi KD, Choi SW, Mason JB, Mannari RK, Gharybian V, et al. DNA methylation changes after 5-aza-2'-deoxycytidine therapy in patients with leukemia. *Cancer Res.* (2006) 66:5495–503. doi: 10.1158/0008-5472.CAN-05-2385
72. Daskalakis M, Nguyen TT, Nguyen C, Guldberg P, Köhler G, Wijermans P, et al. Demethylation of a hypermethylated P15/INK4B gene in patients with myelodysplastic syndrome by 5-Aza-2'-deoxycytidine (decitabine) treatment. *Blood.* (2002) 100:2957–64. doi: 10.1182/blood.V100.8.2957
73. Falkenberg KJ, Johnstone RW. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat Rev Drug Discov.* (2014) 13:673–91. doi: 10.1038/nrd4360
74. Duvic M, Talpur R, Ni X, Zhang C, Hazarika P, Kelly C, et al. Phase 2 trial of oral vorinostat (suberoylanilide hydroxamic acid, SAHA) for refractory cutaneous T-cell lymphoma (CTCL). *Blood.* (2007) 109:31–9. doi: 10.1182/blood-2006-06-025999
75. Piekarz RL, Frye R, Turner M, Wright JJ, Allen SL, Kirschbaum MH, et al. Phase II multi-institutional trial of the histone deacetylase inhibitor romidepsin as monotherapy for patients with cutaneous T-cell lymphoma. *J Clin Oncol.* (2009) 27:5410–7. doi: 10.1200/JCO.2008.21.6150

76. Whittaker SJ, Demierre MF, Kim EJ, Rook AH, Lerner A, Duvic M, et al. Final results from a multicenter, international, pivotal study of romidepsin in refractory cutaneous T-cell lymphoma. *J Clin Oncol.* (2010) 28:4485–91. doi: 10.1200/JCO.2010.28.9066
77. San-Miguel JF, Hungria VT, Yoon SS, Beksac M, Dimopoulos MA, Elghandour A, et al. Panobinostat plus bortezomib and dexamethasone versus placebo plus bortezomib and dexamethasone in patients with relapsed or relapsed and refractory multiple myeloma: a multicentre, randomised, double-blind phase 3 trial. *Lancet Oncol.* (2014) 15:1195–206. doi: 10.1016/S1470-2045(14)70440-1
78. Azad N, Zahnow CA, Rudin CM, Baylin SB. The future of epigenetic therapy in solid tumours–lessons from the past. *Nat Rev Clin Oncol.* (2013) 10:256–66. doi: 10.1038/nrclinonc.2013.42
79. Jones PA, Issa JP, Baylin S. Targeting the cancer epigenome for therapy. *Nat Rev Genet.* (2016) 17:630–41. doi: 10.1038/nrg.2016.93
80. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O, et al. Selective inhibition of BET bromodomains. *Nature.* (2010) 468:1067–73. doi: 10.1038/nature09504
81. Nicodeme E, Jeffrey KL, Schaefer U, Beinke S, Dewell S, Chung CW, et al. Suppression of inflammation by a synthetic histone mimic. *Nature.* (2010) 468:1119–23. doi: 10.1038/nature09589
82. Daigle SR, Olhava EJ, Therkelsen CA, Majer CR, Sneeringer CJ, Song J, et al. Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell.* (2011) 20:53–65. doi: 10.1016/j.ccr.2011.06.009
83. Mohammad HP, Smitheman KN, Kamat CD, Soong D, Federowicz KE, Van Aller GS, et al. A DNA hypomethylation signature predicts antitumor activity of LSD1 inhibitors in SCLC. *Cancer Cell.* (2015) 28:57–69. doi: 10.1016/j.ccell.2015.06.002
84. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell.* (2010) 141:69–80. doi: 10.1016/j.cell.2010.02.027

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JZ declared a shared affiliation, though no other collaboration, with the author to the handling editor.

frontiers in Oncology

ORIGINAL RESEARCH
published: 06 September 2019
doi: 10.3389/fonc.2019.00788

Identification of Eight Small Nucleolar RNAs as Survival Biomarkers and Their Clinical Significance in Gastric Cancer

*Xuning Wang, Maolin Xu, Yongfeng Yan, Yanshen Kuang, Peng Li, Wei Zheng, Hongyi Liu and Baoqing Jia**

Department of General Surgery, Chinese PLA General Hospital, Beijing, China

Gastric cancer is one of most common cancers worldwide. Studies have shown that small nucleolar RNAs (snoRNAs) play important roles in several cancers. In this study, we analyzed the snoRNAs that were differentially expressed between gastric tumors and normal tissues, identified survival-associated snoRNAs, and developed an eight-snoRNA signature to predict overall survival of patients with gastric cancer. Furthermore, we explored the clinical significance of the eight signature snoRNAs. The risk biomarker established by the eight snoRNA signature was an independent prognostic factor (hazard ratio = 3.43, 95% confidence interval: 1.93–6.09, P = 2.72e-05). Furthermore, we validated the expression pattern of those snoRNAs in different gastric cancer cell lines and 5 paired normal and tumor tissues by using real time quantification PCR. Knocking down U66, one of the eight snoRNAs, inhibited the cell proliferation. In conclusion, we identified an eight-snoRNA risk signature to predict overall survival of gastric cancer patients. Seven of these snoRNAs were associated with clinical features of the disease. Knocking down U66 inhibited cell proliferation. These findings provide new clues with prognostic and therapeutic implications in gastric cancer.

Keywords: small nucleolar RNA, biomarker, gastric cancer, survival, risk signature

OPEN ACCESS

Edited by:
Daoud Meerzaman, George Washington University, United States

Reviewed by:
Longqiang Wang, University of Texas MD Anderson Cancer Center, United States
Tania Lee Slatter, University of Otago, New Zealand

****Correspondence:***
Baoqing Jia
baoqingjia@126.com

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *22 April 2019*
Accepted: *05 August 2019*
Published: *06 September 2019*

Citation:
Wang X, Xu M, Yan Y, Kuang Y, Li P, Zheng W, Liu H and Jia B (2019) Identification of Eight Small Nucleolar RNAs as Survival Biomarkers and Their Clinical Significance in Gastric Cancer. Front. Oncol. 9:788. doi: 10.3389/fonc.2019.00788

INTRODUCTION

Gastric cancer (GC) is one of the leading causes of cancer-related death around the world and is the second and third most common cancer in men and women, respectively, in China (1). Many factors contribute to the genesis of GC such as methylation of genes (2), copy number variation (3, 4), positive family history of GC, cigarette smoking, and low consumption of fruits (5). Compared with other cancers, the prognosis is poor with a 5-year survival rate less than 40% (6). This is in part because there are no strong genetic biomarkers for GC. As a result, new biomarkers to improve the predictive value of the incidence and prognosis of GC are desperately needed. Such biomarkers could help to understand cancer pathogenesis and provide personalized treatment.

Small nucleolar RNAs (snoRNAs) are a class of small non-coding RNA molecules, 60–300 base pairs in length. They are encoded predominantly in introns of host genes in vertebrates, and guide site-specific chemical modifications of ribosomes, transfer RNAs, and small nuclear RNAs. There are two main classes of snoRNAs based on sequence motifs and secondary structural elements: C/D box and H/ACA box snoRNAs. Because of advances in next generation

sequencing and experimental and computational approaches, many snoRNAs and their functions are being identified. However, there are many orphan snoRNAs that have no known targets or specific functions.

Recent studies described snoRNAs that displayed unique characteristics and expression patterns, as well as interacting with corresponding protein partners and performing various functions. Increasing attention is being paid to cancer-related snoRNAs. For example, growth arrest-specific transcript 5-associated snoRNAs correlated with TP53 expression and DNA damage in colorectal cancer (7). In addition, C/D-box snoRNAs are associated with metastatic progression and malignant transformation in prostate cancer (8). Finally, snoRNAs and fibrillarin, an enzymatic small nucleolar ribonucleoprotein, are frequently upregulated in human breast and prostate cancers, and those upregulated snoRNAs play crucial roles in tumorigenicity both *in vivo* and *in vitro* (9).

Overall, the results of these studies support the importance of snoRNAs in cell biological processes. Understanding the molecular mechanisms underlying the development of GC is essential for cancer diagnosis and therapy. However, the functions of snoRNAs in GC remain elusive. In the current study, we identified differentially expressed snoRNAs, developed a snoRNA-based signature to predict overall survival of patients with GC, and explored the potential clinical significance of snoRNAs.

MATERIALS AND METHODS

Data Collection and Processing

SnoRNA expression data (fragments per million kilobases for each snoRNA) were downloaded from SNORic, a website used to explore snoRNAs in different cancers with data from The Cancer Genome Atlas (10), and corresponding clinical follow-up data from The Cancer Genome Atlas data portal. **Figure 1** shows the main workflow. We filtered snoRNAs that were expressed at least 30% of samples and removed patients without complete clinical information. In total, 37 normal tissues and 349 tumor samples were included in this study. These tumor samples were assigned randomly into a training set (50%, 174), that was used to develop a risk signature and a test set (50%, 175), to verify the performance of the snoRNA signature. There was no significant difference in demographic characteristics between the training and test sets. The basic clinical information is shown in **Table 1**. Overall, 324 snoRNA profiles were acquired for all patients. This study meets the publication guideline of TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/citing-tcga). As the data used in the study was obtained from public datasets, there was no need for additional written consent.

Identification of Differentially Expressed and Survival-Related snoRNAs

The presence of snoRNAs that were differentially expressed between normal and tumor tissues was analyzed by the *t*-test. A univariate COX proportional regression was applied to identify survival-related snoRNAs. The 30 snoRNAs with the lowest *P*-values were put into a robust likelihood model by the rbsurvR package (11). Firstly, the model placed N*(1 – p) samples randomly into the internal training set, and N*p samples into the validation set. Here, we chose $p = 1/2$. Secondly, the model placed a snoRNA into the training set and calculated the parameter for this snoRNA. Then the logLik for each snoRNA was evaluated with the above parameter, including validation in the internal validation samples. Finally, this model computed the Akaike information criterion, which is an estimator of the relative quality of statistical models for a given data set. We chose the optimal model with the smallest Akaike information criterion. $P < 0.05$ was considered statistically significant.

Establishment and Validation of the Risk Formula

SnoRNAs were chosen with the criteria mentioned above and a multivariate Cox analysis was used to calculate coefficients in the training set to establish risk formula by which a risk score for each sample was calculated. All patients were classified into two different groups (high and low risk) based on the median of the risk score. The Kaplan-Meier method and log-rank test were applied to analyze the overall survival of the two groups by using the R package of survival (12, 13). To evaluate the predictive value of the risk model, a receiver operating characteristic (ROC) curve was constructed using the R package of survivalROC (14). Figures were plotted by ggplot2 (15) and ggfortify (16).

Exploration of the Clinical Significance

We analyzed the expression patterns of snoRNAs that were identified by the risk formula signature. Clinical correlation [Lauren class molecular (17), neoplasm histologic grade, and pathologic stage subtypes] analyses were obtained from SNORic (10).

Experiment Validation

Real-time quantitative PCR was used to measure the expression prolife of snoRNAs in five gastric cancer cell lines (SGC-7901, BGC-823, NCI-N87, MGC-803, and AGS) and one normal gastric mucosal cell line (GES-1). The primer sequence of the snoRNAs was presented in **Supplementary Table 1**. The PCR product was sequenced by Sanger method and blast in NCBI, which indicated seven of eight primers work well (**Supplementary Figures S1, S2**). We collected five patients' tumor and adjacent tissue from surgical specimens which has been approved by Ethics Committee of our hospital. According the expression profile, we selected U66 to test its function. Small interfering RNA (SiRNA) was used to knock down U66. The effect of U66 on cell proliferation was measured by Cell Counting Kit-8.

RESULTS

We identified 259 snoRNAs that were differentially expressed in GC compared with normal tissues (**Supplementary Table 2**). Primarily, we used a univariate COX proportional regression to select survival-related snoRNAs in the training set. The 30 snoRNAs with the lowest *P*-values were used to develop the risk formula to predict overall survival. The risk formula was as follows: (0.0496)*(expression of U66) +

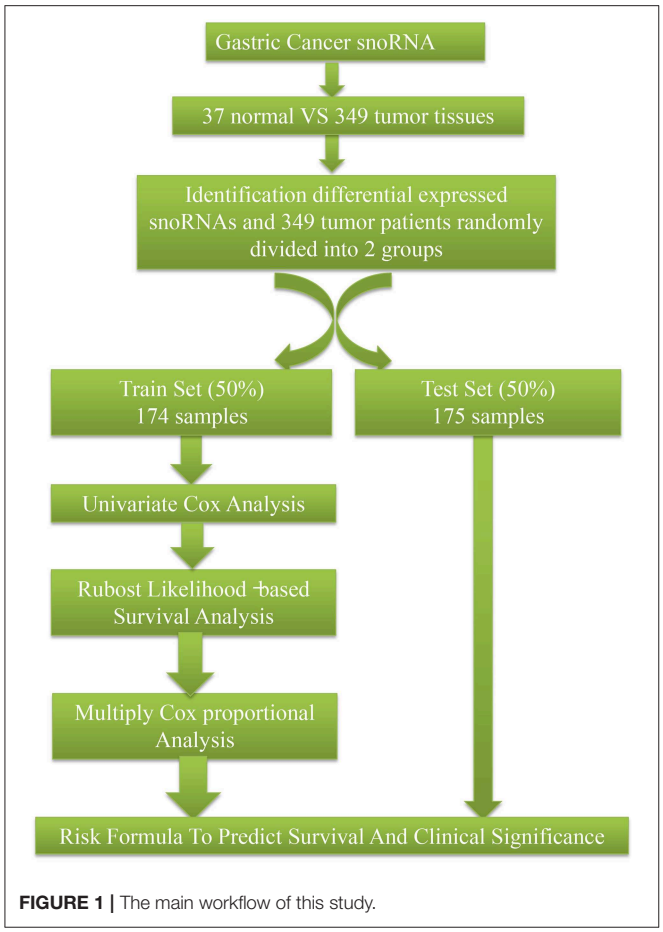


FIGURE 1 | The main workflow of this study.

(−0.0191)*(expression of ACA47) + (0.0363)*(expression of ACA10) + (−0.1711)*(expression of E2) + (0.0650)*(expression of SNORA58) + (0.0953)*(expression of HBII-316) + (−0.4749)*(U70) + (−0.2352)*(expression of U8).

Figures 2A,C show details of the normal and GC tissue groups based on risk score calculated by the risk formula. Survival analysis revealed a significant difference between the two groups (**Figures 2B,D**). The high risk group had significantly shorter overall survival than the low risk group ($p < 0.0001$). The hazard ratio of this risk formula as a prognostic biomarker, was 3.43 (95% confidence interval: 1.93–6.09, $P = 2.72e\text{-}05$). The area under the ROC curve (AUC) of the risk formula was up to 0.828 (**Figure 3A**).

The optimal cutoff was identified as 0.94 with the best Youden's index: 0.64 (sensitivity: 80.1%, specificity: 84.1%). With this cutoff, patients in the test set were divided into two groups (high risk and low risk). Kaplan-Meier curves of the validation data set indicated a significantly prolonged survival time in low-risk compared to high-risk patients (**Figure 3B**; $P < 0.05$). Results from the test set were highly consistent with results from the training set. This suggested that the snoRNA-based signature had good performance in predicting overall survival.

Figure 4A shows the snoRNA expression patterns between normal and tumor tissues. We found eight snoRNAs (ACA47, E2, ACA10, SNORA58, HBII-316, U70, U8, and U66) that were upregulated in tumor compared with normal tissues ($P < 0.05$). Furthermore, there was a correlation between the eight snoRNAs and clinical factors (**Figure 4B**). Seven (ACA47, ACA10, SNORA58, HBII-316, U70, U8, and U66) of the eight snoRNAs were associated with the Lauren classification that divides GC into three types: intestinal, diffuse, and mixed. Seven (ACA47, E2, ACA10, SNORA58, HBII-316, U8, and U66) of eight snoRNAs correlated with the molecular subtype (18). Four (ACA47, HBII-316, U8, and U66) of eight snoRNAs were related with the neoplasm histologic grade. However, none of these eight snoRNAs were statistically correlated with pathologic stage.

Those seven (ACA47, E2, ACA10, SNORA58, HBII-316, U70, and U66) of eight snoRNAs were detected in cell lines. **Figures 5A–G** showed the expression profile of the snoRNAs. Compared with normal tissue, the expression of seven snoRNAs was upregulated in patients (**Figure 5H**). The effect of siRNA of U66 was validated in NCI-N87 (**Figure 5I**). Knocking down U66 inhibited the cell proliferation of NCI-N87 (**Figure 5J**).

TABLE 1 | Clinical covariates for included patients.

Covariate		Total set $n = 349$	Training set $n = 174$	Testing set $n = 175$	P-value[#]
Age, *n*	≥65	202	97	105	$P = 0.421$
	<65	147	77	70	
Gender	Male	134	67	67	$P = 0.966$
	Female	215	107	108	
Pathological stage, *n*	I + II	156	77	79	$P = 0.867$
	III + IV	193	97	96	

[#] x^2-test.

DISCUSSION

Because of advances in high throughput sequencing, numerous snoRNAs have been identified and are emerging as important RNAs, thereby attracting the attention of researchers. Studies have shown that some snoRNAs play important roles in biological processes, and dysfunction of snoRNAs may lead to oncogenesis (19). These studies also indicated that snoRNAs could serve as biomarkers in several diseases, including cancers (20).

In the current study, we used a risk-based formula through multivariate Cox coefficients to identify eight snoRNAs that were differentially expressed between normal and GC tissues (ACA47, E2, ACA10, SNORA58, HBII-316, U70, U8, and U66). The high risk group classified by the risk score had a shorter survival time than the low risk group. These results suggested the eight-snoRNA signature had potential predictive value, and may play a crucial role in the molecular pathogenesis, progression, and prognosis of GC.

The AUC of the ROC was up to 0.828. This indicated that this risk signature had good performance to predict the overall survival of GC patients. Furthermore, Kaplan-Meier survival analysis demonstrated that patients in the high risk group had a shorter overall survival time than those in the low risk

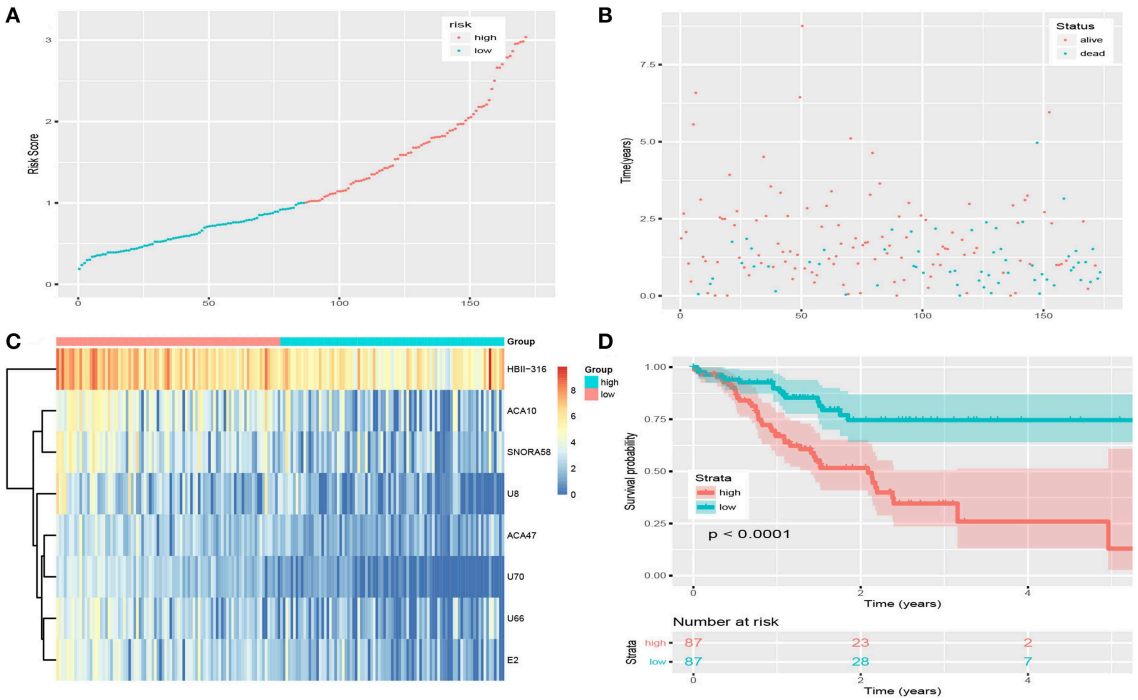


FIGURE 2 | Risk score of snoRNAs-based signature in training set. **(A)** The risk score of patients in train set based on risk signature. **(B)** The distribution of patients' survival status and time. **(C)** The expression profile of eight snoRNAs in train set. **(D)** Survival curve of low-risk group and high-risk group based on median risk score via Kaplan-Meier method.

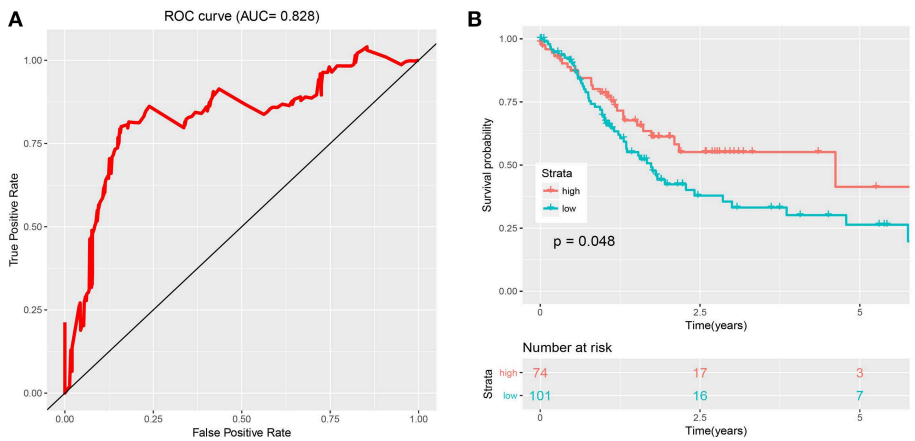


FIGURE 3 | Evaluation and validation of the risk signature. **(A)** The ROC curve of eight-snoRNAs' model. **(B)** The survival curve of two groups (low risk and high risk) based on optimal cutoff in testing set.

group. Thus, the risk biomarker established by the eight-snoRNA signature served as an independent prognostic factor (hazard ratio = 3.43, 95% confidence interval: 1.93–6.09, $P = 2.72e-05$). To our knowledge, this is the first time a risk formula signature was developed using a snoRNA expression profile to predict overall survival of GC patients. These results imply that this risk formula may be used as a novel biomarker.

We also explored the clinical significance of snoRNAs in GC. A clinical features association analysis revealed that seven snoRNAs correlated with the Lauren classification. This classification places GC into three histological subtypes, and has an important influence on prognosis in GC because survival varies depending upon the subtype (21). Seven snoRNAs also correlated with the molecular subtype that classifies GC into four groups: Epstein-Barr virus positive tumors, microsatellite unstable tumors, genomically stable tumors, and tumors with chromosomal instability (17, 18). Therefore, upregulated snoRNAs may be involved in important biological processes such

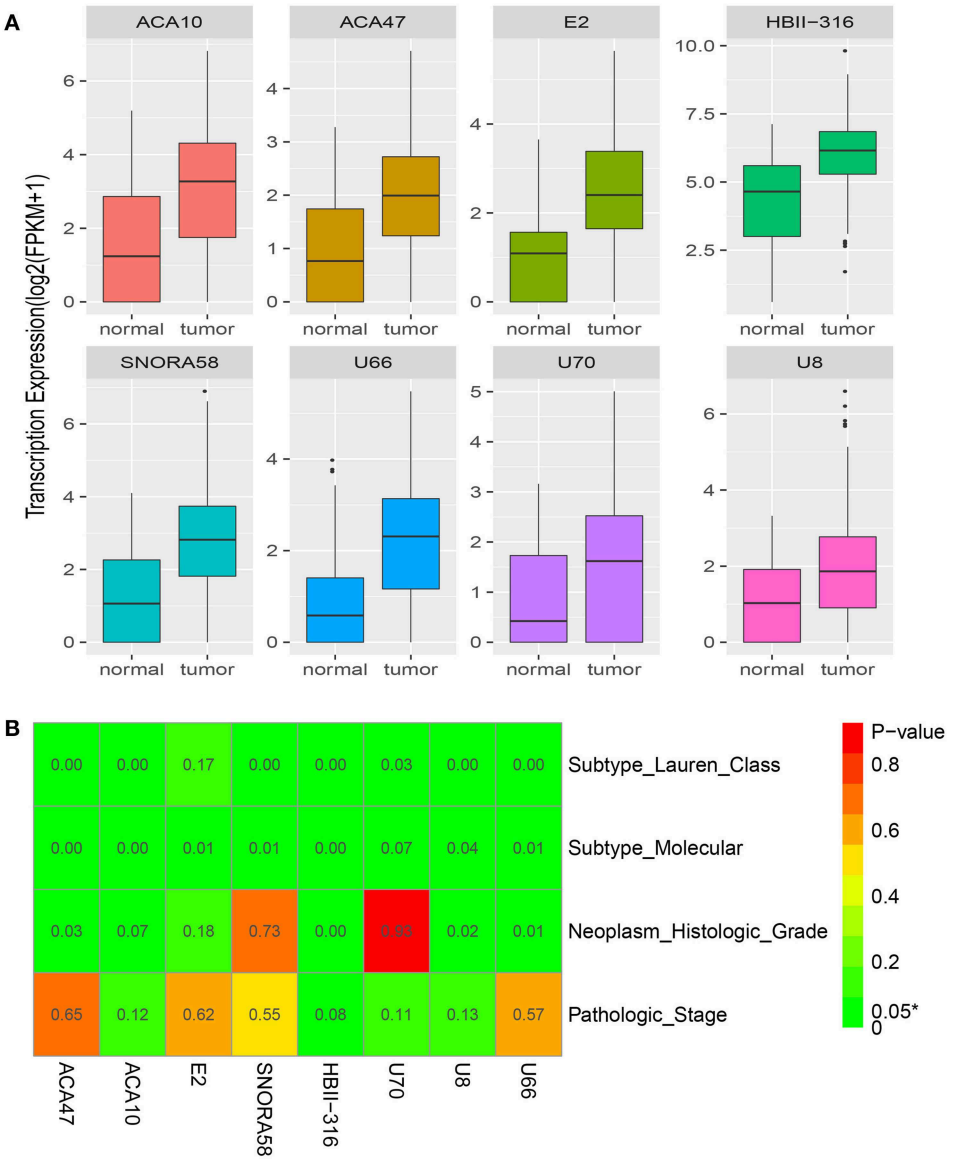


FIGURE 4 | Clinical significance of the eight snoRNAs. **(A)** The expression profile of eight snoRNAs between normal and tumor tissues. **(B)** The correlation between clinical features and the eight snoRNAs.

as microsatellite instability, genomic stability, and chromosomal instability. Although none of the eight snoRNAs correlated statistically with pathologic stage, they may still play important roles in GC biological processes.

This work provides some new clues with clinical implications for the development of novel prognostic factors in GC. Although these eight prognostic snoRNAs have not been investigated previously in cancers, the results indicate that they may be involved in tumorigenesis. We validated seven of eight snoRNAs expression profile both in cell lines and patients' tissue. We validated the function of one snoRNA, U66, which may promote cell proliferation.

A limitation of this study was the analysis of only a single data set because other snoRNA datasets are lacking. Thus, further experiments and more samples are needed to validate these findings.

CONCLUSIONS

In conclusion, 259 differentially expressed snoRNAs were identified and used to develop an eight-snoRNA signature from prognosis-related snoRNAs to predict the overall survival of GC with an AUC up to 0.828. We also explored the potential clinical significance of the eight snoRNAs and found that most were correlated with clinical factors. Overall these results provide further insight into the role of snoRNAs in GC. Further experiment indicated that U66 may promote cell proliferation. Importantly,

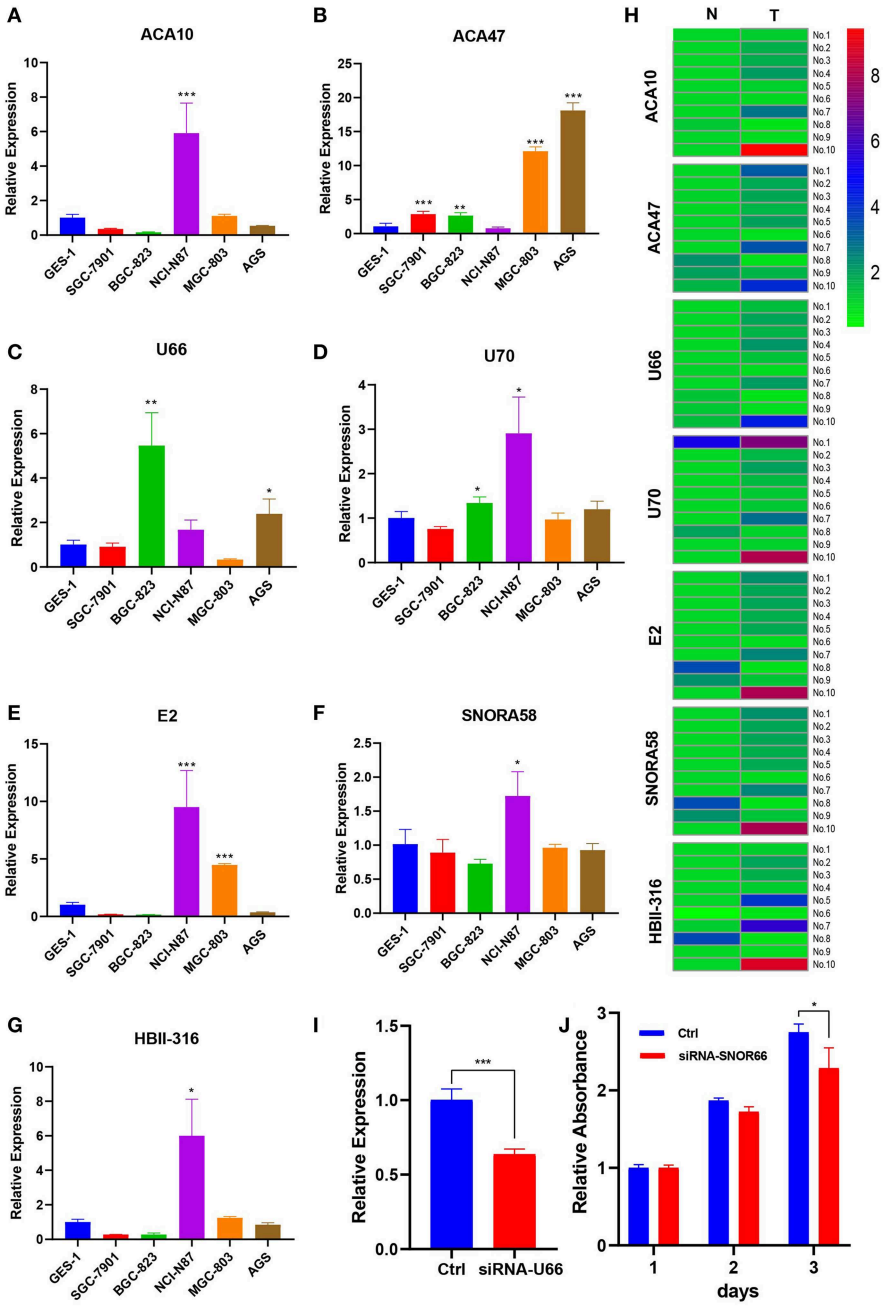


FIGURE 5 | Expression profile of snoRNAs in cell lines and paired tissues. **(A–G)** Expression profile of snoRNAs upregulated in certain cell lines. **(H)** SnoRNAs were mainly upregulated in five patient samples. **(I)** U66 was knocking down by siRNA. **(J)** Knocking down U66 inhibited cell proliferation. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

they may have potential prognostic and therapeutic implications for GC, and serve as predictive biomarkers of overall survival.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://bioinfo.life.hust.edu.cn.

AUTHOR CONTRIBUTIONS

XW, MX, YY, YK, and PL performed the research study and collected the data. XW and MX analyzed the data. BJ and HL designed the research study. XW, MX, and BJ wrote the paper. WZ and PL prepared all the tables. All authors reviewed the manuscript and contributed significantly to this work. In addition, all authors have read and approved the manuscript.

ACKNOWLEDGMENTS

This work was supported by the major scientific instruments and equipment of the state (2013YQ03065110). We thank all people who give us support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.00788/full#supplementary-material

REFERENCES

1. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China (2015). *CA Cancer J Clin*. (2016) 66:115–32. doi: 10.3322/caac.21338
2. Maeda M, Nakajima T, Oda I, Shimazu T, Yamamichi N, Maekita T, et al. High impact of methylation accumulation on metachronous gastric cancer: 5-year follow-up of a multicentre prospective cohort study. *Gut*. (2017) 66:1721–3. doi: 10.1136/gutjnl-2016-313387
3. Wang K, Yuen S, Xu J, Lee S, Yan H, Shi S, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet*. (2014) 46:573–82. doi: 10.1038/ng.2983
4. He J, Zhuo ZJ, Zhang A, Zhu J, Hua RX, Xue WQ, et al. Genetic variants in the nucleotide excision repair pathway genes and gastric cancer susceptibility in a southern Chinese population. *Cancer Manag Res*. (2018) 10:765–74. doi: 10.2147/CMAR.S160080
5. Yu GP, Hsieh CC. Risk factors for stomach cancer: a population-based case-control study in Shanghai. *Cancer Causes Control*. (1991) 2:169–74. doi: 10.1007/BF00056210
6. Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. (2015) 385:977–1010. doi: 10.1016/S0140-6736(14)62038-9
7. Krell J, Frampton AE, Mirnezami R, Harding V, De Giorgio A, Roca Alonso L, et al. Growth arrest-specific transcript 5 associated snoRNA levels are related to p53 expression and DNA damage in colorectal cancer. *PLoS ONE*. (2014) 9:e98561. doi: 10.1371/journal.pone.0098561
8. Martens-Uzunova ES, Hoogstrate Y, Kalsbeek A, Pigmans B, Vredenbregt-van den Berg M, Dits N, et al. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget*. (2015) 6:17430–44. doi: 10.18632/oncotarget.4172
9. Su H, Xu T, Ganapathy S, Shadfan M, Long M, Huang T, et al. Elevated snoRNA biogenesis is essential in breast cancer. *Oncogene*. (2014) 33:1348–58. doi: 10.1038/onc.2013.89
10. Gong J, Li Y, Liu CJ, Xiang Y, Li C, Ye Y, et al. A pan-cancer analysis of the expression and clinical relevance of small nucleolar RNAs in human cancer. *Cell Rep*. (2017) 21:1968–181. doi: 10.1016/j.celrep.2017.10.070
11. Cho HJ, Yu A, Kim S, Kang J. Robust likelihood-based survival modeling with microarray data. *J Stat Softw*. (2008) 29:1–16. doi: 10.18637/jss.v029.i01
12. Therneau TM, Grambsch PM. Modeling survival data: extending the cox model. *Technometrics*. (2000) 44:85–6. doi: 10.1007/978-1-4757-3294-8
13. Therneau T. *A Package for Survival Analysis in S. version 2.38* (2015). Available online at: https://CRAN.R-project.org/package=survival
14. Heagerty PJ. *survivalROC: Time-Dependent ROC Curve Estimation from Censored Survival Data* (2013). Available online at: https://cran.r-project.org/web/packages/survivalROC/index.html
15. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company, Incorporated (2009).
16. Tang Y, Horikoshi M, Li W. *ggfortify: unified interface to visualize statistical results of popular R packages. R J*. (2016) 8:478–89. doi: 10.32614/RJ-2016-060
17. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*. (2014) 513:202–9. doi: 10.1038/nature13480
18. Zhang W. TCGA divides gastric cancer into four molecular subtypes: implications for individualized therapeutics. *Chin J Cancer*. (2014) 33:469–70. doi: 10.5732/cjc.014.10117
19. Williams GT, Farzaneh F. Are snoRNAs and snoRNA host genes new players in cancer? *Nat Rev Cancer*. (2012) 12:84–8. doi: 10.1038/nrc3195
20. Thorenoor N, Slaby O. Small nucleolar RNAs functioning and potential roles in cancer. *Tumour Biol*. (2015) 36:41–53. doi: 10.1007/s13277-014-2818-8
21. Chen YC, Fang WL, Wang RF, Liu CA, Yang MH, Lo SS, et al. Clinicopathological variation of lauren classification in gastric cancer. *Pathol Oncol Res*. (2016) 22:197–202. doi: 10.1007/s12253-015-9996-6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 September 2019
doi: 10.3389/fgene.2019.00786

Specific Glioma Prognostic Subtype Distinctions Based on DNA Methylation Patterns

Xueran Chen[1,2]*, *Chenggang Zhao*[1,3], *Zhiyang Zhao*[1,3], *Hongzhi Wang*[1,2] *and Zhiyou Fang*[1,2]*

[1] Anhui Province Key Laboratory of Medical Physics and Technology; Center of Medical Physics and Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China, [2] Hefei Cancer Hospital, Chinese Academy of Sciences, Hefei, China, [3] University of Science and Technology of China, Hefei, China

DNA methylation is an important regulator of gene expression and may provide an important basis for effective glioma diagnosis and therapy. Here, we explored specific prognosis subtypes based on DNA methylation status using 653 gliomas from The Cancer Genome Atlas (TCGA) database. Five subgroups were distinguished by consensus clustering using 11,637 cytosines preceding a guanosine (CpGs) that significantly influenced survival. The specific DNA methylation patterns were correlated with age, tumor stage, and prognosis. Additionally, weighted gene co-expression network analysis (WGCNA) analysis of CpG sites revealed that 11 of them could distinguish the samples into high- and low-methylation groups and could classify the prognostic information of samples after cluster analysis of the training set samples using the hierarchical clustering algorithm. Similar results were obtained from the test set and 12 glioma patients. Moreover, *in vitro* experiments revealed an inverse relationship between methylation level and migration ability or insensitivity to temozolomide (or radiotherapy) of glioma cells based on the final prognostic predictor. Thus, these results suggested that the model constructed in this study could provide guidance for clinicians regarding the prognosis of various epigenetic subtypes.

Keywords: glioma, consensus clustering, DNA methylation, molecular subtypes, prognosis

OPEN ACCESS

Edited by:
Barbara Karen Dunn, National Institutes of Health (NIH), United States

Reviewed by:
Qi Zhao, Sun Yat-sen University Cancer Center (SYSUCC), China
Tania Lee Slatter, University of Otago, New Zealand

***Correspondence:**
Xueran Chen
xueranchen@cmpt.ac.cn
Zhiyou Fang
z.fang@cmpt.ac.cn

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *20 February 2019*
Accepted: *24 July 2019*
Published: *12 September 2019*

Citation:
Chen X, Zhao C, Zhao Z, Wang H and Fang Z (2019) Specific Glioma Prognostic Subtype Distinctions Based on DNA Methylation Patterns. Front. Genet. 10:786. doi: 10.3389/fgene.2019.00786

INTRODUCTION

Glioma derives from glial cells and is the most prevalent primary central nervous system malignant tumor (Aldape et al., 2003; Aquilanti et al., 2018). The overall survival time continues to be unsatisfactory, especially for high-grade glioma, although treatment strategies, including surgical resection, radiation, and chemotherapy, for glioma patients have been greatly improved (Jain, 2018; Zang et al., 2018). It is therefore urgent to elucidate the molecular mechanisms underlying glioma tumorigenesis for developing novel therapies.

Epigenetics is recognized as heritable alterations in gene expression not connected to an alteration in DNA sequence but plays a crucial role in carcinogenesis (El-Osta, 2004; Issa, 2007; Hao et al., 2017). Cancer epigenetics covers aspects of aberrant DNA methylation, dysregulated

Abbreviations: CDF, consensus cumulative distribution function; CpG, cytosine preceding a guanosine; GBM, glioblastoma multiforme; knn, k-nearest neighbors; LGG, lower-grade glioma; SD, standard deviation; TCGA, The Cancer Genome Atlas; WGCNA, weighted gene co-expression network analysis.

non-coding RNA, and altered post-translational histone modification, among which aberrant DNA methylation is most widely investigated (Dawson and Kouzarides, 2012; Kanwal et al., 2015). Aberrant DNA methylation could influence the key genes that are involved in glioma carcinogenesis and progression and may especially influence some tumor suppressor genes by altering their expression and inhibiting their function (Liu et al., 2016; Charlet et al., 2017). Thus, biological processes, specifically alterations in DNA methylation, can provide an important basis for early diagnosis and prognosis of cancer and development of new approaches for further clinical applications. Although the effects of certain genes with aberrant DNA methylation on glioma have been reported extensively, the comprehensive profile of the interaction network still needs further elucidation.

During the last decades, bioinformatics analysis and microarray technology have been widely used to identify general genetic or epigenetic alterations in carcinogenesis and screen biomarkers for prognosis and diagnosis of cancer (Crispatzu et al., 2017; Yang et al., 2019). Several single genes whose global methylation status correlates with glioma outcome and gene expression level have already been identified (Fanelli et al., 2008; Hill et al., 2014). Additionally, some research on aberrant DNA methylation has been conducted to identify glioma DNA methylation subtypes by DNA methylation profile (Gustafsson et al., 2018; Johannessen et al., 2018); however, this classification was not detailed enough, and the specific sites that are associated with each category are unclear.

In this study, we addressed glioma classification by identifying specific prognosis subtypes based on DNA methylation profiles of glioma obtained from The Cancer Genome Atlas (TCGA) database. This classification system may help identify molecular subtypes or new glioma markers to subdivide glioma patients more accurately. Moreover, our classification system provides guidance for clinicians on personalized treatments and diagnoses by identifying differences in prognosis for each epigenetic subtype.

MATERIALS AND METHODS

Data Pre-processing and the Initial Screening of DNA Methylation Loci in Glioma

Lower-grade glioma (LGG) and glioblastoma multiforme (GBM) DNA methylation data generated with the Illumina Infinium HumanMethylation450 BeadChip array were downloaded from the TCGA data portal (Weinstein et al., 2013). Methylation level of each probe was represented by the β-value, which ranges from 0 to 1, corresponding to unmethylated and fully methylated, respectively. Probes with missing data in more than 70% of the samples were removed. The remaining probes that were not available (NAs) were imputed using the k-nearest neighbors (knn) imputation procedure. The ComBat algorithm in sva *R* package was used to remove batch effects by incorporating patient ID information and batch and integrating all the DNA methylation array data. Unstable genomic sites, including cytosines preceding a guanosine (CpGs) in single nucleotide polymorphisms and sex chromosomes, were removed. We selected CpGs in promoter regions because DNA methylation in promoter regions influences gene expression strongly. Promoter regions were defined as 2 kb upstream to 0.5 kb downstream from transcription start sites. Finally, we selected samples having gene expression profiles. In total, 653 gliomas were used for the analysis.

Next, we separated the data set into two cohorts: a training set and a test set. The criteria for this grouping were as follows: a) random division of samples into two groups and b) similar age distribution, staging, follow-up time, and death ratio in the two groups.

Determining Classification Features by COX Proportional Risk Regression Models

CpG sites influencing survival significantly were used as classification features. First, univariate COX proportional risk regression models were constructed with methylation levels of each CpG site, age, and stage, and survival data of the cases. Then, the significant CpGs obtained from univariate COX proportional risk regression models were introduced into multivariate COX proportional risk regression models, using tumor stage and age as covariates, which were also significant in the univariate models. Finally, the CpG sites that were still significant were used as classification features. COX proportional hazard models were fitted with methylation levels of CpGs using the coxph function in survival package *R*, with clinical and demographic attributes (stage and age) as covariates in the multivariate analysis.

Consensus Clustering to Obtain Molecular Subtypes Associated With Glioma Prognosis

Consensus clustering was performed with the ConsensusClusterPlus package in *R* to determine subgroups of gliomas based on the most variable CpG sites (Wilkerson and Hayes, 2010). In this study, 80% of the samples were sampled 100 times by adopting the resampling program; the similarity distance between samples was estimated by the Euclidean distance (Ghosh and Barman, 2016), and kmdist was used as the clustering algorithm to search for the reliable and stable subgroup classification. After executing ConsensusClusterPlus, the item-consensus results and cluster consensus were obtained. The criteria to determine the number of clusters were as follows: relatively high consistency within clusters, relatively low variation coefficient, and no appreciable rise in the area under the cumulative distribution function (CDF) curve. Variation coefficient was calculated according to the following formula: coefficient of Variation (CV) = (SD/MN)*100%, where MN represents the average of samples and SD represents the standard deviation. The category number was selected as the area under the CDF curve and showed no significant change. The heat map corresponding to the consensus clustering was generated by pheatmap *R* package.

Survival and Clinical Characteristic Analyses

Kaplan–Meier plots were used to determine overall survival among glioma subgroups defined by DNA methylation profiles. The log-rank test was used to measure the significant differences among the clusters. Survival analyses were performed with the survival package in *R* software. Associations between biological and clinical characteristics and DNA methylation clustering were analyzed with the chi-square test. All tests were two-sided, and

for all statistical tests, $p < 0.05$ was considered to be significant unless otherwise noted.

Glioma Cell Survival and Migration Assays

After receiving informed consent, glioma specimens were obtained from patients undergoing surgery at the Hefei Cancer Hospital, Chinese Academy of Sciences, in accordance with the Institutional Review Board. Within hours after surgical removal, tumor specimens were enzymatically dissociated into single cells, following previously reported procedures (Chen et al., 2016). For cell survival assay, the cells were plated at a seeding density of 10,000 cells/plate in a 60 mm plate, treated with or without temozolomide or 6 Gy radiotherapy, grown for 48 h in a standard growth medium, and washed with phosphate buffer saline (PBS). For cell migration assay, cell suspension in serum-free medium was added to the upper Transwell chamber and then incubated for 18 h. The cells were fixed in cold methanol for 20 min, washed, and stored. Fixed cell colonies were visualized by incubating the cells with 0.5% (w/v) crystal violet for 0.5 h. Excess crystal violet was removed by washing with PBS. Cells that survived or migrated were counted. Differences in means were considered statistically significant when $p < 0.05$ using a two-tailed t test.

RESULTS

DNA Methylation Features for Classification Based on Prognosis

To identify the specific CpG sites that were significantly correlated with survival in glioma, we set up the workflow shown in **Figure 1**. The 450 k methylation profiles were downloaded from *TCGA*; 485,577 CpG sites in 685 samples and clinical follow-up information from 1,148 cases were obtained. There were 653 matched samples between clinical data and methylation profiles. The samples were evenly divided into a training set (n = 327) and test set (n = 326); four properties (including age, follow-up period, proportion of death cases, and clinical stage) between the training set and test set samples were observed, and they were found to be similar in the training set and test set (**Supplementary Figure 1**). Firstly, the univariate COX proportional hazard regression model was used to analyze each methylation site and survival data. When $p < 0.05$ was selected as the threshold, a total of 12,264 methylation sites significantly correlated with survival were obtained. Age ($p = 0.0043$) and tumor stage ($p = 0.0012$) were also significant factors. Age and grade were included in the COX proportional hazard regression model as covariates, and 13,739 methylation sites significantly correlated with survival were obtained, including 11,637 matching sites between the two analyses.

Consensus Clustering of Glioma Identified Distinct DNA Methylation Prognosis Subgroups

The methylation profiles of the 11,637 CpG sites from the 327 samples in the training set were employed for the consensus clustering of samples using the ConsensusClusterPlus *R* software package to obtain the glioma molecular subtypes. To determine the appropriate cluster number, we calculated the average cluster consistency and inter-cluster variation coefficient for the number of each cluster, respectively. Typically, the area under the CDF curve tended to be stable after five clusters (**Figure 2A**), the

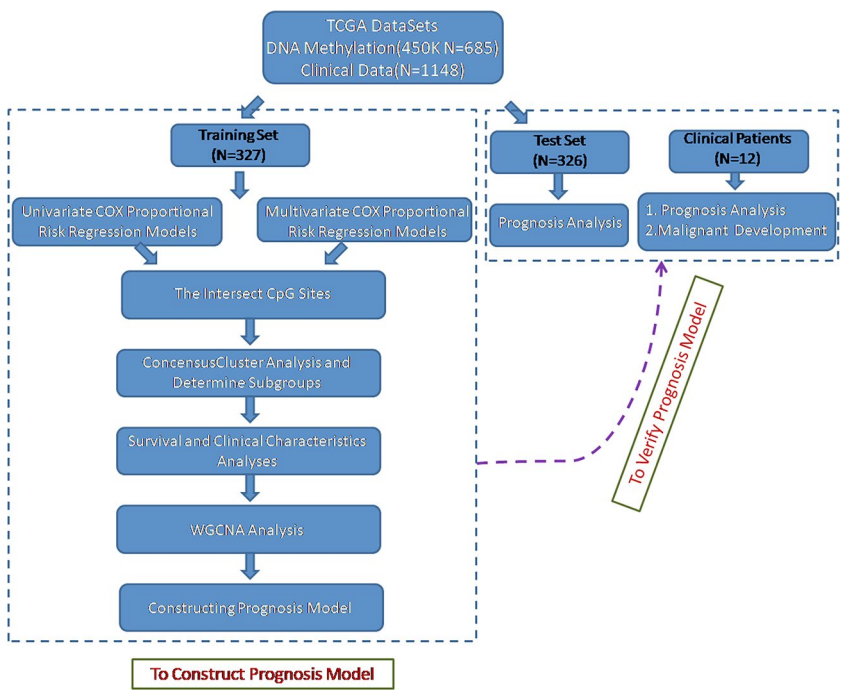


FIGURE 1 | Flowchart describing the schematic overview of the study design.

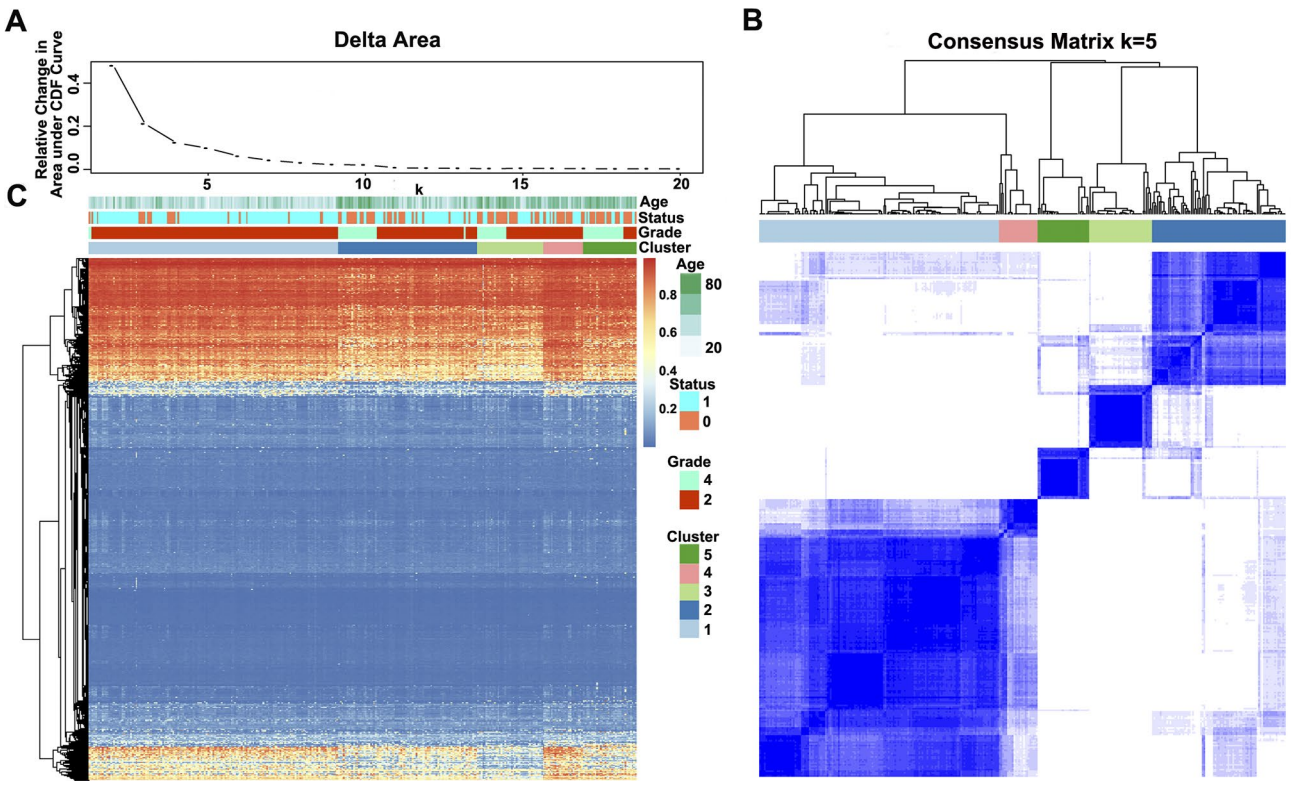


FIGURE 2 | Consensus matrix for DNA methylation classification with the corresponding heat map. **(A)** Delta area curve of consensus clustering, indicating the relative change in area under the cumulative distribution function (CDF) curve for each category number ***k*** compared with ***k*** – 1. The horizontal axis represents the category number ***k***, and the vertical axis represents the relative change in area under the CDF curve. **(B)** Color-coded heat map corresponding to the consensus matrix for ***k*** = 5 obtained by applying consensus clustering. The color gradients were from 0 to 1, representing the degree of consensus, with white corresponding to 0 and dark blue to 1. **(C)** Heat map corresponding to the dendrogram in (B), which was generated using the pheatmap function with DNA methylation classification, tumor stage, age, and prognostic status as the annotations.

smallest variation coefficient among all clusters was 0.076, and the sample cluster number was 5 (**Supplementary Table 1**). Therefore, five was selected as a suitable cluster number for further analysis in this study (**Figure 2B**).

Notably, most methylation sites displayed low DNA methylation levels in each sample; additionally, there were also differences in the DNA methylation profile among the five clusters, and the DNA methylation levels of Cluster2, Cluster3, and Cluster5 were lower than those of Cluster1 and Cluster4 (**Figure 2C**).

Indeed, the methylation levels of these five subgroups were significantly related to some molecular genetic features. For example, the methylation levels were positively associated with TP53 mutant but were negatively associated with co-deletion of 1p/19q in Cluster1 (**Supplementary Table 2**). In Cluster2, tumor protein p53 (TP53) mutant, isocitrate dehydrogenase [NADP(+)] 1 (IDH1) mutant, and co-deletion of 1p/19q have been reported to be negatively associated with methylation levels (**Supplementary Table 3**). The methylation levels were positively related to O-6-methylguanine-DNA methyltransferase (MGMT) promoter unmethylation but were negatively associated with TP53 mutant, α-thalassemia mental retardation X-linked (ATRX) mutant, and co-deletion of 1p/19q in Cluster3 (**Supplementary Table 4**). In Cluster4, the methylation levels have been associated with IDH1 mutant, ATRX mutant, and *MGMT* promoter unmethylation (**Supplementary Table 5**). TP53 mutant, telomerase reverse transcriptase (TERT) mutant, and *MGMT* promoter unmethylation were associated with methylation levels in Cluster5 (**Supplementary Table 6**). Thus, the five subgroups based on the methylation levels may reflect changes in some molecular genetic features.

Characterizing Different Characteristics of DNA Methylation Clustering

Furthermore, we analyzed the prognosis, grade and age distribution, and survival of each sample in the five molecular subtypes. It was discovered through Kaplan–Meier and log-rank tests that there were significant differences in prognosis among samples of these five molecular subtypes (p = 0.00039) (**Figure 3A**); Cluster4 had favorable prognosis, while Cluster2 and Cluster3 were associated with poor prognosis and relatively lower DNA methylation levels, revealing that the prognosis for low-methylated samples was poorer than that for highly methylated samples. It was also noted that patients in Cluster1 were generally between 30 and 45 years of age (**Figure 3B**) and were younger than patients in the other clusters. Comparing the tumor grades

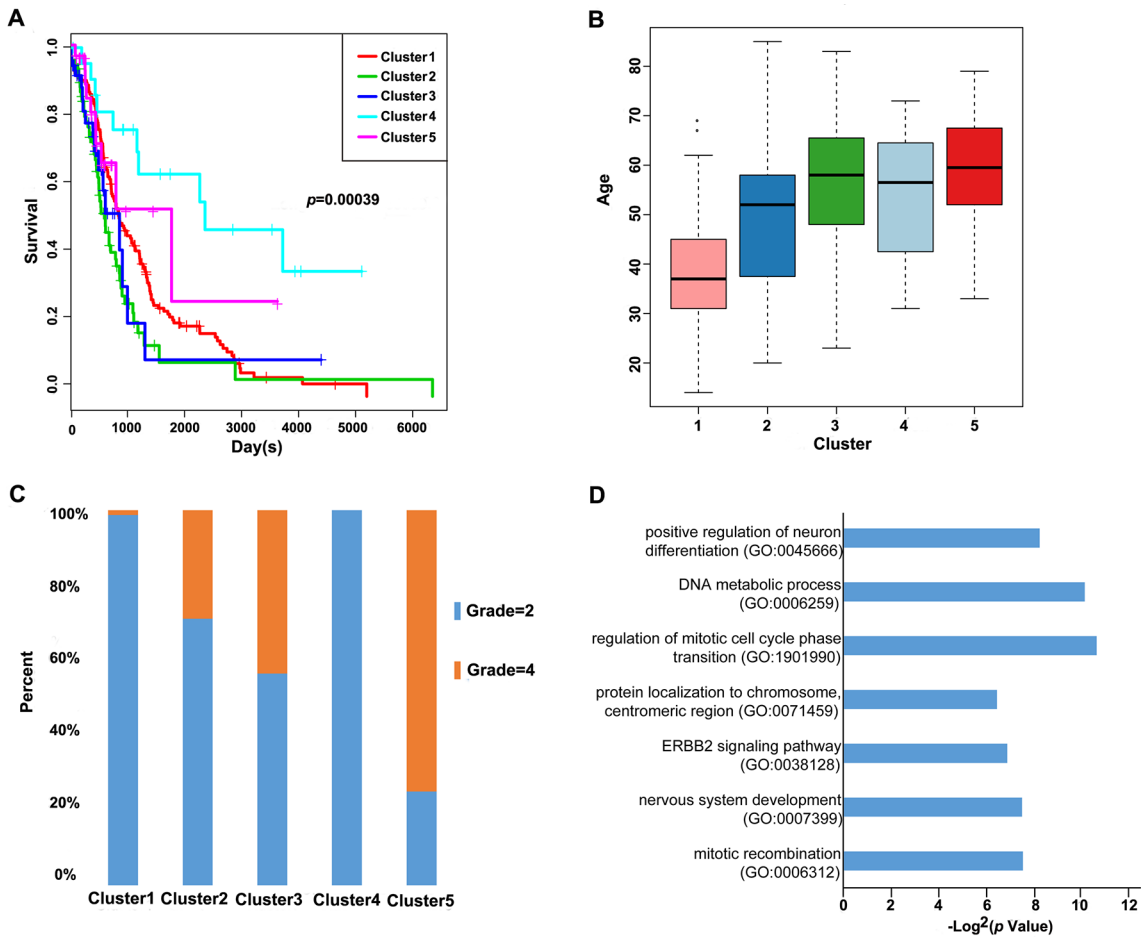


FIGURE 3 | Prognosis, grade, age distribution, and survival of each sample in the molecular subtypes. **(A)** Survival curves of DNA methylation subtypes in the training set. The horizontal axis represents the survival time (days), and the vertical axis represents the probability of survival. The numbers in parentheses in the legend represent the number of samples in each cluster. The log-rank test was used to assess the statistical significance of the differences. **(B)** Age distributions of nine DNA methylation clusters in the training set. The horizontal axis represents the DNA methylation clustering. **(C)** Grade distributions of nine DNA methylation clusters in the training set. The horizontal axis represents the DNA methylation clustering. **(D)** The online network tool Enrichr was utilized for functional enrichment analysis of genes corresponding to the gene promoter regions annotated by the CpG sites that were significantly correlated with survival.

of the subgroups, 98.7% and 100% of the samples in Cluster1 and Cluster4 corresponded to glioma grade 2, respectively, while 71.1%, 56.4%, and 25% of the samples in Cluster2, Cluster3, and Cluster5 corresponded to grade 2, respectively (**Figure 3C**). Taken together, these results indicated that these DNA methylation sites could serve as important markers for prognosis.

Next, the online network tool Enrichr was utilized for functional enrichment analysis of genes corresponding to the gene promoter regions annotated by the CpG sites that were significantly correlated with survival (Chen et al., 2013). It was found that these genes were enriched in the biological processes related to glioma, which included basic cancer-related biological processes, as well as glioma-related specific biological processes, including mitotic recombination, DNA metabolism, and ErbB2 signaling pathway (**Figure 3D**), suggesting that the methylation sites revealed in this study might affect gliomagenesis and development. The weight co-expression network was constructed using the weighted gene co-expression network analysis (WGCNA) *R* software package (Langfelder and Horvath, 2008), and to guarantee that the network was scale-free, the soft threshold â = 6 was selected (**Figure 4A**). Five modules were obtained after further analysis (**Figure 4B**), among which the gene numbers included in each module were 80, 67, 52, 637, 1,319, and 59, respectively (**Supplementary Table 7**). Analysis of the module–trait relationship showed that several of the modules displayed significant correlation or anti-correlation with the five glioma molecular subtypes (**Figure 4C**).

Identifying Specific DNA Methylation Markers

Cluster4 was linked to the best prognosis among all clusters; therefore, all CpG sites in the turquoise module that was most correlated with Cluster4 were selected. The CpG sites (connectivity > 1000) in the network were selected as the feature

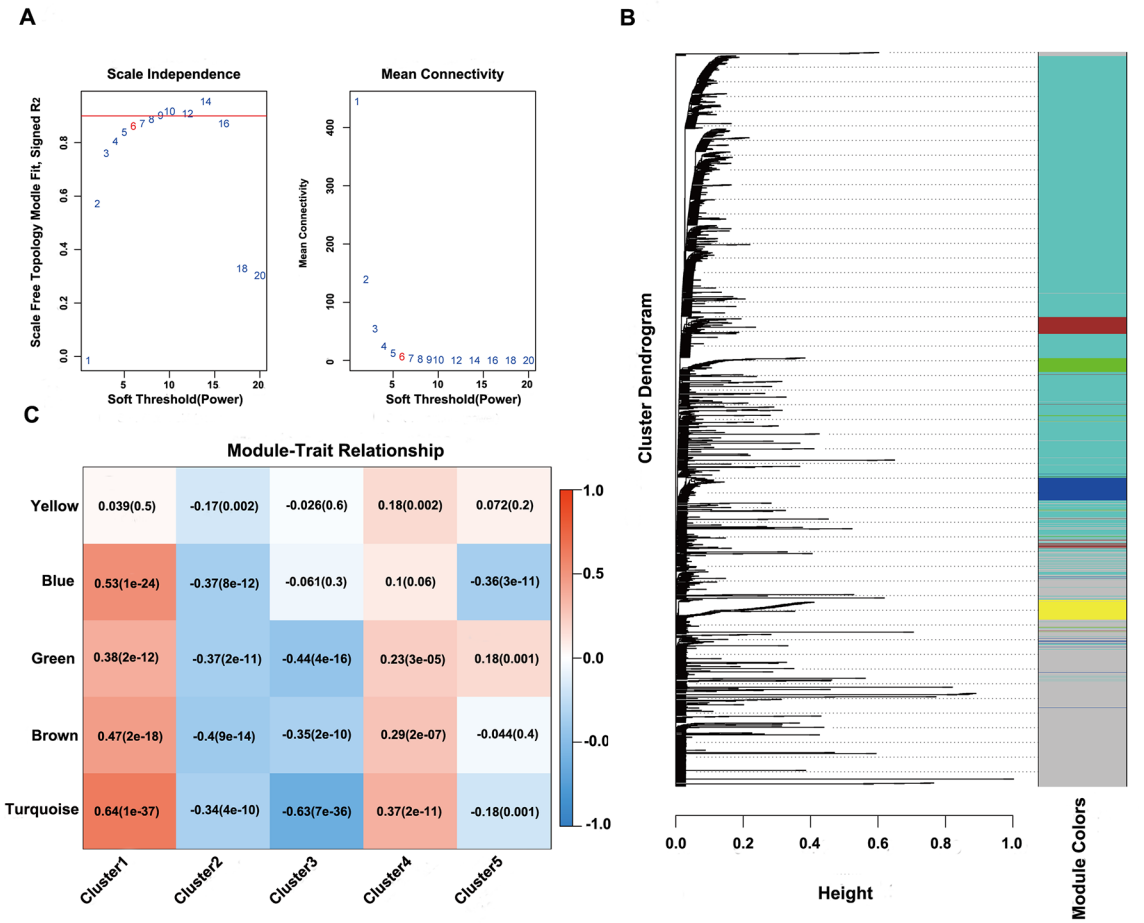


FIGURE 4 | WGCNA analysis of CpG sites. **(A)** Scale-free topology index and mean connectivity were used to determine the soft threshold (â = 6). **(B)**, Clustering dendrogram of CpG sites. The dissimilarity of CpG sites is based on topological overlap. The genes are assigned to different modules and are identified using different colors. **(C)** Module–trait correlation analysis showed that five modules were significantly correlated with each cluster.

methylation sites of Cluster4 samples, and the correlation among 108 CpG loci was significantly higher than that among other loci using Pearson correlation analysis. Ultimately, we chose 11 CpG loci, which intersected the 2 loci (**Supplementary Figure 2** and **Supplementary Table 8**).

Constructing and Evaluating the Prognosis Prediction Model

These 11 CpG methylation profiles were selected for further unsupervised cluster analysis; the similarity between samples was calculated by the Euclidean distance. The results suggested that the methylation levels of these 11 CpG sites could divide the samples into two groups, namely, Cluster1 and Cluster2, of which Cluster2 was the high-methylation group, while Cluster1 was the low-methylation group (**Figure 5A**). The difference in prognosis between the two groups was further analyzed, which revealed that the prognosis in the high-methylation group was worse than that in the low-methylation group (**Figure 5B**). The methylation profiles of these 11 CpG sites were extracted from the methylation profiles in the test set for further hierarchical cluster analysis. It was observed that the methylation profiles of these 11 CpG methylation sites could be clearly grouped into two clusters, among which the methylation level in Cluster1 samples was markedly lower than that in Cluster2 samples (**Figure 5C**). The distinct high-methylation and low-methylation samples were selected for survival analysis and demonstrated that the prognosis in highly methylated samples was notably worse than that in low-methylated samples (**Figure 5D**), which was consistent with the training set results.

Based on the final prognostic predictor, we analyzed the clinical follow-up data of these 12 glioma patients, which were divided into the high-methylation group (n = 6) and low group (n = 6) (**Supplementary Figure 3** and **Figure 6A**). There was a positive correlation between the methylation level and overall survival (p = 0.0162) (**Figure 6B**), with an area under curve (AUC) of 0.8542 (**Figure 6C**). Consistent with these, there was an inverse relationship between the methylation level and insensitivity to temozolomide (or radiotherapy) (**Figures 6D**, **E**) or migration ability (**Figure 6F**) of glioma cells derived from GBM patients. Thus, we concluded that this prognostic predictor showed great promise for application in clinical practice.

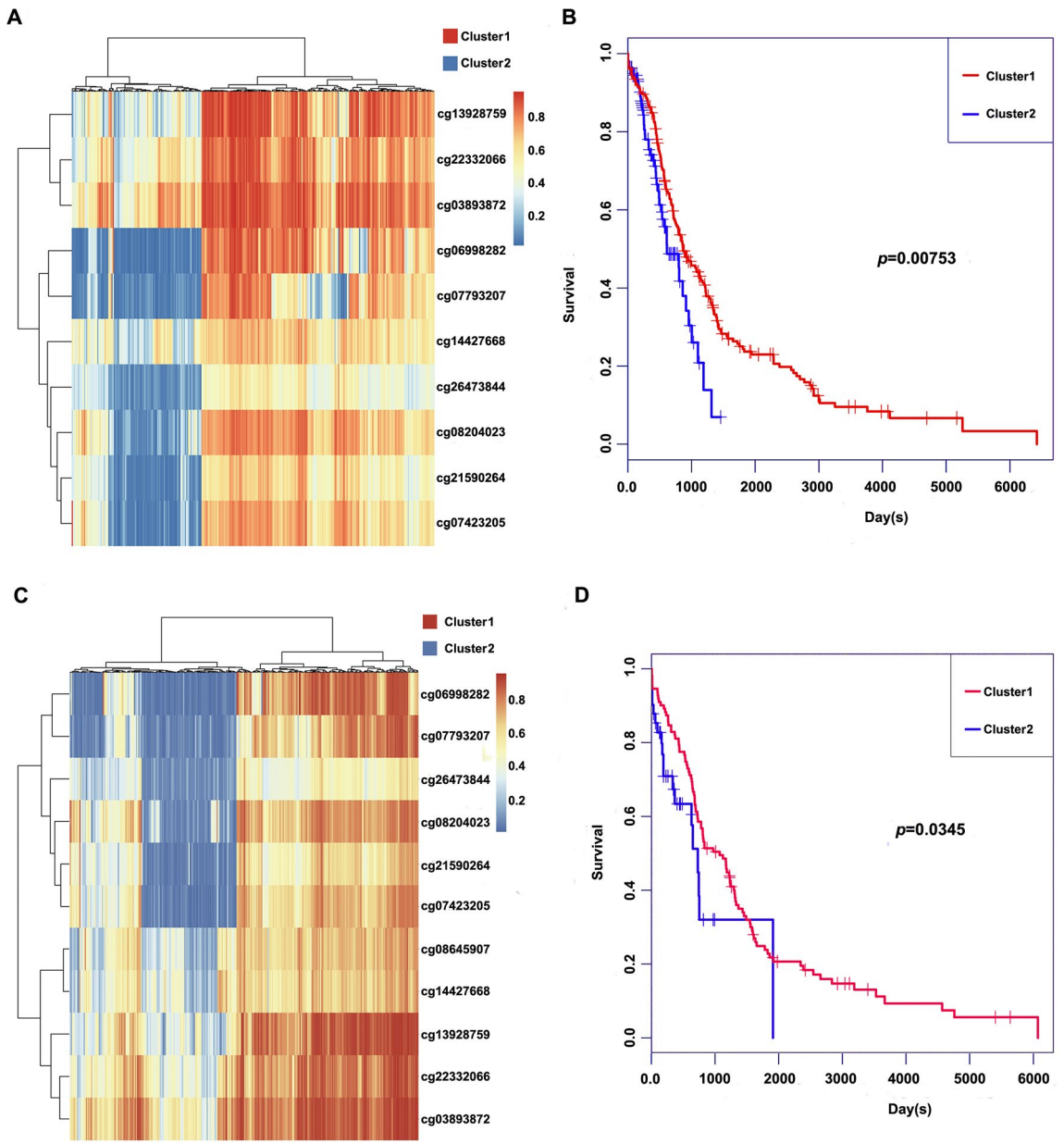


FIGURE 5 | Clustering and survival results of the 11 CpG sites in the training and test set. **(A)** Consensus clustering of the 11 CpG sites in the training set. **(B)** Survival curves of two clusters predicted from the training set using the prognosis model. The log-rank test was used to assess the statistical significance of the difference. **(C)** Consensus clustering of the 11 CpG sites in the test set. **(D)** Survival curves of two clusters predicted from the test set using the prognosis model. The log-rank test was used to assess the statistical significance of the difference.

DISCUSSION

Aberrant DNA methylation is one of the hallmarks of cancer tissues (Klutstein et al., 2016; Witt et al., 2018). Recent developments in sequencing technologies have made it possible to analyze genome-wide DNA methylation profiles at high resolution. Whole genome bisulfate sequencing is the best method to investigate DNA methylation; its efficacy, however, is limited by high analytic burden and cost. DNA methylation arrays are a good alternative for investigating genome-wide DNA methylation in a large collection of tumors. The TCGA database is a publicly available resource that covers a wide variety of data types in a variety of cancers; thus, the large sample sizes allowed us to explore glioma molecular subtypes more comprehensively.

Global loss of methylation and gene-specific DNA promoter methylation occur frequently during carcinogenesis, and these methylation alterations have been regarded as potential molecular markers for cancer initiation and progression (Dor and Cedar, 2018;

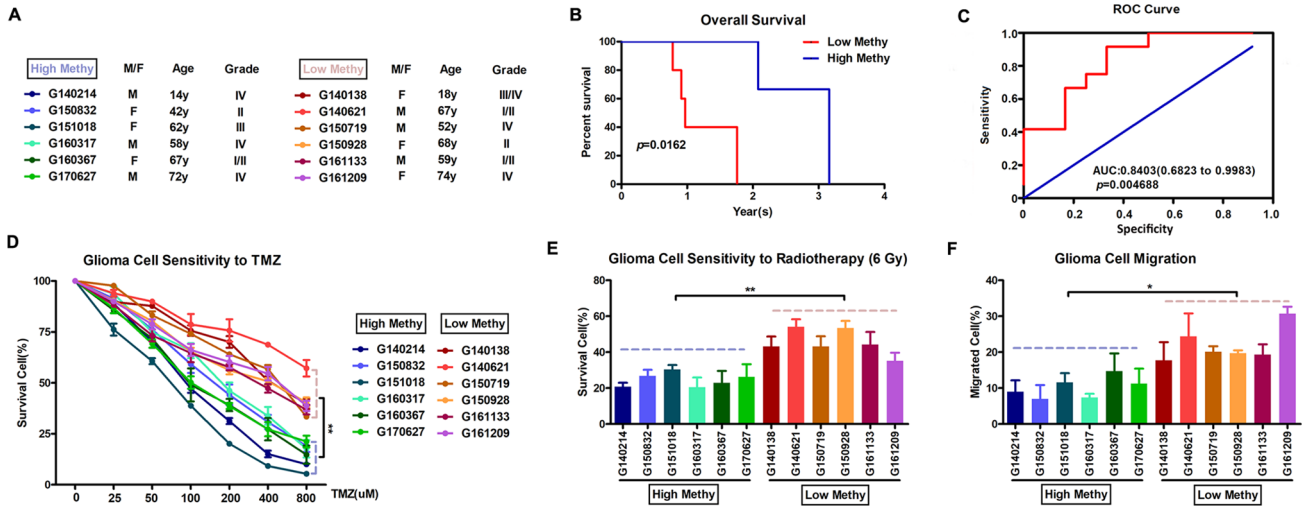


FIGURE 6 | Application in clinical practice of the final prognostic predictor on 11 feature genes. **(A)** The clinical characteristics of the 12 glioma patients. **(B)** Survival curves of two clusters predicted from 12 glioma patients using the prognosis model. The log-rank test was used to assess the statistical significance of the difference. The red line indicates the low-methylation group (high-risk group), while the blue line indicates the high-methylation group (low-risk group), based on the final prognostic predictor. **(C)** receiveroperating characteristic (ROC) curve with AUC under the final prognostic predictor. **(D)** The proportion of surviving glioma cells derived from glioma patients after treatment with temozolomide with indicated concentration. **(E)** The proportion of surviving glioma cells derived from glioma patients after 6 Gy of irradiation. **(F)** The proportion of migrated glioma cells derived from glioma patients.

Koch et al., 2018). DNA methylation in mammals mostly occurs at position 5′ of the cytosine ring in CpGs through a covalent bond of the methyl group (Arber and Linn, 1969; Yarus, 1969). Non-CpG sequences can also get methylated but with less frequency. In normal tissue, CpG island methylation usually increases with age, although the total genomic content of methylcytosine decreases (Perez et al., 2018). During carcinogenesis, a global loss of DNA methylation, together with tumor suppressor gene silencing by promoter DNA methylation, has been observed in most tumor types. Promoter methylation in tumor suppressor gene CpG islands has been demonstrated as a hallmark of cancer. Earlier research has profiled gene-specific promoter methylation in neck squamous cell carcinoma and head, bladder, lung, and liver cancers, among others.

Molecular mechanistic study based on bioinformatics analysis is a significant method in cancer research. Previous studies indicated that glioma could be classified into three groups based on patterns of global DNA methylation: glioma CpG island methylator phenotype (G-CIMP) (highly methylated), intermediately methylated, or low-methylated tumors (Verhaak et al., 2010). One problem associated with the use of clustering algorithms to classify tumors into subgroups is the failure to realize the "true" number of subgroups that are present in a data set. Here, we explored specific prognosis subtypes based on DNA methylation status using 653 gliomas from the TCGA database. To determine the appropriate cluster number, we calculated the average cluster consistency and inter-cluster variation coefficient for the number of each cluster, respectively. Typically, the area under the CDF curve tended to be stable after five clusters, the smallest variation coefficient among all clusters was 0.076, and the sample cluster number was 5. Thus, five subgroups were distinguished by consensus clustering using 11,637 CpGs that significantly influenced survival. Similar to recent studies (Ceccarelli et al., 2016; De Souza et al., 2018), the subgroups based DNA methylation was associated with patient age, advanced stage, and prognosis. Importantly, the methylation levels of different subgroups could reflect different molecular genetic features.

Multifold molecular analyses have been used to take advantage of tumor biology in response to prediction or risk stratification (Krajewska et al., 2017; Masci, 2017). It is known that transcriptional activity is regulated by methylation of cytosine residues, which constitutes a rather stable DNA modification. Reports on DNA methylation signature, which predicts cancer risk, are rare, however. It is important to discover tumor-specific prognostic factors for glioma to predict outcome and improve treatments. Here, WGCNA analysis of the CpG sites revealed that 11 of them could distinguish the samples into high- and low-methylation groups and could classify the prognostic information of samples after cluster analysis of the training set samples using the hierarchical clustering algorithm. It is worth noting that four CpG sites were found in the glial cell line–derived neurotrophic factor (GDNF) gene, a member of the transforming growth factor-â (TGF-â) superfamily, which signals *via* the tyrosine kinase receptor c-Ret and the Glial cell line-derived neurotrophic factor receptor(GDNF)-alpha (GFRá); meanwhile, it is well documented that GDNF also supports neuronal differentiation and dopaminergic development. Limited availability of clinical data and fresh tumor specimens symbolizing transitional steps from tumor initiation to progression is an important barrier to improving the clinical outcomes and therapeutic strategies for glioma patients. Now, we could analyze epigenomic profiles to understand the epigenome-based evolution of gliomas. At first recurrence, the IDH-wild-type stem cell–like GBM phenotype

by G-CIMP-low showed molecular similarity to glial cell differentiation (De Souza et al., 2018). In our study, we found a series of CpG sites at genes involved in brain development or neuronal differentiation. These results could provide clues to the mechanism of the evolution of glioma. Indeed, genes involved in brain development and neuronal differentiation were strongly enriched among genes frequently methylated in tumors, for example, choline O-acetyltransferase (CHAT), GS homeobox 2 (GSX2), NK6 homeobox 1 (NKX6-1), paired box 6 (PAX6), retina and anterior neural fold homeobox (RAX) and distal-less homeobox 2 (DLX2) (Wu et al., 2010; Yu et al., 2013). The methylation of the genes involved in neuronal differentiation, in cooperation with other oncogenic events, may shift the balance from regulated differentiation towards gliomagenesis.

A recent report emphasized the relevance of DNA methylation profiles in somatic TERT pathway alterations (Ceccarelli et al., 2016). Indeed, functional enrichment analysis by Enrichr in our study found that these genes were enriched in the basic cancer-related biological processes, including mitotic recombination, DNA metabolism, and ErbB2 signaling pathway. These biological processes were significantly associated with telomere maintenance. Based on the final prognostic predictor, we analyzed the clinical follow-up data of these 12 glioma patients and found a positive correlation between methylation level and overall survival. Using *in vitro* experiments, we also confirmed that glioma cells with low methylation level would have higher migration ability and show resistance to temozolomide (or radiotherapy) compared to cells with high methylation level. Thus, these results suggested that the model constructed in this study could provide guidance for clinicians regarding the prognosis of various epigenetic subtypes.

CONCLUSION

Our research identified five different prognosis subgroups using glioma data in TCGA that differed either at the molecular level or in epidemiology, providing a more detailed explanation for glioma heterogeneousness. Additionally, our criteria will provide more targets for glioma precision medicine by identifying specific molecular markers for each subtype. Changes in DNA methylation can be used as markers to diagnose special subgroups, and clinicians can develop personalized treatments following these prognoses. Our approaches can also be used to study other tumors.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://cancergenome.nih.gov/

ETHICS STATEMENT

The protocol of this article was approved by the Institutional Review Board of the Hefei Institutes of Physical Science, Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

XC and ZF: conceived and designed the experiments. CZ and ZZ: collected the data. XC and CZ: performed the analysis. XC, HW, and ZF: participated in the discussion of the algorithm. XC and CZ: prepared and edited the manuscript. All authors have read and approved the final manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (81872066, 31571433, and 81773131), the innovative program of the Development Foundation of Hefei Center for Physical Science and Technology (2018CXFX004 and 2017FXCX008), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2018487).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00786/full#supplementary-material

REFERENCES

Aldape, K. D., Okcu, M. F., Bondy, M. L., and Wrensch, M. (2003). Molecular epidemiology of glioblastoma. *Cancer J.* 9, 99–106. doi: 10.1097/00130404-200303000-00005

Aquilanti, E., Miller, J., Santagata, S., Cahill, D. P., and Brastianos, P. K. (2018). Updates in prognostic markers for gliomas. *Neuro. Oncol.* 20, vii17–vii26. doi: 10.1093/neuonc/noy158

Arber, W., and Linn, S. (1969). DNA modification and restriction. *Annu. Rev. Biochem.* 38, 467–500. doi: 10.1146/annurev.bi.38.070169.002343

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* 164, 550–563. doi: 10.1016/j.cell.2015.12.028

Charlet, J., Tomari, A., Dallosso, A. R., Szemes, M., Kaselova, M., Curry, T. J., et al. (2017). Genome-wide DNA methylation analysis identifies MEGF10 as a novel epigenetically repressed candidate tumor suppressor gene in neuroblastoma. *Mol. Carcinog.* 56, 1290–1301. doi: 10.1002/mc.22591

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128. doi: 10.1186/1471-2105-14-128

Chen, X., Hao, A., Li, X., Du, Z., Li, H., Wang, H., et al. (2016). Melatonin inhibits tumorigenicity of glioblastoma stem-like cells *via the* AKT-EZH2-STAT3 signaling axis. *J. Pineal Res.* 61, 208–217. doi: 10.1111/jpi.12341

Crispatzu, G., Kulkarni, P., Toliat, M. R., Nurnberg, P., Herling, M., Herling, C. D., et al. (2017). Semi-automated cancer genome analysis using high-performance computing. *Hum. Mutat.* 38, 1325–1335. doi: 10.1002/humu.23275

Dawson, M. A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. *Cell* 150, 12–27. doi: 10.1016/j.cell.2012.06.013

De Souza, C. F., Sabedot, T. S., Malta, T. M., Stetson, L., Morozova, O., Sokolov, A., et al. (2018). A distinct DNA methylation shift in a subset of glioma CpG island

methylator phenotypes during tumor recurrence. *Cell Rep.* 23, 637–651. doi: 10.1016/j.celrep.2018.03.107

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *Lancet* 392, 777–786. doi: 10.1016/S0140-6736(18)31268-6

El-Osta, A. (2004). The rise and fall of genomic methylation in cancer. *Leukemia* 18, 233–237. doi: 10.1038/sj.leu.2403218

Fanelli, M., Caprodossi, S., Ricci-Vitiani, L., Porcellini, A., Tomassoni-Ardori, F., Amatori, S., et al. (2008). Loss of pericentromeric DNA methylation pattern in human glioblastoma is associated with altered DNA methyltransferases expression and involves the stem cell compartment. *Oncogene* 27, 358–365. doi: 10.1038/sj.onc.1210642

Ghosh, A., and Barman, S. (2016). Application of Euclidean distance measurement and principal component analysis for gene identification. *Gene* 583, 112–120. doi: 10.1016/j.gene.2016.02.015

Gustafsson, J. R., Katsioudi, G., Degn, M., Ejlerskov, P., Issazadeh-Navikas, S., and Kornum, B. R. (2018). DNMT1 regulates expression of MHC class I in post-mitotic neurons. *Mol. Brain* 11, 36. doi: 10.1186/s13041-018-0380-9

Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., et al. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7414–7419. doi: 10.1073/pnas.1703577114

Hill, V. K., Shinawi, T., Ricketts, C. J., Krex, D., Schackert, G., Bauer, J., et al. (2014). Stability of the CpG island methylator phenotype during glioma progression and identification of methylated loci in secondary glioblastomas. *BMC Cancer* 14, 506. doi: 10.1186/1471-2407-14-506

Issa, J. P. (2007). DNA methylation as a therapeutic target in cancer. *Clin. Cancer Res.* 13, 1634–1637. doi: 10.1158/1078-0432.CCR-06-2076

Jain, K. K. (2018). A critical overview of targeted therapies for glioblastoma. *Front. Oncol.* 8, 419. doi: 10.3389/fonc.2018.00419

Johannessen, L. E., Brandal, P., Myklebust, T. A., Heim, S., Micci, F., and Panagopoulos, I. (2018). MGMT gene promoter methylation status—assessment of two pyrosequencing kits and three methylation-specific PCR methods for their predictive capacity in glioblastomas. *Cancer Genomics Proteomics* 15, 437–446. doi: 10.21873/cgp.20102

Kanwal, R., Gupta, K., and Gupta, S. (2015). Cancer epigenetics: an introduction. *Methods Mol. Biol.* 1238, 3–25. doi: 10.1007/978-1-4939-1804-1_1

Klutstein, M., Nejman, D., Greenfield, R., and Cedar, H. (2016). DNA methylation in cancer and aging. *Cancer Res.* 76, 3446–3450. doi: 10.1158/0008-5472.CAN-15-3278

Koch, A., Joosten, S. C., Feng, Z., De Ruijter, T. C., Draht, M. X., Melotte, V., et al. (2018). Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* 15, 459–466. doi: 10.1038/s41571-018-0004-4

Krajewska, J., Chmielik, E., and Jarzab, B. (2017). Dynamic risk stratification in the follow-up of thyroid cancer: what is still to be discovered in 2017? *Endocr. Relat. Cancer* 24, R387–R402. doi: 10.1530/ERC-17-0270

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. doi: 10.1186/1471-2105-9-559

Liu, B., Song, J., Luan, J., Sun, X., Bai, J., Wang, H., et al. (2016). Promoter methylation status of tumor suppressor genes and inhibition of expression of DNA methyltransferase 1 in non–small cell lung cancer. *Exp. Biol. Med. (Maywood)* 241, 1531–1539. doi: 10.1177/1535370216645211

Masci, P. G. (2017). Negative risk markers for improving prediction of heart failure: risk stratification implementation or simply the other side of existing risk scores? *Int. J. Cardiol.* 249, 328–329. doi: 10.1016/j.ijcard.2017.09.196

Perez, R. F., Tejedor, J. R., Bayon, G. F., Fernandez, A. F., and Fraga, M. F. (2018). Distinct chromatin signatures of DNA hypomethylation in aging and cancer. *Aging Cell* 17, e12744. doi: 10.1111/acel.12744

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

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

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573. doi: 10.1093/bioinformatics/btq170

Witt, H., Gramatzki, D., Hentschel, B., Pajtler, K. W., Felsberg, J., Schackert, G., et al. (2018). DNA methylation-based classification of ependymomas in adulthood: implications for diagnosis and treatment. *Neuro. Oncol.* 20, 1616–1624. doi: 10.1093/neuonc/noy118

Wu, X., Rauch, T. A., Zhong, X., Bennett, W. P., Latif, F., Krex, D., et al. (2010). CpG island hypermethylation in human astrocytomas. *Cancer Res.* 70, 2718–2727. doi: 10.1158/0008-5472.CAN-09-3631

Yang, H., Wu, J., Zhang, J., Yang, Z., Jin, W., Li, Y., et al. (2019). Integrated bioinformatics analysis of key genes involved in progress of colon cancer. *Mol. Genet. Genomic Med.* 7, e588. doi: 10.1002/mgg3.588

Yarus, M. (1969). Recognition of nucleotide sequences. *Annu. Rev. Biochem.* 38, 841–880. doi: 10.1146/annurev.bi.38.070169.004205

Yu, Z. Q., Zhang, B. L., Ren, Q. X., Wang, J. C., Yu, R. T., Qu, D. W., et al. (2013). Changes in transcriptional factor binding capacity resulting from promoter region methylation induce aberrantly high GDNF expression in human glioma. *Mol. Neurobiol.* 48, 571–580. doi: 10.1007/s12035-013-8443-5

Zang, L., Kondengaden, S. M., Che, F., Wang, L., and Heng, X. (2018). Potential epigenetic-based therapeutic targets for glioma. *Front. Mol. Neurosci.* 11, 408. doi: 10.3389/fnmol.2018.00408.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 September 2019
doi: 10.3389/fgene.2019.00810

GTn Repeat Microsatellite Instability in Uterine Fibroids

Bineta Kénémé[1] and Mbacké Sembène[1,2]*

[1] GenGesPop, Cheikh Anta Diop University, Animal Biology, Dakar-Fann, Senegal, [2] Biopass, IRD, Dakar-Bel Air, Senegal

Background: *Type I collagen* is a triple helix structure with two α1 and one α2 chains. Coordinated biosynthesis of α1 and α2 subunits is very important for tissue morphogenesis, growth, and repair. In contrast, abnormal deposition in response to proinflammatory cytokines is associated with organ dysfunction. In humans, *COL1A2* contains two microsatellite loci: one located at the 5'-flanking region is composed of poly CA and poly CG; the other located in the 1st intron is constituted of poly GT. Expression of *COL1A2* has been noted in gastric cancer and was positively correlated with degree of invasion and metastases. But no genetic study taking into account polymorphism of *COL1A2* in uterine fibroids has been undertaken.

Methods: In this study, repeated dinucleotide $\mathbf{GT_n}$ of intron 1 ***COL1A2*** was highlighted in 55 patients with uterine fibroids (UF). Clinical and pathological data were obtained from patient's records, and other parameters were recorded. Mutation Surveyor version 5.0.1, DnaSP version 5.10, MEGA version 7.0.26, and Arlequin version 3.5.1.3 were used to determine genetics parameters. To estimate genetic variation according to epidemiological parameters, index of genetic differentiation (Fst) and genetic structure (AMOVA) were determined with Arlequin version.

Results: Based on reference microsatellite pattern $\mathbf{(GT)_{14}CT(GT)_3CT(GT)_3}$, 15 haplotypes were found. Among the 15 haplotypes, 12 have mutation at position **2284C > G** and 7 at position **2292C > G**. Insertions of repeated dinucleotide GT_n were found on three haplotypes against eight haplotypes in which they are deletions. Intron 1 of *COL1A2* gene exhibits high genetic diversity in uterine fibroids with 35.34% polymorphic sites, 95.74% of which were parsimoniously variable and an average number of nucleotide difference of 10.442, which reflects an important genetic variability. According to epidemiological parameters, our results showed, for the first time, a genetic structuring of uterine fibroids according to ethnicity, marital status, use of contraception, diet, and physical activity, beyond confirming the involvement dinucleotide length polymorphism GT_n in occurrence of uterine fibroids in Senegalese women.

Conclusion: Results obtained open up avenues for understanding the mechanisms involved in the racial variation in the prevalence of uterine fibroids as well as the predisposing factors.

Keywords: uterine fibroid, COL1A2 polymorphism, risk factors, Senegal, microsatellite genetic marker

OPEN ACCESS

Edited by:
Barbara Karen Dunn, National Institutes of Health (NIH), United States

Reviewed by:
Pawel Buczkowicz, Gene42, Inc., Canada
Jun Zhong, National Cancer Institute (NCI), United States

****Correspondence:***
Bineta Keneme
bineta.keneme@ucad.edu.sn

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *28 January 2019*
Accepted: *02 August 2019*
Published: *20 September 2019*

Citation:
Kénémé B and Sembène M (2019) GTn Repeat Microsatellite Instability in Uterine Fibroids. Front. Genet. 10:810. doi: 10.3389/fgene.2019.00810

INTRODUCTION

Uterine fibroids (UF), more commonly known as myomas or uterine leiomyomas, are the most common benign tumors of female reproductive organs. They are associated with significant morbidity and therefore constitute a real public health problem. UF, which are highly variable within uterus, develop at the expense of smooth muscle and are often separated from the myometrium by a pseudocapsule associated with connective tissue condensation (Audebert, 1990). Heterogeneity of UF localization and their progression in the same patient illustrate the complex biological mechanism involved in their development. Clinically, UF are firm, stiff nodular tumours, a fact confirmed by biomechanical studies. Proteins of extracellular matrix, especially interstitial collagens, are responsible for this property of "firmness" and mechanical strength of tissue. Indeed, UF have an accumulation of altered collagens and different amounts of glycosaminoglycans and a proliferation of cells, which is by definition a fibrosis. A complete understanding of the role of extracellular matrix proteins, in particular collagen, and their effect on the growth and development of UF becomes an important issue for elucidating molecular mechanisms involved in their etiology.

Located on chromosome 7, *COL1A2* is an essential component of matrix tissue. It is predominantly produced by mesenchymal cells such as fibroblasts, osteoblasts, and smooth muscle cells. Transcription of *COL1A2* is under control of a regulatory complex that includes several DNA elements and several trans-activating factors. During the last two decades, *type I alpha chain collagen 2* (*COL1A2*) has been considered as an informative model for studying principles that govern the control of extracellular matrix transcription for normal and fibrotic tissues (Kirkland, 2009; Krasny et al., 2010; Trojonowska, 2002; Yasul et al., 2004). In humans, *COL1A2* contains two microsatellite loci: one located at the 5'-flanking region of the gene is composed of poly CA and poly CG; the other located in the 1st intron is constituted of poly GT. In a study led by Akai et al. (1999), it has been shown that complete transcription of *COL1A2* gene is regulated by these repeated dinucleotides. Analysis of polymorphism in these two regions indicates that these two sequences show a variation in their repetition number, suggesting that these dinucleotides constitute microsatellites. Lei et al. (2005) hypothesized that GT_n polymorphism triggers transcription of the gene, and variation in the number of repetitions can partly be responsible for the difference in transcriptional activity. In this study, we evaluate instability of repeated dinucleotide $\mathbf{GT_n}$ in Senegalese patients with UF.

MATERIALS AND METHODS

Clinical Sampling

Tumor tissue samples were collected from 55 patients with UF (from Military Hospital of Ouakam and General Hospital of Grand Yoff). Clinical and pathological data were recorded including age, ethnicity, age at menarche, marital status, number of pregnancies, number of childbirth, hormonal contraception, diet, and physical activity (**Table 1**). None of the patients surveyed claimed to consuming alcohol and using tobacco, which is why these factors are not included in this study.

DNA Extraction, Amplification, and Sequencing of Intron 1 *COL1A2* Gene

Total DNA of each sample was extracted using Qiagen protocol (Qiagen Dneasy Tissue kit). After extraction, repeated dinucleotide $\mathbf{GT_n}$ were amplified using forward 5'-TGTCT ACCACTGCATAATTTC-3 and reverse 5'-AATATGAACTCG GTAATGTGA-3' primers (Lei et al., 2005). The 35 cycle PCR for *COL1A2* intron 1 amplification was carried out using 4 µl of human genomic DNA in a 50 µl reaction mixture, which contained 0.1 µl of Taq DNA polymerase, 2.5 µl of forward and reverse primers, 1 µl of magnesium chloride, 2 µl of mix dNTPs, and 5 µl of 10X ammonium sulfate buffer. Thermal cycle conditions for amplification PCR consisted of 1st step-3 min

TABLE 1 | Clinical and pathological characteristics of 55 cases analyzed.

Epidemiological factors	Number of patients (%)
Age (n = 36)	
≤35	11 (30.55%)
]35–45]	18 (50%)
> 45	7 (19.45%)
Ethnicity (n = 39)	
Wolof	13 (33.33%)
Sérère	4 (10.26%)
Lébou	7 (17.95%)
Bambara	3 (7.69%)
Diola	5 (12.82%)
Alpulaar	7 (17.95%)
Marital status (n = 31)	
Single	8 (25.80%)
Married	20 (64.52%)
Divorced	3 (9.68%)
Age at menarche (n = 18)	
≤12	1 (5.56%)
]12–15]	13 (72.22%)
> 15	4 (22.22%)
Number of pregnancies (n = 31)	
0	20 (64.51%)
I	4 (12.91%)
II	4 (12.91%)
III	1 (3.22%)
> III	2 (6.45%)
Number of childbirth (n = 33)	
0	23 (69.70%)
I	7 (21.21%)
II	1 (3.03%)
III	2 (6.06%)
> III	0 (0%)
Hormonal contraception (n = 23)	
Yes	2 (8.69%)
No	21 (91.31)
Diet (n = 23)	
Meat preference	7 (30.43%)
Vegetarian preference	6 (26.09%)
No preference	10 (43.48%)
Physical activity (n = 23)	
Yes	5 (21.74%)
No	18 (78.26%)

cycle of initial denaturation at a temperature of 94°C, followed by 2nd step consisting of 35 cycles each of 45 s of denaturation at 94°C, annealing at 60°C/1min and primer extension at 72°C/1 min, and 3rd step: final extension or polymerization at 72°C for 10 min. After PCR reaction, all products were electrophoresed on 1.5% agarose gel, followed by its analysis in an UVitec Gel Documentation system for imaging the gel and to determine the amplicon lengths. Sequencing reactions were performed in a thermal cycler MJ Research PTC-225 Peltier type with ABI PRISM BigDye TM Terminator Cycle kit. Each sample was sequenced using forward primer. Fluorescent fragments were purified with the BigDye Xterminator purification protocol. The samples were suspended in distilled water and subjected to electrophoresis in 3730xl ABI sequencer (Applied Biosystems).

Molecular Analysis

To determine length polymorphism of dinucleotide **GT_n** of intron 1 *COL1A2* gene, the raw sequencing data were submitted to Mutation Surveyor software version 5.0.1 (www.softgenetics.com). This program can directly compare chromatograms with genomic DNA of reference sequence of *COL1A2* (NT_007933_94023373). Alignment of the sequences was carried out using BioEdit software version 8.0.5 and ClustalW algorithm (Thompson et al., 1994). Sequences obtained (Hall, 1999) were thoroughly checked, cleaned, and aligned to identify homologies among sites, and also to perform other phylogenetic analysis including the determination of variability index and genetic diversity as well as the parameters of genetic differentiation. Genetic variability parameters (number of polymorphic sites, total number of haplotype, average number of nucleotide difference K) were obtained through DnaSP 5.10 software (Librado and Rozas, 2009) and MEGA 7.0.26 (Kumar et al., 2016). To estimate genetic variation according to epidemiological parameters, the factor of genetic differentiation (Fst) and the analysis of molecular variance (AMOVA) were determined with Arlequin software version 3.5.1.3 (Excoffier and Lischer, 2010). Values of P less than 0.05 are considered significant at a 5% confidence interval.

RESULTS

Mutations Status of Microsatellite GT_n

COL1A2 was sequenced in 55 tumour tissues. Of these sequences, five were removed from the genetic analysis because of a strong polymorphism. Based on the microsatellite reference pattern in the form $(GT)_{14}CT(GT)_3CT(GT)_3$, 15 haplotypes were found in 50 Senegalese women with UF (**Table 2**). These haplotypes indicate a variation in GT repetition number ranging from 13 to 25 (**Figure 1**).

Of the 15 haplotypes, 12 have mutation at position **2284C > G** (first site that interrupts GT dinucleotide repeat) and 7 have mutation at **2292C > G** (2nd site that interrupts GT dinucleotide repeat). **Insertions** of repeated dinucleotide GT_n were found on three haplotypes (microsatellite elongation) compared to eight haplotypes in which they were **deletions** (microsatellite shortening). Haplotype 7 representing 20% of the haplotypes was characterized by the presence of two types of transversions **2284C > G** and **2292C > G**. Haplotypes 11 (14%) and 8 (12%) were respectively characterized by deletions at position **2280-2283_DelGTGT** and **2282-2283_DelGT** (**Table 2**). Some microsatellite length polymorphisms of **GT_n** *COL1A2* were summarized in **Figure 2**.

COL1A2 Intron 1 Polymorphisms

Intron 1 of *COL1A2* gene exhibits high genetic diversity in UF with 35.34% polymorphic sites, 95.74% of which were parsimoniously variable. Average number of nucleotide differences was 10.442 (**Table 3**). The high haplotypic diversity (Hd = 0.9984) and the low nucleotide diversity (Pi = 0.0877) showed a rapid evolution of microsatellite polymorphism in UF in Senegalese women.

Microsatellite GT_n Instability and Genetic Differentiation

Depending on epidemiological parameters studied, only repeated dinucleotide is taken into account in analysis. This allows us to highlight the role of the length polymorphism of intron 1 *COL1A2*. Results obtained show a variable expressivity of dinucleotide GT_n in

TABLE 2 | Length and pattern polymorphism of repeated dinucleotide GT_n of intron 1 *COL1A2* gene in uterine fibroids.

Haplotype	Number (%)	Microsatellite pattern	Variants
H1	5 (10%)	$(GT)_{14}CT(GT)_3CT(GT)_3$	Wide type
H2	4 (8%)	$(GT)_{17}CT(GT)_3CT(GT)_3$	2276_2281_InsGTGTGT
H3	2 (4%)	GT_{25}	2276_2281_InsGTGTGT; 2284C > G; 2292C > G
H4	1 (2%)	$(GT)_{21}CT(GT)_3$	2276_2281_InsGTGTGT; 2284C > G
H5	2 (4%)	GT_{23}	2281-2282_InsGT; 2284C > G; 2292C > G
H6	2 (4%)	$(GT)_{18}CT(GT)_3$	2284C > G
H7	10 (20%)	GT_{22}	2284C > G; 2292C > G
H8	6 (12%)	(GT_{21})	2282-2283_DelGT; 2284C > G; 2292C > G
H9	2 (4%)	$(GT)_{13}CT(GT)_3CT(GT)_3$	2282-2283_DelGT
H10	1 (2%)	$(GT)_{17}CT(GT)_3$	2282-2283_DelGT; 2284C > G
H11	7 (14%)	$(GT)_{16}CT(GT)_3$	2280-2283_DelGTGT; 2284C > G
H12	4 (8%)	GT_{20}	2280-2283_DelGTGT; 2284C > G; 2292C > G
H13	1 (2%)	GT_{19}	2276-2281_DelGTGTGT; 2284C > G; 2292C > G
H14	2 (4%)	$(GT)_{15}CT(GT)_3$	2276-2281_DelGTGTGT; 2284C > G
H15	1 (2%)	GT_{18}	2274-2281_DelGTGTGTGT; 2284C > G; 2292C > G

Ins, insertion; Del, deletion.

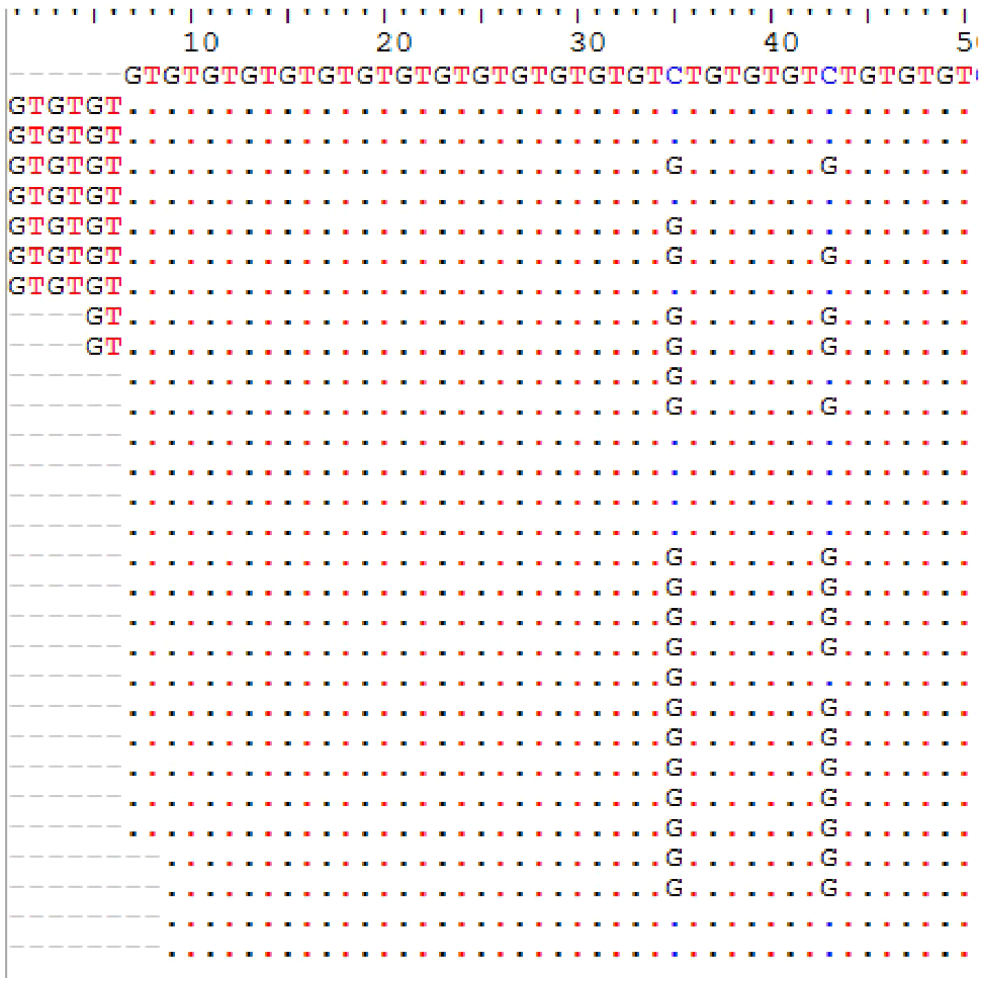


FIGURE 1 | Microsatellite GT_n length and pattern polymorphism in uterine fibroids.

Senegalese women with UF (**Table 4**). For age parameter, tumoral tissues were genetically more different in women under 35 (Fst = 0.04237) and those over 45 (Fst = 0.06574) compared to women aged 35–45 years (Fst = 0.03152). This differentiation is more significant between the two extremes (under 35 and over 45). This heterogeneity of repeated dinucleotide GT_n polymorphism is more noticeable among women of Bambara, Sérère, Lébou and Alpulaar ethnic groups, UF being genetically homogeneous in Wolof and Diola women. Strong genetic differentiation was noted between Wolof and Alpulaar.

According to marital status, UF seem to have the same genetic characteristics in single women (Fst = 0.08964), unlike married women (Fst = 0.10771) and divorced women (Fst = 0.23556), where there is a strong genetic differentiation within each group. However, no statistically significant differentiation is noted between these groups.

For the age at menarche variable, only one woman who reported having menarche before the age of 12 had a different haplotype than the remaining 17 women on which age of menarche data was available. Further investigation with a larger patient cohort is required to determine the significance of this observation. As for the number of pregnancies, no statistically significant differentiation is noted between the sub-groups, but nevertheless, we notice a strong genetic differentiation in three women with three and more than three pregnancies. It is the same for the number of childbirth. Compared to hormonal contraception on the one hand and physical activity on the other hand, we noted an important genetic differentiation between sub-groups.

Polymorphism of repeated dinucleotide GT_n was genetically different within women who have a meat preference and those who have no food preference. Women who are preferably vegetarians were genetically homogeneous (**Table 4**).

Microsatellite GT_n Instability and Molecular Variance Analysis

The Fst values are further explained by molecular variance analysis (**Table 5**). Repeated dinucleotide GT_n analysis showed that UF are genetically structured according to ethnicity ($p = 0.03421^*$), marital status ($p = 0.00782^{**}$), hormonal contraception ($p = 0.00098^{***}$), dietary preference ($p = 0.04301^*$), and physical activity ($p = 0.00684^{***}$). In other words, molecular mechanisms of *COL1A2* involved in etiology of UF in Senegalese women

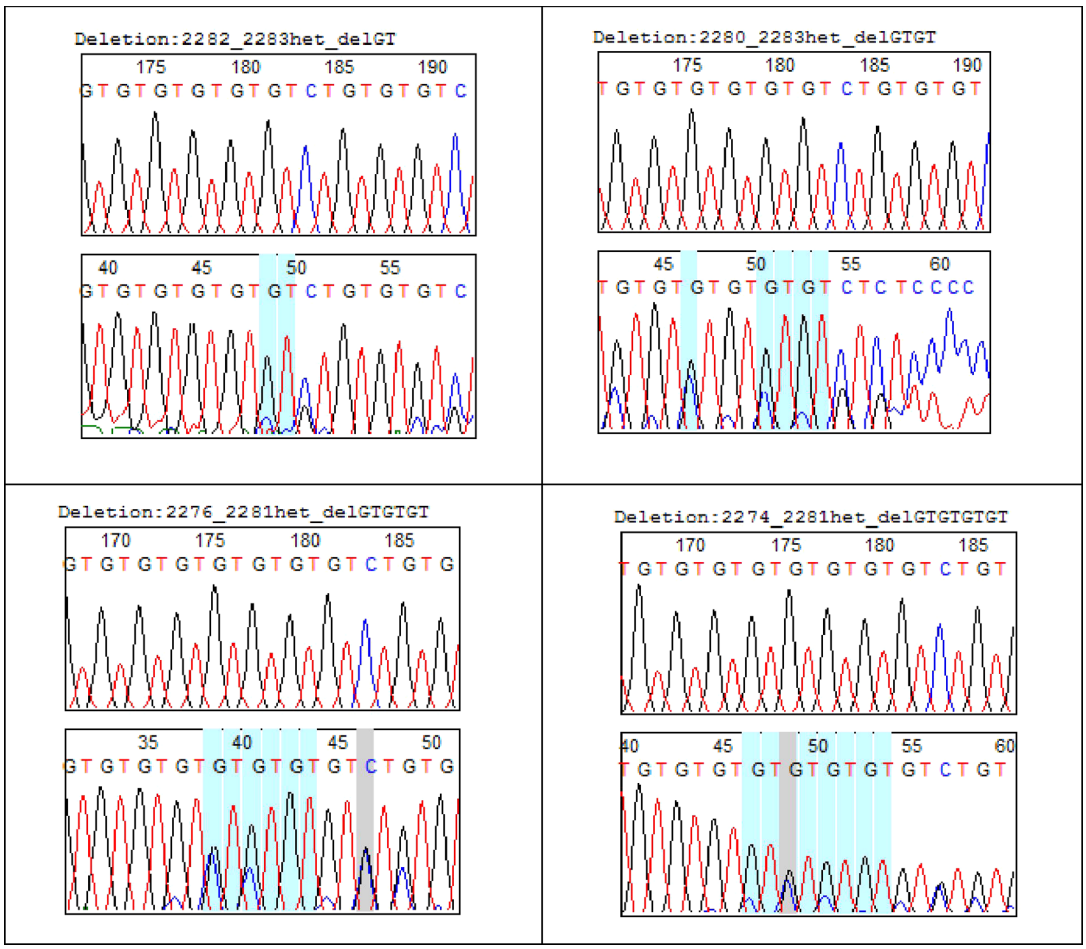


FIGURE 2 | GT_n *COLLA2* deletion in uterine fibroids.

TABLE 3 | Index of variability and genetic diversity of intron 1 *COL1A2* in fibroid cases.

Variability index		
Parameters	**Number**	**Percentage**
Number of sequences	50	
Number of sites	133	
Monomorphic sites	86	64.66%
Polymorphic sites	47	35.34%
Singleton variable sites	2	4.26%
Parsimony informative sites	45	95.74%
Average number of nucleotide differences (k)	10.442	
Genetic diversity index		
Pi ± variance	0.0877 ± 0.00002	
Hd ± variance	0.9984 ± 0.00003	

Hd, haplotypic diversity; Pi, nucleotide diversity.

are modulated by risk factors such as ethnicity, marital status, hormonal contraception, diet, and physical activity. Indeed, the polymorphism of the repeated dinucleotide GT_n of intron 1 *COL1A2* in UF is explained to:

- 8.83% by differentiation between women of different ethnic groups;
- 10.57% by a differentiation between women according to their marital status;
- 24.88% by differentiation according to whether or not use of hormonal contraception;
- 7.94% by dietary preference; and
- 15.10% by a differentiation according to physical activity.

Since there is no multivariate analysis and the sample sizes are small for some of these variables, more research is needed to highlight these results.

DISCUSSION

COL1A2 Polymorphisms in Uterine Fibroids

Located on chromosome 7, *COL1A2* is an essential component of the tissue matrix. It is predominantly produced by mesenchymal cells such as fibroblasts, osteoblasts, and smooth muscle cells (Rossert et al., 2000). Transcription of *COL1A2* is under control

TABLE 4 | Degree of genetic differentiation of repeated dinucleotide GT_n of *COL1A2* gene in relation to the epidemiological parameters studied.

Epidemiological parameters	**Genetic differentiation (Fst)**		
Groups Sub-groups	**Within sub-groups**	**Between sub-groups**	
Age	**Fst**	**Between sub-groups**	**Fst (P-value)**
≤35	0.04237	≤35 &]35–45]	0.01274 (0.32422)
]35–45]	0.03152	≤35 & >45	0.20818 (0.07812)
> 45	0.06574	]35–45] & >45	–0.02089 (0.46582)
Ethnicity			
Wolof	0.00180	Wolof & Sérère	0.15238 (0.17773)
Sérère	0.20123	Wolof & Lébou	0.06714 (0.29785)
Lébou	0.14805	Wolof & Bambara	–0.03106 (0.99902)
Bambara	0.30276	Wolof & Diola	0.10515 (0.19531)
Diola	–0.00183	Wolof & Alpulaar	0.18605 (0.07324)
Alpulaar	0.11805	Sérère & Lébou	–0.16129 (0.76562)
		Sérère et Bambara	0.06667 (0.99902)
		Sérère & Diola	0.15152 (0.15918)
		Sérère & Alpulaar	0.12513 (0.12891)
		Lébou & Bambara	–0.33333 (0.99902)
		Lébou & Diola	–0.09091 (0.60938)
		Lébou & Alpulaar	–0.08691 (0.72852)
		Bambara & Diola	–0.40000 (0.99902)
		Bambara & Alpulaar	–0.52941 (0.99902)
		Diola & Alpulaar	0.13754 (0.11523)
Marital status			
Single	0.08964	Single & Married	0.03216 (0.24805)
Married	0.10771	Single & Divorcée	0.39683 (0.99902)
Divorced	0.23556	Married & Divorced	0.38444 (0.99902)
Age at menarche			
≤12	0.20408	≤12 &]12 – 15]	–0.45799 (0.91992)
]12 – 15]	–0.14031	≤12 & >15	0.00000 (0.39453)
>15	–0.08291	] 12 – 15] & >15	–0.01850 (0.48242)
Number of pregnancies			
0	0.00988	0 & I	–0.06950 (0.70508)
I	0.02840	0 & II	0.00284 (0.38184)
II	–0.11721	0 & III	–0.01720 (0.49707)
III	0.34321	0 & > III	0.42177 (0.99902)
> III	0.34321	I & II	–0.35429 (0.80273)
		I & III	–0.10092 (0.70703)
		I & > III	–0.17647 (0.99902)
		II & III	0.13333 (0.31641)
		II & > III	–1.00000 (0.99902)
		III & > III	1.00000 (0.99902)
Number of childbirth			
0	–0.06871	0 & I	0.07169 (0.18652)
I	–0.13999	0 & II	–0.41520 (0.99902)
II	0.26463	0 & III	–0.41520 (0.99902)
III	0.26463	I & II	–0.16049 (0.99902)
		I & III	–0.16049 (0.99902)
		II & III	0.00000 (0.99902)
Hormonal contraception			
Yes	0,17376	Yes & No	0.24883 (0.07129)
No	–0,07280		
Diet			
1 Meat preference	0.01042	1 & 2	0.17374 (0.07520)
2 Vegetarian preference	–0.01543	1 & 3	–0.11892 (0.92285)
3 No preference	0.01406	2 & 3	0.20431 (0.04785)
Physical activity			
Yes	0.14435	Yes & No	0.15107 (0.06445)
No	0.07661		

TABLE 5 | Genetic structuring of GT_n *COL1A2* according to epidemiological parameters.

Epidemiological parameters	Source of variation	Percentage of variation	Fst (P-value)
Age	Within sub-groups	96.03862	0.03961 (0.06843)
	Between sub-groups	3.96138	
Ethnicity	Within sub-groups	91.16560	0.08834 (0.03421)
	Between sub-groups	8.83440	
Marital status	Within sub-groups	89.42471	0.10575 (0.00782)
	Between sub-groups	10.57529	
Age at menarche	Within sub-groups	108.72040	–0.08720 (0.88368)
	Between sub-groups	–8.72040	
Number of pregnancies	Within sub-groups	95.80665	0.04193 (0.21994)
	Between sub-groups	4.19335	
Number of childbirth	Within sub-groups	105.40304	–0.05403 (0.79570)
	Between sub-groups	–5.40304	
Hormonal contraception	Within sub-groups	75.11658	0.24883 (0.00098)
	Between sub-groups	24.88342	
Diet	Within sub-groups	92.05111	0.07949 (0.04301)
	Between sub-groups	7.94889	
Physical activity	Within sub-groups	84.89311	0.15107 (0.00684)
	Between sub-groups	15.10689	

of a regulatory complex that includes several DNA elements and several trans-activating factors. In humans, *COL1A2* contains two microsatellite loci: one located at the 5'-flanking region of the gene is composed of poly CA and poly CG; the other located in the 1st intron is constituted of poly GT. In this study, microsatellite polymorphism GT_n of intron 1 *COL1A2* was highlighted in cases of UF in Senegalese women. Based on microsatellite reference pattern that is $(GT)_{14}(CT)(GT)_3(CT)(GT)_3$, 15 haplotypes were found. These haplotypes indicate a variation in number of GT repeats ranging from 13 to 25. Of the 15 haplotypes, 12 have the 2284C > G mutation and 7 have the 2292C > G mutation. Insertions of the dinucleotide GT_n were found on three haplotypes (microsatellite elongation) compared to eight haplotypes in which they were deletions (microsatellite shortening). This suggests an altered mechanism of the role of *COL1A2* gene in UF. Indeed, in a study led by Lei et al. (2005), it has been hypothesized that intron 1 GT_n polymorphism triggers transcription of gene and variation in number of repeats may be partly responsible for the difference in transcriptional activity. In addition, about 200 different chromosomal abnormalities have been described in UF including long-arm translocations of chromosome 7 occurring in about 17% of karyotypically abnormal UF (Sandberg, 2005). In contrast to normal tissues where collagen is organized into long, thin, wavy fibrils parallel to the epithelial boundary, collagen fibrils in the tumor stroma are thicker and shorter (Cho et al., 2015). In epithelial ovarian cancer, collagenous pathways perpendicular to the epithelial boundary have been observed (Adur et al., 2014).

Intron 1 of *COL1A2* gene exhibits high genetic diversity in UF with 35.34% polymorphic sites, 95.74% of which are parsimoniously variable and an average number of nucleotide differences of 10.442, which reflects an important genetic variability. This could be explained by the fact that compared to the myometrium, in UF, not only expression of collagen genes increases (Stewart et al., 1994), but also amount of mature reticulated collagen protein is increased and the more important is modified (Leppert et al., 2004). UF are firm, stiff nodular tumors, a fact understood by all clinicians and confirmed by biomechanical studies (Rogers et al., 2008; Jayes et al., 2013). Extracellular matrix (ECM) proteins, including interstitial collagens, are responsible for this property of "firmness" and mechanical strength of tissues. ECM is a structure that has a supporting role, but on the other hand, it provides signals to cells that determines their behavior. The role of ECM and mechanotransduction as an important signaling factor in human uterus is just beginning to be appreciated. ECM is not just substance surrounding cells, but rigidity compresses cells or stretches them into signals converted into chemical changes, depending on amount of collagen, crosslinking and hydration, as well as other components of ECM. Since connective tissue integrity, architecture, and function result from specific interactions between collagen and other components of ECM, the presence of abnormal collagen chains may have a strong influence on metabolism of non-collagenic components (Tenni et al., 1988). According to study by Hauptman et al. (2018) in colorectal cancers, the results showed that 9 of 16 genes that show differential expression in carcinomas compared to adenomas are components of ECM. Among these components, two collagen type I proteins (COL1A1, COL1A2) are significantly over-regulated in cancerous tissues compared to normal tissues. Studies on cell lines suggest that type I collagen adhesion promotes intracellular signaling pathways.

Microsatellite GT_n Instability in Uterine Fibroids: Correlation With Epidemiological Parameters

Ethnicity

In addition to great variability, repeated dinucleotide GT_n of intron 1 *COL1A2* exhibits heterogeneity given to clinico-pathological parameters in women with UF. Heterogeneity of predisposing factors involved in UF illustrates the complex biological mechanism involved in their development. This suggests the involvement of several molecular mechanisms in occurrence of UF. Prospective studies with larger number of samples would strengthen the correlation observed in the current study. Epidemiological data have mentioned racial disparity in occurrence of UF. Ethnicity has a major influence on development and clinical severity of UF. African-American women develop UF at higher frequency and with more severe symptoms. Hispanic women have an intermediate disease profile, and Caucasian women are the least severely affected ethnic group (Velebil et al., 1995; Baird et al., 2003; Wise et al., 2012). It appears that increased incidence and severity of disease in African-American women may be due to a combination of specific genetic and environmental factors that are not independent risk factors for the disease (Commandeur et al., 2015). In this study, we took ethnicity into account, although these women are all black. UF are genetically heterogeneous in Bambara, Sérère, Lébou, and Alpulaar and more homogeneous in Diola and Wolof. In a study conducted by Thiaw (2018) on ethnic diversity of Senegalese population (unpublished data), analysis of GT_n pattern polymorphism shows a genetic differentiation between Diola and Wolof compared to other

ethnic groups. This differentiation may explain in part 8.83% of genetic structure of *COL1A2* observed in UF by ethnicity.

Marital Status

Genetic differentiation observed (10.57%) is also explained by the differentiation between women according to their marital status; greater differentiation is observed among married and divorced women compared to single. This could be explained by the difference in hormonal status in these women. Studies of Barrett et al. (2015) on ovarian steroid status in marital status showed that estradiol was higher among married women than among unmarried women (β = 0.19, 95% CI: 0.02–0.36) as well as progesterone (β = 0.19, 95% CI: 0.01–0.39). In addition, many clinical observations indicate that the development of UF is related to hormonal status (Ross et al., 1986). For example, UF do not occur in prepubertal women and are rarely seen in adolescent girls (Fields and Neinstein, 1996).

Hormonal Contraception

Our results also indicated that 24.88% of differentiation observed in cases of UF is explained by a differentiation following hormonal contraception use. Relationship between oral contraceptives and UF has been largely elucidated. But epidemiological data between contraceptive use and UF seems controversial. Published studies show a reduction or absence of risk between oral contraceptives use combined with appearance of UF (Berisavac et al., 2009). One study has shown that oral contraception may play a role in development of UF. Others have found no association between occurrence of UF and use of contraception (Parazzini et al., 1992).

Diet

In relation to diet, a positive correlation was noted between dietary preference and genetic expression of *COL1A2* in UF (7.94% of genetic differentiation). Genetic differentiation is more observed in patients with meat preference. Recently, Wise and Laughlin-Tommaso (2016) published results on relationship between dietary fat intake and UF risk in African-American women, confirming an increased risk associated with consumption of omega-3 fatty acids long chain. They validated hypothesis that a diet rich in fruits and vegetables reduced risk. According to studies of Chiaffarino (1999), women with UF consume beef, other red meats, and ham more frequently and have less frequent consumption of green vegetables, fruits, and fish. Multivariate rib ratios were 1.7 for beef and other red meats, 1.3 for ham, and 0.8 for fruit consumption. Limitation of this current diet study is the lack of data on total energy intake because information was collected only on the frequency of vegetable consumption compared to red meat and in interviews with patients. Further research would be interesting to evaluate the effect of fat intake on uterine fibroids biology.

Physical Activity

There have been few studies on effect of physical activity on risk of developing UF. Nevertheless, our results showed a genetic structuring of UF according to practice or not of sport (15.10% of genetic differentiation). Since this is a modifiable factor, more research is needed to evaluate effects of physical activity on UF biology.

CONCLUSION

Results obtained show, for the first time, a genetic structuring of UF according to ethnicity, marital status, use of contraception, diet, and physical activity, beyond confirming the involvement of *COL1A2* gene, in particular dinucleotide length polymorphism GT_n in occurrence of UF in Senegalese women. In addition to this, results obtained open up avenues for understanding the mechanisms involved in racial variation in the prevalence of UF as well as the predisposing factors. Given the admitted results, it is clear that more research is needed to determine risk factors associated with appearance and growth of UF, as they cause significant morbidity and affect quality of life. A clear overview of the epidemiology of UF has not yet been realized and future research on modifiable risk factors such as vegetarian diet, contraception, physical activity, among others could inform the prevention of myomas and provide new non-surgical approaches to treatment.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of World Medical Association's Declaration of Helsinki. The protocol was approved by the Institutional Ethics Committee on Human Research of Cheikh Anta Diop University (Reference: Protocol 0267/2017/CER/UCAD). All subjects gave written informed consent according to a standardized form.

AUTHOR CONTRIBUTIONS

BK performed molecular analysis, organized the database, performed data analysis, and wrote the first draft of the manuscript. MS contributed to conception and design of the study, revised the manuscript, and read and approved the submitted version.

ACKNOWLEDGMENTS

We acknowledge the African Center of Excellence for Mother and Child Health (ACE-MCH), Grant number B041715-P00041/2017 and University Cheikh Anta Diop, Dakar, UCAD (http://www.ucad.sn) for technical support. We are most grateful to all Senegalese women who participated in the present study. We are extremely grateful to Dr Daouda CISS Dr Sidy KA and Pr Ahmadou DEM who helped with the collection of samples. Also Pr SEMBENE the head of molecular biology platform of BIOPASS institute of Senegal for all the molecular studies done.

REFERENCES

Adur, J., Pelegati, V. B., and De Thomaz, A. A. (2014). Second harmonic generation microscopy as a powerful diagnostic imaging modality for human ovarian cancer. *J. Biophotonics.* 7, 37–48. doi: 10.1002/jbio.201200108

Akai, J., Kimura, A., and Hata, R. I. (1999). Transcriptional regulation of the human type I collagen a2 (*COL1A2*) gene by the combination of two dinucleotide repeats. *Gene* 239, 65–73. doi: 10.1016/S0378-1119(99)00380-7

Audebert, A. (1990). Endométriose externe: histogénèse, étiologie et évolution naturelle. *Rev. Praticien.* 40, 1077–1081.

Baird, D. D., Dunson, D. B., Hill, M. C., Cousins, D., and Schectman, J. M. (2003). High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. *Am. J. Obstet. Gynecol.* 188, 100–107. doi: 10.1067/mob.2003.99

Barrett, E. S., Tran, V., Thurston, S. W., Frydenberg, H., Lipson, S. F., Thune, I., et al. (2015). Women who are married or living as married have higher salivary estradiol and progesterone than unmarried women. *Am. J. Hum. Biol.* 27 (4), 501–507. doi: 10.1002/ajhb.22676

Berisavac, M., Sparic, R., and Argirovic, R. (2009). Contraception: modern trends and controversies. *Srp. Arh. Celok. Lek.* 137 (5–6), 310–319. doi: 10.2298/SARH0906310B

Chiaffarino, F. (1999). Diet and uterine myomas. *Obstet. Gynecol.* 94 (3), 395–398. doi: 10.1016/S0029-7844(99)00305-1

Cho, A., Howell, V. M., and Colvin, E. K. (2015). The extracellular matrix in epithelial ovarian cancer-a piece of a puzzle. *Front. Oncol.* 5, 1–16. doi: 10.3389/fonc.2015.00245

Commandeur, A. E., Styer, A. K., and Teixeira, J. M. (2015). Epidemiological and genetic clues for molecular mechanisms involved in uterine leiomyoma development and growth. *Hum. Reprod. Update.* 21 (5), 593–615. doi: 10.1093/humupd/dmv030

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resources.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Fields, K. R., and Neinstein, L. S. (1996). Uterine myomas in adolescents: case reports and a review of the literature. *J. Pediatr. Adolesc. Gynecol.* 9, 195–198. doi: 10.1016/S1083-3188(96)70030-X

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41, 95–98.

Hauptman, N., Boštjancic, E., Zlajpah, M., Rankovic, B., and Zidar, N. (2018). Bioinformatics analysis reveals most prominent gene candidates to distinguish colorectal adenoma from adenocarcinoma. *BioMed. Res. Intern.* 2018, article 9416515, 10. doi: 10.1155/2018/9416515

Jayes, F. L., Ma, X., Flannery, E. M., Moutos, F. T., Guilak, F., and Leppert P. C. (2013) Biomechanical evaluation of human uterine fibroids after expo-sure to purified clostridial collagenase. In: *Supplement to Biology of Reproduction for the 46th Annual Meeting of the Society for the Study of Reproduction*; 2013 July 22–26; Montreal, Canada.

Kirkland, S. C. (2009). Type I collagen inhibits differentiation and promotes a stem cell-like phenotype in human colorectal carcinoma cells. *BJC.* 101 (2), 320–326. doi: 10.1038/sj.bjc.6605143

Krasny, L., Shimony, N., and Tzukert, K. (2010). An in-vitro tumour microenvironment model using adhesion to type I collagen reveals Akt-dependent radiation resistance in renal cancer cells. *Neph. Dial. Transpl.* 25 (2), 373–380. doi: 10.1093/ndt/gfp525

Kumar, S., Stecher, G., and Tamura, K. (2016). Molecular evolution genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874.

Lei, S. F., Denga, F. Y., Xiaoa, S. M., Chena, X. D., and Deng, H. W. (2005). Association and haplotype analyses of the *COL1A2* and ER-a gene polymorphisms with bone size and height in Chinese. *Bone* 36, 533–541. doi: 10.1016/j.bone.2004.11.002

Leppert, P. C., Baginski, T., Prupas, C., Catherino, W. H., Pletcher, S., and Segars, J. H. (2004). Comparative ultrastructure of collagen fibrils in uterine leiomyomas and normal myometrium. *Fertil. Steril.* 82 (3), 1182–1187. doi: 10.1016/j.fertnstert.2004.04.030

Librado, P., and Rozas, J. (2009). Dna SP v5: A software for comprehensive analysis of DNA polymrphism data. *Bioinformatics* 25, 1451–1452.

Parazzini, F., Negri, E., La Vecchia, C., Fedele, L., Rabaiotti, M., and Luchini, L. (1992). Oral contraceptive use and risk of uterine fibroids. *Obstet. Gynecol.* 79, 430–433. doi: 10.1097/00006250-199203000-00021

Rogers, R., Norian, J., Malik, M., Abu-Asab, M., Christman, G., Malik, M., et al., et al. (2008). Mechanical homeostasis is altered in uterine leiomyoma. *Am. J. Obstet. Gynecol.* 198 (4), 1–474. doi: 10.1016/j.ajog.2007.11.057

Ross, R. K., Pike, M. C., Vessey, M. P., Bull, D., Yeates, D., and Casagrande, J. T. (1986). Risk factors for uterine fibroids: reduced risk associated with oral contraceptives. *Br. Med. J.* 293, 359–362. doi: 10.1136/bmj.293.6543.359

Rossert, J., Terraz, C., and Dupont, S. (2000). Regulation of type I collagen genes expression. *Nephrol. Dial. Transplant.* 15, 66–68. doi: 10.1093/ndt/15.suppl_6.66

Sandberg, A. A. (2005). Updates on the cytogenetics and molecular genetics of bone and soft tissue tumors: leiomyoma. *Cancer Genet. Cytogenet.* 158, 1–26. doi: 10.1016/j.cancergencyto.2004.08.025

Stewart, E. A., Friedman, A. J., Peck, K., and Nowak, R. A. (1994). Relative overexpression of collagen type I and collagen type III messenger ribonucleic acids by uterine leiomyomas during the proliferative phase of the menstrual cycle. *J. Clin. Endocrin. Metabol.* 79 (3), 900–906. doi: 10.1210/jcem.79.3.8077380

Tenni, R., Cetta, G., Dyne, K., Rossi, A., Quacci, D., Lenzi, L., et al. (1988). Type I procollagen in the severe non-lethal form of osteogenesis imperfecta. Defective pro-α1(I) chains in a patient with abnormal proteoglycan metabolism and mineral deposits in the dermis. *Hum. Genet.* 79, 245–250. doi: 10.1007/BF00366245

Thiaw, M. (2018). Diversité génétique de la population Sénégalaise: comparaison entre groupes ethniques. Mémoire de Diplôme de Master en Biologie Animale (mémoire non publié). Dakar, Sénégal: Faculté des Sciences et Techniques, Université Cheikh Anta Diop.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22 (22), 4673–4680. doi: 10.1093/nar/22.22.4673

Trojonowska, M. (2002). Molecular aspects of scleroderma. *Front. Biosci.* 7, 608–618. doi: 10.2741/A798

Yasul, W., Oue, N., and Ito, R. (2004). Search for new biomarkers of gastric cancer through serial analysis of gene expression and its clinical implications. *Cancer Sci.* 95, 385–392. doi: 10.1111/j.1349-7006.2004.tb03220.x

Velebil, P., Wingo, P. A., Xia, Z., Wilcox, L. S., and Peterson, H. B. (1995). Rate of hospitalization for gynecologic disorders among reproductive-age women in the United States. *Obstet. Gynecol.* 86, 764–769. doi: 10.1016/0029-7844(95)00252-M

Wise, L. A., and Laughlin-Tommaso, S. K. (2016). Epidemiology of uterine fibroids from menarche to menopause. *Clin. Obstet. Gynecol.* 59 (1), 2–24. doi: 10.1097/GRF.0000000000000164

Wise, L. A., Ruiz-Narvaez, E. A., Palmer, J. R., Cozier, Y. C., Tandon, A., Patterson, N., et al. (2012). African ancestry and genetic risk for uterine leiomyomata. *Am. J. Epidemiol.* 176, 1159–1168. doi: 10.1093/aje/kws276

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 October 2019
doi: 10.3389/fgene.2019.00906

A Novel Prognostic Signature of Transcription Factors for the Prediction in Patients With GBM

Quan Cheng[1†], Chunhai Huang[2†], Hui Cao[3], Jinhu Lin[1], Xuan Gong[1], Jian Li[1], Yuanbing Chen[1], Zhi Tian[2], Zhenyu Fang[1] and Jun Huang[1]*

[1] Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, China, [2] Department of Neurosurgery, First Affiliated Hospital of Jishou University, Jishou, China, [3] Clinical Medical Research Center of Hunan Provincial Mental Behavioral Disorder, Clinical Medical School of Hunan University of Chinese Medicine, Hunan Provincial Brain Hospital, Changsha, China

OPEN ACCESS

Edited by:
Barbara Karen Dunn,
National Institutes of Health (NIH),
United States

Reviewed by:
Howard Donninger,
University of Louisville,
United States
Zeyi Liu,
Soochow University, China
Yadi Zhou,
Cleveland Clinic, United States

***Correspondence:**
Jun Huang
xyyyhj@csu.edu.cn

†These authors share authorship

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: 01 February 2019
Accepted: 27 August 2019
Published: 01 October 2019

Citation:
Cheng Q, Huang C, Cao H, Lin J, Gong X, Li J, Chen Y, Tian Z, Fang Z and Huang J (2019) A Novel Prognostic Signature of Transcription Factors for the Prediction in Patients With GBM. Front. Genet. 10:906. doi: 10.3389/fgene.2019.00906

Background: Although the diagnosis and treatment of glioblastoma (GBM) is significantly improved with recent progresses, there is still a large heterogeneity in therapeutic effects and overall survival. The aim of this study is to analyze gene expressions of transcription factors (TFs) in GBM so as to discover new tumor markers.

Methods: Differentially expressed TFs are identified by data mining using public databases. The GBM transcriptome profile is downloaded from The Cancer Genome Atlas (TCGA). The nonnegative matrix factorization (NMF) method is used to cluster the differentially expressed genes to discover hub genes and signal pathways. The TFs affecting the prognosis of GBM are screened by univariate and multivariate COX regression analysis, and the receiver operating characteristic (ROC) curve is determined. The GBM hazard model and nomogram map are constructed by integrating the clinical data. Finally, the TFs involving potential signaling pathways in GBM are screened by Gene Set Enrichment Analysis (GSEA), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis.

Results: There are 68 differentially expressed TFs in GBM, of which 43 genes are upregulated and 25 genes are downregulated. NMF clustering analysis suggested that GBM patients are divided into three groups: Clusters A, B, and C. LHX2, MEOX2, SNAI2, and ZNF22 are identified from the above differential genes by univariate/multivariate regression analysis. The risk score of those four genes are calculated based on the beta coefficient of each gene, and we found that the predictive ability of the risk score gradually increased with the prolonged predicted termination time by time-dependent ROC curve analysis. The nomogram results have showed that the integration of risk score, age, gender, chemotherapy, radiotherapy, and 1p/19q can further improve predictive ability towards the survival of GBM. The pathways in cancer, phosphoinositide 3-kinases (PI3K)–Akt signaling, Hippo signaling, and proteoglycans, are highly enriched in high-risk groups by GSEA. These genes are mainly involved in cell migration, cell adhesion, epithelial–mesenchymal transition (EMT), cell cycle, and other signaling pathways by GO and KEGG analysis.

Conclusion: The four-factor combined scoring model of LHX2, MEOX2, SNAI2, and ZNF22 can precisely predict the prognosis of patients with GBM.

Keywords: glioblastoma, transcription factors, prognostic signature, LHX2, MEOX2, SNAI2, ZNF22

INTRODUCTION

Glioblastomas (GBMs) are the most common malignant tumors in the central nervous system (CNS), which accounts for 14.9% of primary CNS and 47.1% of primary brain tumors. The incidence of GBM increases with age, being most common during 75–84 years of age. It is generally associated with a poor prognosis, in which median overall survival (OS) is 15 months and 5-year survival is only about 5.5% (Ostrom et al., 2017). However, studies have shown that the prognosis varies widely among individuals. The histopathology which is commonly used in the clinic is not an ideal prognosis marker and can even lead to erroneous judgement. For the past 10 years, rapid advancement in bioinformatics has provided better tools to explore the molecular characteristics of cancer. This way, many molecular markers and molecular characterizing systems of GBM have been identified, which offers novel insights into the better understanding of progression mechanisms, diagnosis, and treatment of GBM (Lee et al., 2018). For instance, the prognostic and predictive significance of isocitrate dehydrogenase (IDH)1/2 mutation has been validated by many studies. In these studies, GBM patients with IDH1/2 mutations have notably longer OS compared with patients without (Yan et al., 2009; Beiko et al., 2014). In addition, O^6-methylguanine DNA methyltransferase (MGMT) methylation status is another important molecular marker, predicting the therapeutic effects of temozolomide (TMZ) in GBM patients (Hegi et al., 2005).

Transcription factor (TF), also known as trans-acting factor, is a protein with a unique structure that controls the rate of transcription or the production of messenger RNA (mRNA). TF can act as an activator or repressor by interacting with cis-acting elements. During eukaryotic transcription initiation, RNA polymerase II binds to TFs to form a transcription initiation complex. Transcription is a very complex process which is operated by synchronized multi-protein complexes including TFs. According to the functional characteristics of the TFs, they can be divided into two types; the first type is general TFs such as TFII family proteins which are ubiquitous and bring the RNA polymerase through binding to the promoter region near the transcription start site to turn on genes (Kadonaga, 2004). The second type is sequence-specific TFs that bind upstream of the transcription start site to promote or inhibit the expression of a particular gene. The sequence-specific TFs contain one or more DNA-binding domains and recognize specific DNA motifs near the gene to initiate their functions. TFs are involved in different biological processes such as cell proliferation, growth, differentiation, and apoptosis. Dysfunction of TFs can lead to imbalance in homeostasis, leading to a variety of diseases. Due to the complexity of transcriptional regulation, there are not many systematic studies on transcriptional regulation of GBM. This study mainly focuses on changes of transcriptome profiling in GBM, with the intention to discover key regulatory molecules which can be developed as new markers.

In this study, we have identified, established, and evaluated a scoring system with a combination of four TFs (LHX2, MEOX2, SNAI2, and ZNF22) to assess the prognosis of GBM. To achieve this, we have integrated the analysis of GBM patients' expression profiles or sequencing data from Oncomine, Gene Expression Omnibus (GEO), TCGA, and Chinese Glioma Genome Atlas (CGGA) databases. We also provide an evidence that the expression levels of SNAI and MEOX2 are significantly associated with histopathological grade and survival time in glioma patients, indicating that these two transcriptional factors play a crucial role in the malignancy of glioma.

MATERIALS AND METHODS

Identification of the Differentially Expressed TFs

Gene expression profile data of the SUN brain, Murat brain, GBM, and normal brain tissue in TCGA were obtained from the Oncomine (https://www.oncomine.org/resource/) database. The statistically significant differentially expressed TFs (DETFs) were identified with a fold change larger than 2. The candidate cell-specific TF markers per tissue were derived from the molecular signature database [http://software.broadinstitute.org/gsea/msigdb/gene_families.jsp, Molecular Signatures Database (MSigDB) V6.0]. The overlapped upregulated or downregulated TFs of four groups were defined as the most widely and significantly DETFs.

Datasets

The genome-wide mRNA array expression profile of GBM patients and their corresponding clinical information, including histology, gender, age, survival information and IDH1 gene mutation status, 1p/19q codelet, GeneExp subtype, and others, were downloaded from TCGA (https://xenabrowser.net) (Goldman et al., 2019). These clinical features and mRNA expression profile of TCGA GBM array are utilized as the training dataset which includes 524 patient samples. As for the validation dataset, there are 60 samples from GSE74187, 215 samples from the CGGA GBM RNA-Seq dataset, and 157 samples from the TCGA GBM-seq dataset, which are an independent human glioma gene expression profile. The CGGA GBM RNA-Seq dataset is downloaded from the CGGA (http://cgga.org.cn/index.jsp). The GBM

mRNA-seq dataset was also gained from TCGA (https://xenabrowser.net) (Goldman et al., 2019).

Risk Model Establishment Analysis of the Detfs and Prognosis Survival of GBM

We employed the nonnegative matrix factorization (NMF) method to find the key genes and signal pathways by clustering the DETFs, which were identified in a previous step. The gene function and pathway annotation were performed using the clusterProfiler package in R (Yu et al., 2012). Univariate Cox hazard analysis was used to identify individual single genes that affect the survival of TCGA GBM patients. And then multivariate Cox regression analysis was used to establish a linear joint risk score of gene expression level (expr) using regression coefficient β. The risk score for each sample was calculated as follows: risk score = $\text{expr}_{\text{gene1}} \times \beta_{\text{gene1}} + \text{expr}_{\text{gene2}} \times \beta_{\text{gene2}} + \cdots + \text{expr}_{\text{gene}n} \times \beta_{\text{gene}n}$. The area under the receiver operating characteristic (ROC) curve (AUC) of the time-dependent risk score was calculated using the survivalROC package of R. The samples were then divided into high- and low-risk groups based on the median or the best cutoff of risk scores, for survival analysis. Next, we randomly selected half of the samples from TCGA GBM array training set to validate the efficacy of our model. After that, we conduct the external validation with the GSE74187 dataset, the CGGA dataset, and TCGA GBM RNA-Seq dataset. The correlation analysis between high- and low-risk groups towards clinical features was performed in the training set. The multivariate Cox model was constructed using the survival package for the risk score and clinical features with a P value < 0.05 as cutoff, and the Nomogram chart was drawn using the regplot package. The risk model was assessed by the calibration curve and AUC.

Gene Set Enrichment Analysis

The GSEA was performed *via* the clusterProfiler package of R. The GBM samples in TCGA were divided into downregulated and upregulated groups based on the median of the risk score of the TFs. The absolute value of normalized enrichment score (NES) > 1, P value < 0.05, and false discovery rate (FDR) q value < 0.25 were defined as the statistically significant criteria. The co-expressed genes of the prognostic-related TFs identified in TCGA dataset were identified (|Spearman's r| ≥ 0.4). The genes were then subjected to the clusterProfiler package for GO (biological process) and KEGG enrichment analysis, with P < 0.05 as the cutoff.

Statistical Analysis

All statistical analyses were performed using SPSS 22.0 or R software. Two groups' statistical significance was calculated using the t-test or non-parametric t-test. The chi-square test was used to analyze the correlation of the classified data. In this study, P < 0.05 was defined as a statistically significant cutoff. For the Cox regression analysis, the time-dependent Cox model variable test was verified using the proportional hazard hypothesis (PH hypothesis).

RESULTS

Identification of the DETFs

A total of 68 significantly DETFs were identified from TCGA/SUN brain/Murat brain database, of which 43 were upregulated and 25 were downregulated (**Table 1** and **Figures 1A, B**). Furthermore, we have obtained the gene expression profile matrix of TCGA GBM patients and have found that GBM patients can be divided into three categories using the NMF clustering method (**Figure 1C**). Representative genes of each group are shown in **Table 1**. Among them, the proneural patients in the Cluster A group were the most (56%) accounted for. Mesenchymal and classical patients were mostly in the Cluster C group, accounting for 44% and 48%, respectively. The proportions of three subtypes in the

TABLE 1 | Differentially expressed transcription factors (TFs).

GBM *vs* normal braint		Representative genes		
Up	**Down**	**Cluster A**	**Cluster B**	**Cluster C**
ASCL1	ARNT2	ARNT2	CBX6	HIF1A
BAZ1A	BCL11A	ASCL1	CBX7	MEF2A
CBX3	CBX6	ETV1	CHD5	MEOX2
ETV1	CBX7	HEY1	FEZF2	PDLIM5
EZH2	CHD5	LHX2	HIVEP2	PRRX1
FOXM1	FEZF2	LIMA1	HLF	RELA
HEY1	HIVEP2	RNF41	LDB2	RUNX1
HIF1A	HLF	SOX11	LDOC1	SHOX2
HMGB2	LDB2	SOX2	LMO3	SMAD1
HOXA10	LDOC1	TRIM24	MEF2C	SNAI2
HOXA5	LHX2	ZNF207	MYT1L	SNAPC1
HOXA7	LMO3	ZNF22	OPTN	TBX2
HOXB2	MED14	BAZ1A	PRDM2	TGFB1I1
HOXC10	MEF2A	BCL11A	RIMS3	TGIF1
HOXC6	MEF2C	CBX3	RUNX1T1	ZNF217
ILF3	MYT1L	EZH2	STON1	
LIMA1	NFYB	FOXM1	ULK2	
MBD2	OPTN	HMGB2	ZMYND11	
MEOX2	PRDM2	ILF3		
PDLIM5	PSIP1	MBD2		
PRRX1	RIMS3	MED14		
RARA	RNF41	NFYB		
RELA	RUNX1T1	PSIP1		
RUNX1	ULK2	RARA		
SHOX2	ZMYND11	SOX4		
SMAD1		TCF3		
SNAI2		TFAP2A		
SNAPC1		WHSC1		
SOX11		HOXA10		
SOX2		HOXA5		
SOX4		HOXA7		
STON1		HOXB2		
TBX2		HOXC10		
TCF3		HOXC6		
TFAP2A				
TGFB1I1				
TGIF1				
TRIM24				
WHSC1				
ZFAND6				
ZNF207				
ZNF217				
ZNF22				

GBM, glioblastoma.

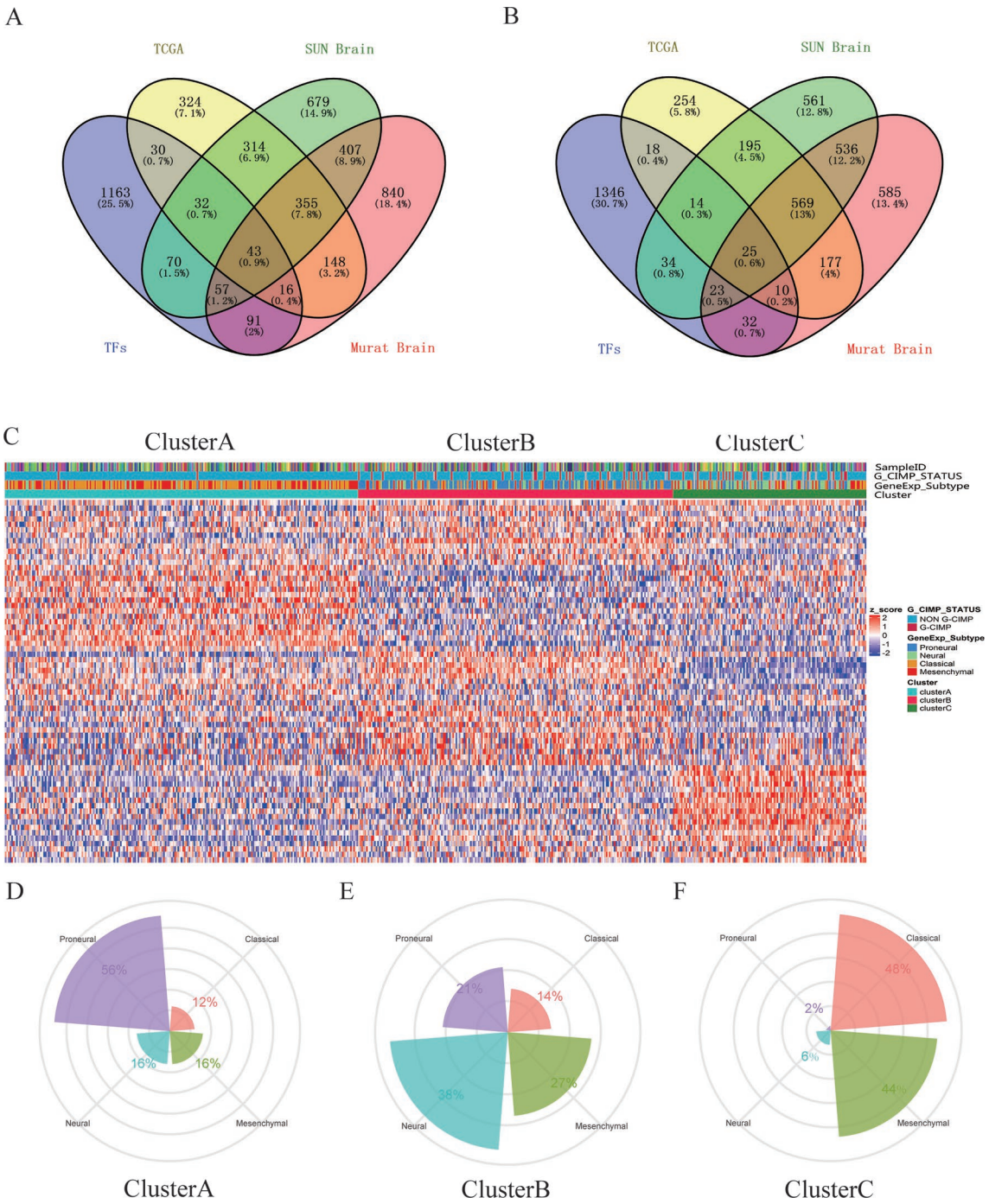


FIGURE 1 | Identification of differentially expressed transcription factors (DETFs). **(A)** total of 43 significantly upregulated transcription factors were screened from the three databases of The Cancer Genome Atlas (TCGA)/SUN brain/Murat brain. **(B)** A total of 25 significantly downregulated transcription factors were screened from the three databases of TCGA/SUN brain/Murat brain. **(C)** Clusters A–C of glioblastoma (GBM) patients through 68 transcription factors using the nonnegative matrix factorization (NMF) clustering method. **(D–F)** Proportions of proneural, mesenchymal, classical, and neural in Clusters A–C.

Cluster B group were very close, and the neural type accounted for a large proportion (38%) (**Figures 1D, E, F**). Correlation analysis between the 68 identified TFs has revealed that the genetic correlation among three clusters was quite good (**Figure 2A**). Patients in the Cluster A group had the best prognosis, with a median OS of 493 days. Patients in the Cluster B group had a median OS of 457 days; while patients in the Cluster C group had the worst prognosis with a median OS of 419 days (**Figure 2B**).The gene function and pathway annotation analysis by the clusterProfiler package have revealed that the most

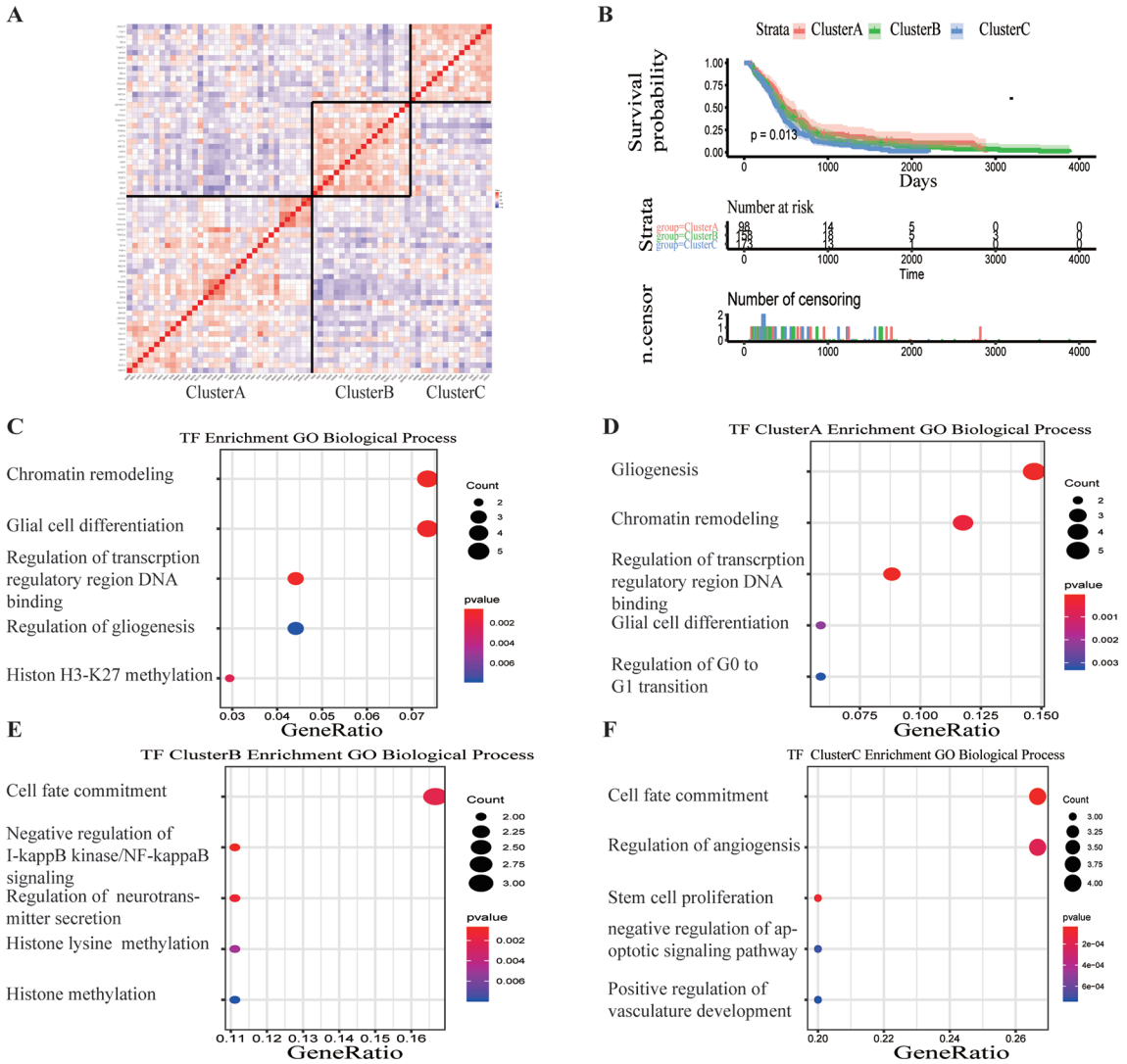


FIGURE 2 | Survival analysis and gene function enrichment of Clusters A–C. **(A)** Gene expression correlation of Clusters A–C in The Cancer Genome Atlas (TCGA) glioblastoma (GBM) data. **(B)** Survival analysis of the three groups, Clusters A–C: the patients in Cluster A had the best prognosis, while those in Cluster C had the worst prognosis. **(C)** Gene Ontology (GO) (biological process) enrichment results of 68 transcription factors. **(D–F)** GO (biological process) enrichment results of Clusters A–C.

enriched pathways of the 68 TFs were chromatin remodeling, glial cell differentiation, regulation of transcription regulatory region DNA binding, and regulation of gliogenesis (**Figure 2C**). The most enriched pathways of Cluster A were gliogenesis, chromatin remodeling, regulation of transcription regulatory region DNA binding, glial cell proliferation, and regulation of G0-to-G1 transition (**Figure 2D**). The most enriched pathways of Cluster B were cell fate commitment, negative regulation of 1-kappa B kinase/nuclear factor (NF)-kappa B signaling, histone lysine methylation, and histone methylation (**Figure 2E**). The most enriched pathways of Cluster C were cell fate commitment, regulation of angiogenesis, stem cell proliferation, negative regulation of apoptotic pathway, and positive regulation of vasculature development (**Figure 2F**).

Construction of Prognostic Classifier From the Training Sets and Validation

The GBM expression profile of TCGA was used as a train dataset to screen the DETFs. Univariate Cox hazard analysis was used to identify individual single genes from 68 TFs that affect the survival of TCGA GBM patients, in which we obtained 12 statistically significant genes: ASCL1, HOXB, HOXC1, LHX2, MEOX2, RARA, RUNX1, SNAI2, SOX4, TCF3, TGIF1, and ZNF22. The 12 TFs were entered into the multivariate regression analysis. The four TFs (LHX2, MEOX2, SNAI2, and ZNF22) were inputted to the final equation, and the results indicated that these four TFs can be used as independent predictors for the prognosis of GBM. The β-cofactors of LHX2, MEOX2, SNAI2, and ZNF22 were 0.318, 0.264, 0.332, and -0.349, respectively.

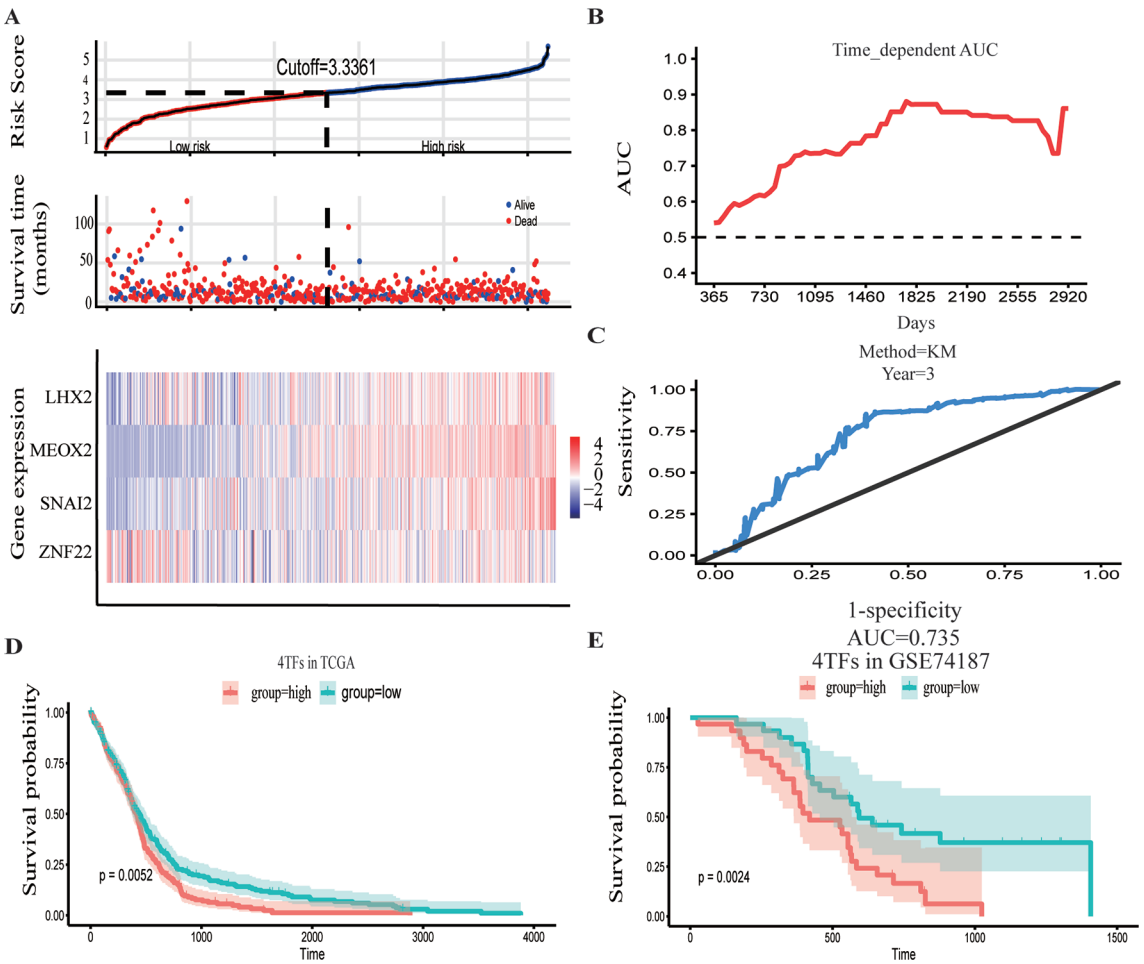


FIGURE 3 | Construction and verification of the hazard assessment system. **(A)** The distribution of risk score, patient survival time and status in The Cancer Genome Atlas (TCGA) set, and heatmap of the gene risk assessment model in TCGA dataset. **(B, C)** The area under the curve (AUC) for the risk assessment model in TCGA set and time-dependent receiver operating characteristic (ROC) for predicting the 3-year survival. **(D, E)** Kaplan–Meier curves of the high-risk group and low-risk group of TCGA dataset and GSE74187 dataset.

The joint risk score of the four TFs was calculated by substituting the coefficient into the formula. The median value was 3.3361 by ranking the risk score from low to high, which was used to divide the samples into low- and high-risk groups (**Figure 3A**). Through time-dependent ROC curve analysis, it was found that the predictive ability of the joint risk score of the four TFs for the patients' survival prognosis gradually increased with the predicted termination time (**Figure 3B**), and the AUC of the risk score ROC curve at the predicted termination time of 3 years was 0.735 (**Figure 3C**). GBM patients were divided into high- and low-risk groups by the median value of the risk score, and the results showed that the OS time between the low- and high-risk groups was very significant ($P = 0.0052$) (**Figure 3D**). While the results of twice internal validations and the ROC curve are satisfied (**Figures S1A–D**), to validate the risk model with the external dataset, the GSE74187 dataset, the CGGA dataset, and TCGA dataset, we used the β-cooperative coefficient to calculate the joint risk score of the four TFs in each dataset that will predict the prognosis of GBM patients. With these taken together, these results manifested that the OS of GBM patients in the high- and low-risk groups was significantly different (GSE74187 $P = 0.0024$, the CGGA dataset $P < 0.0001$, and TCGA dataset $P = 0.0055$). The ROC curve also corresponds with our expectation (**Figures 3E** and **S1E–H**).

Prognostic Value of the Integrated Classifier Is Independent of the Clinical Feature

To assess whether the prognostic classifier was an independent indicator in GBM patients, we analyzed the effect of each clinicopathological feature towards survival by using the Cox regression model. The multivariate regression analysis, the risk score based on TFs, age, gender, chemotherapy, radiotherapy, and 1p/19q codelet were entered into the final equation of the Cox regression model (**Table 2**). We found that the risk score based on TFs was strong and an independent predictive factor in the GBM data of TCGA (**Table 2**). Next,

TABLE 2 | Univariate and multivariate Cox regression analysis of factors affecting overall survival of patients in The Cancer Genome Atlas (TCGA) glioblastoma (GBM) cohort.

	Univariate analysis			Multivariate analysis		
	P	HR	95%CI	*P*	HR	95%CI
Risk score	<0.001	1.37	1.20–1.57	0.005	1.23	1.06–1.43
Age group (> 45)	<0.001	2.291	1.632–3.216	<0.001	2.01	1.39–2.91
Gender (Female)	0.094	0.810	0.634–1.036	0.001	0.64	0.50–0.84
Subtype						
Proneural	0.118	0.769	0.552–1.069			
Mesenchymal	0.267	1.193	0.874–1.629			
Neural	0.588	1.101	0.778–1.558			
Chemotherapy (Yes)	<0.001	0.378	0.283–0.505	<0.001	0.48	0.33–0.69
Radiotherapy (Yes)	<0.001	0.131	0.094–0.183	<0.001	0.19	0.130–0.28
IDH status (WT)	<0.001	0.321	0.196–0.524			
1p/19q codelet (non-codel)	0.046	4.24	1.026–17.52	0.005	8.54	1.89–38.5

HR, hazard ratio; IDH, isocitrate dehydrogenase; WT, wild type.

we constructed a nomogram that integrated TF classifiers and clinicopathological features to predict the 1-year and 3-year survival of GBM patients (**Figure 4A**). The calibration curve showed that the predicted 1-year and 3-year survival rates were closely related to the actual observed ratio (**Figure 4B**). GBM patients were divided into high- and low-risk groups by the median value of the new classifier based on the TF risk score and the clinical features. The outcome of this analysis shows that the OS of GBM patients in the high- and low-risk groups was significantly different ($P < 0.0001$) (**Figure 4C**). By calculating the AUC of the new classifier, we found that the AUC value was 0.819 at the predicted 3-year end time and 0.734 at 1-year end time (**Figure 4D**), which was higher than that using the TF classifier alone. By calculating ROC values of different times, the ROC value of the new classifier was significantly higher than that using the TF classifier only (**Figure 4E**). These results demonstrated the robust and predictive power of the new classifier based on the TF risk score and the clinical features performed better.

Functional Analysis for the Prognostic Classifier of Genes

To identify the potential functional mechanisms that led to different prognosis in high- and low-risk groups, we applied functional enrichment analysis (GSEA) on identified TFs. The top 30 pathways were shown in **Figure 5A** where $|NES| > 1$, P value < 0.05, and FDR q value < 0.25 were used as the cutoff for identifying differentially enriched signal pathways. Pathways in cancer such as the phosphoinositide 3-kinases (PI3K)–Akt signaling pathway, hippo signaling pathway, proteoglycans in cancer, and other signaling pathways (**Figures 5B–E**) were significantly enriched in high-risk groups, which may partly explain the reason for poor prognosis in high-risk group patients. The co-expressed genes of LHX2 were mainly involved in pathways such as glial cell differentiation and cell adhesion. The co-expressed genes of MEOX2 were mainly related to the glial cell differentiation, extracellular matrix composition, cell adhesion, PI3K–Akt pathway, and other functional and pathways. The co-expressed genes of SNAI2 were mainly involved in extracellular matrix, cell invasion, cell adhesion, PI3K–Akt pathway, and NF-kappa B pathway. The co-expressed genes of SNAI2 were mainly involved in mRNA processing, histone modification, chromosome segregation, cell cycle, and Notch signaling pathway.

DISCUSSION

Glioma is the most common type of tumor in the brain, and its OS is still not satisfactory. In particular, the GBM patients with high-grade malignancy still have a high mortality rate (Ostrom et al., 2017). New studies are focusing on better classification, prognosis prediction, molecular mechanism, and targeted drug therapy for GBM (Touat et al., 2017). TFs play an important role in turning genes "on" and "off," yet there are few systematic studies focusing on their roles in gliomas. By analyzing the DETFs in GBM using TCGA, SUN brain, and Murat brain datasets, we identified 68 TFs that were differentially expressed in GBM patients compared to the normal brain tissues. Using TCGA dataset as a training dataset, we found that GBM patients can be divided into three distinct subpopulations based on 68 TFs. It is well known that there is a significant heterogeneity within the malignant tumor, which leads to a large difference in its prognosis and response to various treatments. From the perspective of TFs' expression profile, we elucidated the intrinsic differences in GBM patients, which indicated the underlining mechanisms of tumor development in different subtypes of GBM that are regulated by different signaling pathways. Our analysis showed that gliogenesis, chromatin remodeling, regulation of transcription regulatory region DNA binding, glial cell proliferation, and regulation of G0-to-G1 transition may play a major role in cancer progression in Cluster A. Cell fate commitment, negative regulation of 1-kappa B kinase/NF-kappa B signaling, histone lysine methylation, histone methylation, and so on were mainly involved in Cluster B. In the subtype of Cluster C, cell fate commitment, regulation of angiogenesis, stem cell proliferation, negative regulation of the apoptotic pathway, positive regulation

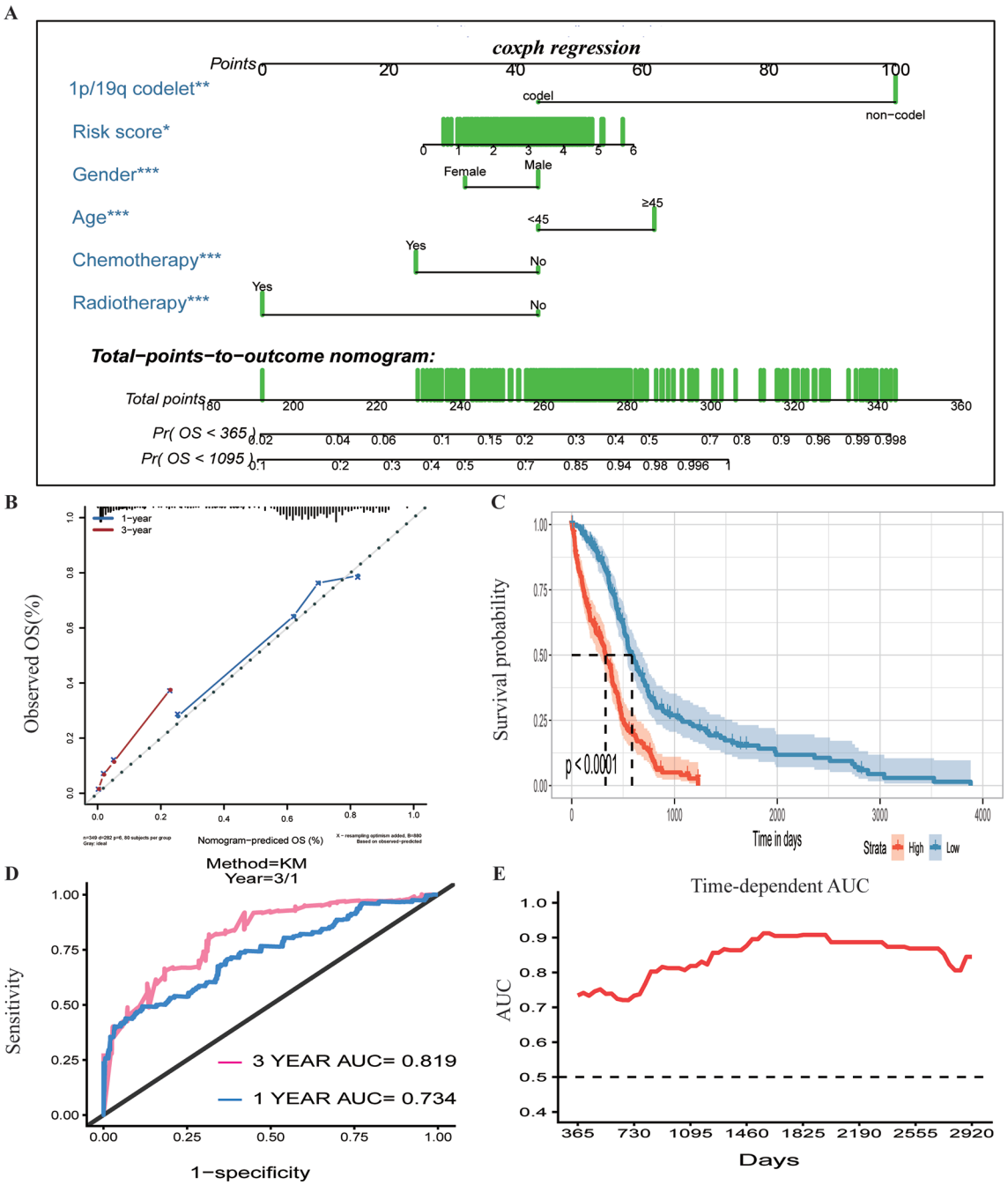


FIGURE 4 | Prognostic value of the integrated classifier is independent of clinical feature. **(A)** Prognostic nomogram for glioblastoma (GBM) patients with six chief characteristics. **(B)** The calibration curve of overall survival (OS) at 1/3 year. Nomogram-predicted probability of the OS is plotted on the *x*-axis, and the observed OS is plotted on the *y*-axis. **(C)** Comparison of OS between high-risk-score group and low-risk-score group. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **(D, E)** The time-dependent receiver operating characteristic (ROC) for predicting the 1/3-year survival and area under the curve (AUC) for the risk assessment model in The Cancer Genome Atlas (TCGA) set.

of vasculature development, and so on played a more critical role. These different mechanisms of tumor progression of GBM can also explain the complex heterogeneity and differences in the prognosis. According to the subtypes of Verhaak et al. (2010), Cluster A contained more proneural subtypes, and its prognosis was better. The proportions of subtypes in Cluster B were roughly the same, while Cluster C had the most mesenchymal and classical subtypes, which may cause poor prognosis.

In order to find out which of these factors plays a key role in the prognosis of GBM, we used the Cox hazard ratio model to analyze and finally determined four independent factors (LHX2, MEOX2, SNAI2, and ZNF22) as predictors of GBM prognosis.

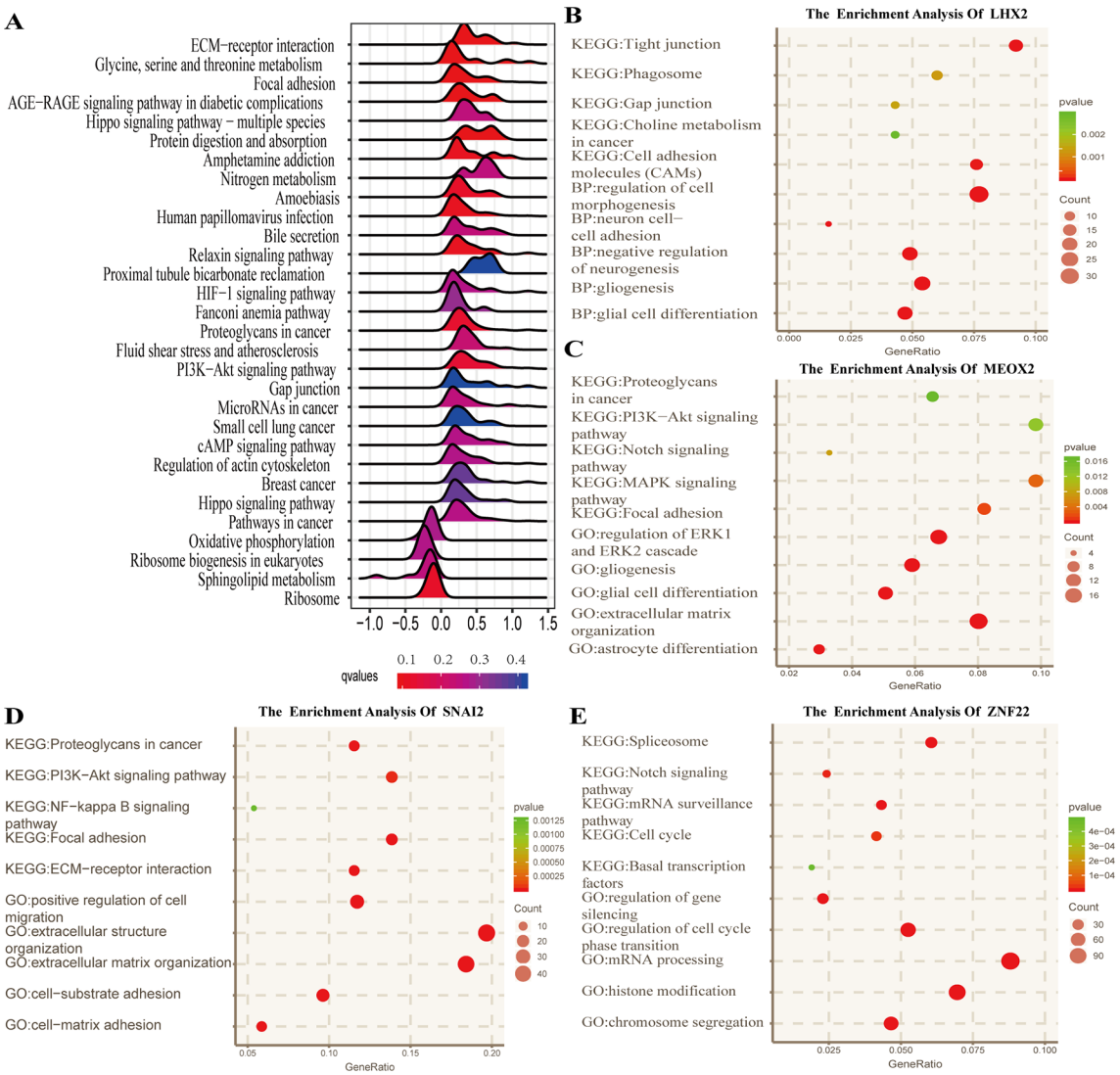


FIGURE 5 | Functional analysis for the prognostic classifier of genes. **(A)** Gene Set Enrichment Analysis (GSEA) based on risk score of transcription factors is performed to identify associated pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets. **(B–E)** Gene Ontology (GO) (biological process) terms and KEGG pathway related to co-expressed genes of LHX2, MEOX2, SNAI2, and ZNF22 in The Cancer Genome Atlas (TCGA) dataset.

Time-dependent ROC analysis and survival analysis found that the joint risk score based on the four TFs can accurately predict the survival prognosis. LHX2 is the major "cortical selection gene" in the cerebral cortex and plays multiple roles in different organs including the development of CNS (Chou and Tole, 2018). The relationship between LHX2 and tumor development has not yet been identified. It has been recently found that miR-124 can inhibit the migration and invasiveness of lung cancer cells by inhibiting LHX2 expression (Yang et al., 2017). Zakrzewski et al. (2015) discovered that LHX2 expressed differentially in different regions that were associated with disease progression in the underlying fibroma astrocytoma by bioinformatics, and studies have shown that this factor may play an important regulatory role in the development of tumors. The mesenchymal homeobox (MEOX) family includes two homeodomain protein+s, MEOX1 and MEOX2, with 95% sequence identity in the homologous domain, which are required for proper bone and muscle development in mouse embryos. MEOX2 is also known as a growth arrest 65-specific homeobox protein (Gax) (Northcott et al., 2017). Abnormal gene expression of MEOX2 has been found in a variety of diseases, including hepatic portal hypertension, Alzheimer disease, and cancer (Wu et al., 2005; Zeng et al., 2006). Additionally, in these diseases, MEOX2 has also been found to be associated with vascular dysfunction. MEOX2 inhibits cell proliferation and epithelial–mesenchymal transition (EMT) of vascular smooth muscle and endothelial cells (Valcourt et al., 2007). Tachon et al. (2019) demonstrated that MEOX2 expression was associated with IDH1/2 wild-type molecular subtype and was significantly correlated with the OS of all gliomas, especially in lower-grade gliomas. The Snail family

of zinc finger transcriptional repressors includes three members: snai1/snail, snai2/slug, and snai3/smuc, which play key roles in EMT (Nieto, 2002; Strobl-Mazzulla and Bronner, 2012; Liu et al., 2014; Liu et al., 2017). It has been found that mRNA expression of SNAI2 was associated with histological grade and invasive phenotype in primary human glioma specimens and can be induced by epidermal growth factor receptor (EGFR) activation in human GBM cells. The overexpression of SNAI2/Slug increased the proliferation and invasion of GBM cells *in vitro* and promoted angiogenesis and tumor growth *in vivo*. Importantly, knockdown of endogenous SNAI2/Slug in GBM cells reduced invasion and increased survival in the mouse intracranial human GBM xenograft model (Yang et al., 2010). Liao et al. (2015) found that miR-203 can target SNAI2 to inhibit EMT and promote drug sensitivity and implied that targeting SNAI2 may be a potential therapeutic approach to overcome chemoresistance in GBM. In this study, we found that SNAI2 was overexpressed, and SNAI2 overexpression is characteristic for interstitial transformation, of Cluster C, proving the precision of cluster classification. ZNF22 is thought to be involved in the development of teeth (Gao et al., 2003), and its role in tumors has not been studied thoroughly. In this study, GO and KEGG analysis of DETFs revealed that these genes were mainly enriched in signal pathways such as cell migration, cell adhesion, EMT, and cell cycle, which are consistent with the studies mentioned above.

By dividing the GBM patients into high- and low-risk groups based on the four-factor joint risk score, we found that the signal pathways involved in different groups were quite different. Pathways in cancer, PI3K–Akt signaling pathway, hippo signaling pathway, and proteoglycans in cancer signaling pathways were mainly enriched in high-risk patients. These enriched malignant pathways can lead to significantly greater tumor proliferation and invasion in the high-risk group than in the low-risk group. PI3K is responsible for the conversion of PIP2 to PIP3, which activates the downstream target PKB/Akt (Chang et al., 2017; Fu et al., 2017). The PI3K pathway is usually activated by EGFR and other growth factor receptors (Zoncu et al., 2011). It was shown that the PI3K pathway was activated in almost all GBM, although only less than 15% of GBM showed activating mutations in the PI3K gene. The activation of the PI3K/Akt/mTOR pathway led to the development of GBM resistance, thereby inhibiting the therapeutic effect of chemotherapy (Li et al., 2016). The prognosis of GBM patients with activation of the PI3K–Akt pathway was terribly poor (Chakravarti et al., 2004). The hedgehog (Hh) signaling pathway, also known as hedgehog-patched (Hh-Ptch), hedgehog-Gli (Hh-Gli), or hedgehog-patched-smoothened (Hh-Ptch-Smo), is an evolutionarily conservative signaling pathway from the cell membrane to the nucleus (Skoda et al., 2018). Dysfunction or abnormal activation of the Hh signaling pathway is associated with developmental malformations and cancer, such as basal cell nevus syndrome (BCNS), sporadic basal cell carcinoma (BCC), medulloblastoma (MB), rhabdomyosarcoma, meningioma, and glioma (Taipale and Beachy, 2001; Xu et al., 2012; Skoda et al., 2018). Xu et al. (2010) found that CD44 promoted the resistance of glioma cells to reactive oxygen species-induced and cytotoxic agent-induced stress by attenuating the activation of the hippo signaling pathway. Lu et al. (2017) found that IKBKE regulated cell proliferation, invasion, and EMT of malignant glioma cells *in vitro* and *in vivo* by affecting the hippo pathway. Proteoglycans, including heparan sulfate and chondroitin sulfate proteoglycans (HSPG and CSPG, respectively), regulate the activity of many signaling pathways as well as cellular–microenvironment interactions (Nagarajan et al., 2018). Proteoglycans are the main component of the extracellular environment of the brain and regulate cell signaling and cell migration. The abnormality of proteoglycans and their modification enzymes in GBM leads to the changes of EGFR or PDGFRα signaling pathways (Wade et al., 2013). Proteoglycans are very critical for the mechanistic understanding of proteoglycan function in carcinogenic signaling and tumor microenvironment interactions in GBM and can be used to identify the new tumor biomarkers and druggable targets. These genes involved functions and pathways that are coincident with the results we found.

We analyzed the effect of each clinicopathological feature and TF risk model affecting survival by using the Cox regression model. In the multivariate regression analysis, the risk score based on TFs, age, gender, chemotherapy, radiotherapy, and 1p/19q codelet was entered into the final equation of the Cox regression model. The calibration curve of this model and AUC values indicate that the model has satisfactory accuracy. Next, we constructed a nomogram that integrated TF classifiers and clinicopathological features to predict the 1-year and 3-year survival of GBM patients. This nomogram can be used to guide doctors in judging the prognosis of GBM patients and to help them better communicate with patients.

In summary, the significantly DETFs in GBM that promote malignant progression of the tumors are mainly involved in the PI3K-Akt signaling pathway, hippo signaling pathway, proteoglycans in cancer, and other related signaling pathways. We believe that these pathways lead to poor prognosis and resistance to treatment in GBM. We have established a four-factor predictive joint risk score model that can be used to predict the prognosis of patients with GBM effectively. Based on this, two TFs closely related to the malignant progression of glioma are identified, which will provide a foundation to develop new biomarkers and targeted therapies in GBM.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://xenabrowser.net. Accession: GSE74187.

AUTHOR CONTRIBUTIONS

QC, CH, and JH conceived and designed the idea for the manuscript and wrote the paper. HC, JLin, JLi, ZF, and XG contributed the collection, analysis, and interpretation of data. ZT and YC provided analysis tools. All authors gave approval for

this version of the manuscript to be published and agree to be accountable for all aspects of the work.

FUNDING

This research was supported by the National Natural Science Foundation of China (no. 81703622 and no. 81560414), China Postdoctoral Science Foundation (no. 2018M633002), Hunan Provincial Natural Science Foundation of China (no. 2018JJ3838), and Hunan Provincial Health and Health Committee Foundation of China (C2019186), and Science and Technology Department of Hunan Province (NO.2015SK2032-2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00906/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | The result of the internal validation dataset 1 and 2, time-dependent receiver operating characteristic (ROC) **(A**,**C)**, kaplan-Meier curves of high-risk group and low-risk group **(B**,**D)** from the TCGA GBM array dataset. As for the external validation, time-dependent receiver operating characteristic (ROC) **(E**,**G)**, kaplan-Meier curves of high-risk group and low-risk group **(F**,**H)** of the CGGA dataset and the TCGA GBM RNA-seq dataset respectively.

REFERENCES

Beiko, J., Suki, D., Hess, K. R., Fox, B. D., Cheung, V., Cabral, M., et al. (2014). IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. *Neuro. Oncol.* 16, 81–91. doi: 10.1093/neuonc/not159

Chakravarti, A., Zhai, G., Suzuki, Y., Sarkesh, S., Black, P. M., Muzikansky, A., et al. (2004). The prognostic significance of phosphatidylinositol 3-kinase pathway activation in human gliomas. *J. Clin. Oncol.* 22, 1926–1933. doi: 10.1200/JCO.2004.07.193

Chang, H., Li, X., Cai, Q., Li, C., Tian, L., Chen, J., et al. (2017). The PI3K/Akt/mTOR pathway is involved in CVB3-induced autophagy of HeLa cells. *Int. J. Mol. Med.* 40, 182–192. doi: 10.3892/ijmm.2017.3008

Chou, S. J., and Tole, S. (2018). Lhx2, an evolutionarily conserved, multifunctional regulator of forebrain development. *Brain Res.* 1705, 1–14. doi: 10.1016/j.brainres.2018.02.046

Fu, Y. F., Liu, X., Gao, M., Zhang, Y. N., and Liu, J. (2017). Endoplasmic reticulum stress induces autophagy and apoptosis while inhibiting proliferation and drug resistance in multiple myeloma through the PI3K/Akt/mTOR signaling pathway. *Oncotarget* 8, 61093–61106. doi: 10.18632/oncotarget.17862

Gao, Y., Kobayashi, H., and Ganss, B. (2003). The human KROX-26/ZNF22 gene is expressed at sites of tooth formation and maps to the locus for permanent tooth agenesis (He-Zhao deficiency). *J. Dent. Res.* 82, 1002–1007. doi: 10.1177/154405910308201213

Goldman, M., Craft, B., Hastie, M., Repečka, K., Kamath, A., Mcdade, F., et al. (2019). The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv* doi: 10.1101/326470

Hegi, M. E., Diserens, A. C., Gorlia, T., Hamou, M. F., De Tribolet, N., Weller, M., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003. doi: 10.1056/NEJMoa043331

Kadonaga, J. T. (2004). Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116, 247–257. doi: 10.1016/S0092-8674(03)01078-X

Lee, E., Yong, R. L., Paddison, P., and Zhu, J. (2018). Comparison of glioblastoma (GBM) molecular classification methods. *Semin. Cancer Biol.* 53, 201–211. doi: 10.1016/j.semcancer.2018.07.006

Li, X., Wu, C., Chen, N., Gu, H., Yen, A., Cao, L., et al. (2016). PI3K/Akt/mTOR signaling pathway and targeted therapy for glioblastoma. *Oncotarget* 7, 33440–33450. doi: 10.18632/oncotarget.7961

Liao, H., Bai, Y., Qiu, S., Zheng, L., Huang, L., Liu, T., et al. (2015). MiR-203 downregulation is responsible for chemoresistance in human glioblastoma by promoting epithelial–mesenchymal transition *via* SNAI2. *Oncotarget* 6, 8914–8928. doi: 10.18632/oncotarget.3563

Liu, K., Tang, Z., Huang, A., Chen, P., Liu, P., Yang, J., et al. (2017). Glyceraldehyde-3-phosphate dehydrogenase promotes cancer growth and metastasis through upregulation of SNAIL expression. *Int. J. Oncol.* 50, 252–262. doi: 10.3892/ijo.2016.3774

Liu, S., Liao, G., Ding, J., Ye, K., Zhang, Y., Zeng, L., et al. (2014). Dysregulated expression of snail and E-cadherin correlates with gastrointestinal stromal tumor metastasis. *Eur. J. Cancer Prev.* 23, 329–335. doi: 10.1097/CEJ.0000000000000072

Lu, J., Yang, Y., Guo, G., Liu, Y., Zhang, Z., Dong, S., et al. (2017). IKBKE regulates cell proliferation and epithelial–mesenchymal transition of human malignant glioma *via the* Hippo pathway. *Oncotarget* 8, 49502–49514. doi: 10.18632/oncotarget.17738

Nagarajan, A., Malvi, P., and Wajapeyee, N. (2018). Heparan sulfate and heparan sulfate proteoglycans in cancer initiation and progression. *Front. Endocrinol. (Lausanne)* 9, 483. doi: 10.3389/fendo.2018.00483

Nieto, M. A. (2002). The snail superfamily of zinc-finger transcription factors. *Nat. Rev. Mol. Cell Biol.* 3, 155–166. doi: 10.1038/nrm757

Northcott, J. M., Czubryt, M. P., and Wigle, J. T. (2017). Vascular senescence and ageing: a role for the MEOX proteins in promoting endothelial dysfunction. *Can. J. Physiol. Pharmacol.* 95, 1067–1077. doi: 10.1139/cjpp-2017-0149

Ostrom, Q. T., Gittleman, H., Liao, P., Vecchione-Koval, T., Wolinsky, Y., Kruchko, C., et al. (2017). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro. Oncol.* 19, v1–v88. doi: 10.1093/neuonc/nox158

Skoda, A. M., Simovic, D., Karin, V., Kardum, V., Vranic, S., and Serman, L. (2018). The role of the hedgehog signaling pathway in cancer: a comprehensive review. *Bosn. J. Basic Med. Sci.* 18, 8–20. doi: 10.17305/bjbms.2018.2756

Strobl-Mazzulla, P. H., and Bronner, M. E. (2012). A PHD12-Snail2 repressive complex epigenetically mediates neural crest epithelial-to-mesenchymal transition. *J. Cell Biol.* 198, 999–1010. doi: 10.1083/jcb.201203098

Tachon, G., Masliantsev, K., Rivet, P., Petropoulos, C., Godet, J., Milin, S., et al. (2019). Prognostic significance of MEOX2 in gliomas. *Mod. Pathol.* 32, 774–786. doi: 10.1038/s41379-018-0192-6

Taipale, J., and Beachy, P. A. (2001). The hedgehog and Wnt signalling pathways in cancer. *Nature* 411, 349–354. doi: 10.1038/35077219

Touat, M., Idbaih, A., Sanson, M., and Ligon, K. L. (2017). Glioblastoma targeted therapy: updated approaches from recent biological insights. *Ann. Oncol.* 28, 1457–1472. doi: 10.1093/annonc/mdx106

Valcourt, U., Thuault, S., Pardali, K., Heldin, C. H., and Moustakas, A. (2007). Functional role of Meox2 during the epithelial cytostatic response to TGF-beta. *Mol. Oncol.* 1, 55–71. doi: 10.1016/j.molonc.2007.02.002

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Wade, A., Robinson, A. E., Engler, J. R., Petritsch, C., James, C. D., and Phillips, J. J. (2013). Proteoglycans and their roles in brain cancer. *FEBS J.* 280, 2399–2417. doi: 10.1111/febs.12109

Wu, Z., Guo, H., Chow, N., Sallstrom, J., Bell, R. D., Deane, R., et al. (2005). Role of the MEOX2 homeobox gene in neurovascular dysfunction in Alzheimer disease. *Nat. Med.* 11, 959–965. doi: 10.1038/nm1287

Xu, M., Li, X., Liu, T., Leng, A., and Zhang, G. (2012). Prognostic value of hedgehog signaling pathway in patients with colon cancer. *Med. Oncol.* 29, 1010–1016. doi: 10.1007/s12032-011-9899-7

Xu, Y., Stamenkovic, I., and Yu, Q. (2010). CD44 attenuates activation of the hippo signaling pathway and is a prime therapeutic target for glioblastoma. *Cancer Res.* 70, 2455–2464. doi: 10.1158/0008-5472.CAN-09-2505

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

Yang, H. W., Menon, L. G., Black, P. M., Carroll, R. S., and Johnson, M. D. (2010). SNAI2/Slug promotes growth and invasion in human gliomas. *BMC Cancer* 10, 301. doi: 10.1186/1471-2407-10-301

Yang, Q., Wan, L., Xiao, C., Hu, H., Wang, L., Zhao, J., et al. (2017). Inhibition of LHX2 by miR-124 suppresses cellular migration and invasion in non-small cell lung cancer. *Oncol. Lett.* 14, 3429–3436. doi: 10.3892/ol.2017.6607

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zakrzewski, K., Jarzab, M., Pfeifer, A., Oczko-Wojciechowska, M., Jarzab, B., Liberski, P. P., et al. (2015). Transcriptional profiles of pilocytic astrocytoma are related to their three different locations, but not to radiological tumor features. *BMC Cancer* 15, 778. doi: 10.1186/s12885-015-1810-z

Zeng, J. H., Yang, Z., Xu, J., Qiu, M. L., and Lin, K. C. (2006). Down-regulation of the gax gene in smooth muscle cells of the splenic vein of portal hypertension patients. *Hepatobiliary Pancreat. Dis. Int.* 5, 242–245.

Zoncu, R., Efeyan, A., and Sabatini, D. M. (2011). mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat. Rev. Mol. Cell Biol.* 12, 21–35. doi: 10.1038/nrm3025

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

MINI REVIEW
published: 02 October 2019
doi: 10.3389/fonc.2019.00984

AI Meets Exascale Computing: Advancing Cancer Research With Large-Scale High Performance Computing

*Tanmoy Bhattacharya[1†], Thomas Brettin[2], James H. Doroshow[3], Yvonne A. Evrard[4], Emily J. Greenspan[5], Amy L. Gryshuk[6], Thuc T. Hoang[7], Carolyn B. Vea Lauzon[8], Dwight Nissley[9], Lynne Penberthy[10], Eric Stahlberg[11], Rick Stevens[2,12], Fred Streitz[13], Georgia Tourassi[14], Fangfang Xia[15] and George Zaki[11]**

[1] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, United States, [2] Computing, Environment and Life Sciences Directorate, Argonne National Laboratory, Lemont, IL, United States, [3] Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, United States, [4] Applied Development and Research Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, United States, [5] Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD, United States, [6] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, United States, [7] National Nuclear Security Administration, U.S. Department of Energy, Advanced Simulation and Computing, Washington, DC, United States, [8] Office of Science, U.S. Department of Energy, Advanced Scientific Computing Research, Washington, DC, United States, [9] NCI RAS Initiative, Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, Frederick, MD, United States, [10] Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, United States, [11] Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, United States, [12] Computer Science Department, University of Chicago, Chicago, IL, United States, [13] High Performance Computing Innovation Center, Lawrence Livermore National Laboratory, Livermore, CA, United States, [14] Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [15] Data Science and Learning Division, Argonne National Laboratory, Lemont, IL, United States

OPEN ACCESS

Edited by:
Daoud Meerzaman, George Washington University, United States

Reviewed by:
Yuriy Gusev, Georgetown University, United States
Amir Bahmani, Stanford University, United States

***Correspondence:**
George Zaki
george.zaki@nih.gov

[†]*Authors list in alphabetic order*

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *06 May 2019*
Accepted: *16 September 2019*
Published: *02 October 2019*

Citation:
Bhattacharya T, Brettin T, Doroshow JH, Evrard YA, Greenspan EJ, Gryshuk AL, Hoang TT, Lauzon CBV, Nissley D, Penberthy L, Stahlberg E, Stevens R, Streitz F, Tourassi G, Xia F and Zaki G (2019) AI Meets Exascale Computing: Advancing Cancer Research With Large-Scale High Performance Computing. Front. Oncol. 9:984. doi: 10.3389/fonc.2019.00984

The application of data science in cancer research has been boosted by major advances in three primary areas: (1) Data: diversity, amount, and availability of biomedical data; (2) Advances in Artificial Intelligence (AI) and Machine Learning (ML) algorithms that enable learning from complex, large-scale data; and (3) Advances in computer architectures allowing unprecedented acceleration of simulation and machine learning algorithms. These advances help build *in silico* ML models that can provide transformative insights from data including: molecular dynamics simulations, next-generation sequencing, omics, imaging, and unstructured clinical text documents. Unique challenges persist, however, in building ML models related to cancer, including: (1) access, sharing, labeling, and integration of multimodal and multi-institutional data across different cancer types; (2) developing AI models for cancer research capable of scaling on next generation high performance computers; and (3) assessing robustness and reliability in the AI models. In this paper, we review the National Cancer Institute (NCI) -Department of Energy (DOE) collaboration, *Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)*, a multi-institution collaborative effort focused on advancing computing and data technologies to accelerate cancer research on three levels: molecular, cellular,

and population. This collaboration integrates various types of generated data, pre-exascale compute resources, and advances in ML models to increase understanding of basic cancer biology, identify promising new treatment options, predict outcomes, and eventually prescribe specialized treatments for patients with cancer.

Keywords: cancer research, high performance computing, artificial intelligence, deep learning, natural language processing, multi-scale modeling, precision medicine, uncertainty quantification

INTRODUCTION

Predictive computational models for patients with cancer can in the future support prevention and treatment decisions by informing choices to achieve the best possible clinical outcome. Toward this vision, in 2015, the national Precision Medicine Initiative (PMI) (1) was announced, motivating efforts to target and advance precision oncology, including looking ahead to the scientific, data and computational capabilities needed to advance this vision. At the same time, the horizon of computing was changing in the life sciences, as the capabilities and transformations enabled by exascale computing were coming into focus, driven by the accelerated growth in data volumes and anticipated new sources of information catalyzed by new technologies and initiatives such as PMI.

The National Strategic Computing Initiative (NSCI) in 2015 named the Department of Energy (DOE) as a lead agency for "advanced simulation through a capable exascale computing program" and the National Institutes of Health (NIH) as one of the deployment agencies to participate "in the co-design process to integrate the special requirements of their respective missions." This interagency coordination structure opened the avenue for a tight collaboration between the NCI and the DOE. With shared aims to advance cancer research while shaping the future for exascale computing, the NCI and DOE established the JDACS4C in June of 2016 through a 5-year memorandum of understanding with three co-designed pilot efforts to address both national priorities. The high-level goals of these three pilots were to push the frontiers of computing technologies in specific areas of cancer research: (1) Cellular-level: advance the capabilities of patient-derived pre-clinical models to identify new treatments; (2) Molecular-level: further understand the basic biology of undruggable targets; and (3) Population-level: gain critical insights on the drivers of population cancer outcomes. The pilots would also develop new Uncertainty Quantification (UQ) methods to evaluate confidence in the AI model predictions.

Using co-design principles, each of the pilots in the JDACS4C collaboration is based on—and driven by—team science, which is the hallmark of the collaboration's success. Enabled by deep learning, Pilot One (cellular-level) combines data in innovative ways to develop computationally predictive models for tumor response to novel therapeutic agents. Pilot Two (molecular-level) combines experimental data, simulation, and AI to provide new windows to understand and explore the biology of RAS-related cancers. Pilot Three (population-level) uses AI and clinical information at unprecedented scales to enable precision cancer surveillance to transform cancer care.

AI AND LARGE-SCALE COMPUTING TO PREDICT TUMOR TREATMENT RESPONSE

After years of efforts within the research and pharmaceutical sectors, many patients with cancer still do not respond to standard-of-care treatments, and emergence of therapy resistance is common. Efforts in precision medicine may someday change this by using a targeted therapeutics approach, individually tailored to each patient based on predictive models that use molecular and drug signatures. The *Predictive Modeling for Pre-Clinical Screening Pilot* (Pilot One) aims to develop predictive capabilities of drug response in pre-clinical models of cancer to improve and expedite the selection and development of new targeted therapies for patients with cancer. Highlights of the work done in Pilot One is shown in **Figure 1**.

As omics data continues to accumulate, computational models integrating multimodal data sources become possible. Multimodal deep learning (2) aims to enhance learned features for one task by learning features over multiple modalities. Early Pilot One work (3) measured performance of multi-modal deep neural network drug pair response models with 5-fold cross validation. Using the NCI-ALMANAC (4) data, best model performance was demonstrated when gene expression, microRNA, proteome, and Dragon7 drug descriptors (5) were combined obtaining an R-squared value of 0.944, which indicates that over 94% of the variation in tumor response is explained by the variation among the contributing gene expression, micro RNA expression, proteomics and drug property data.

Mechanistically informed feature selection is an alternative approach that has the potential to increase predictive model performance. The LINCS landmark genes (6) for example has been used to train deep learning models to predict gene expression of non-landmark genes (7) and to classify drug-target interactions (8). Ongoing work in Pilot One is exploring the impact on prediction using gene sets like that of the LINCS landmark genes and other mechanistically defined gene sets. The potential of employing mechanistically informed feature selection extends beyond improving prediction accuracy, to building models on the basis of existing biological knowledge.

Transfer learning is another area of important research activity. The goal of transfer learning is to improve learning in the target learning task by leveraging knowledge from an existing source task (9). Given challenges in obtaining sufficient data

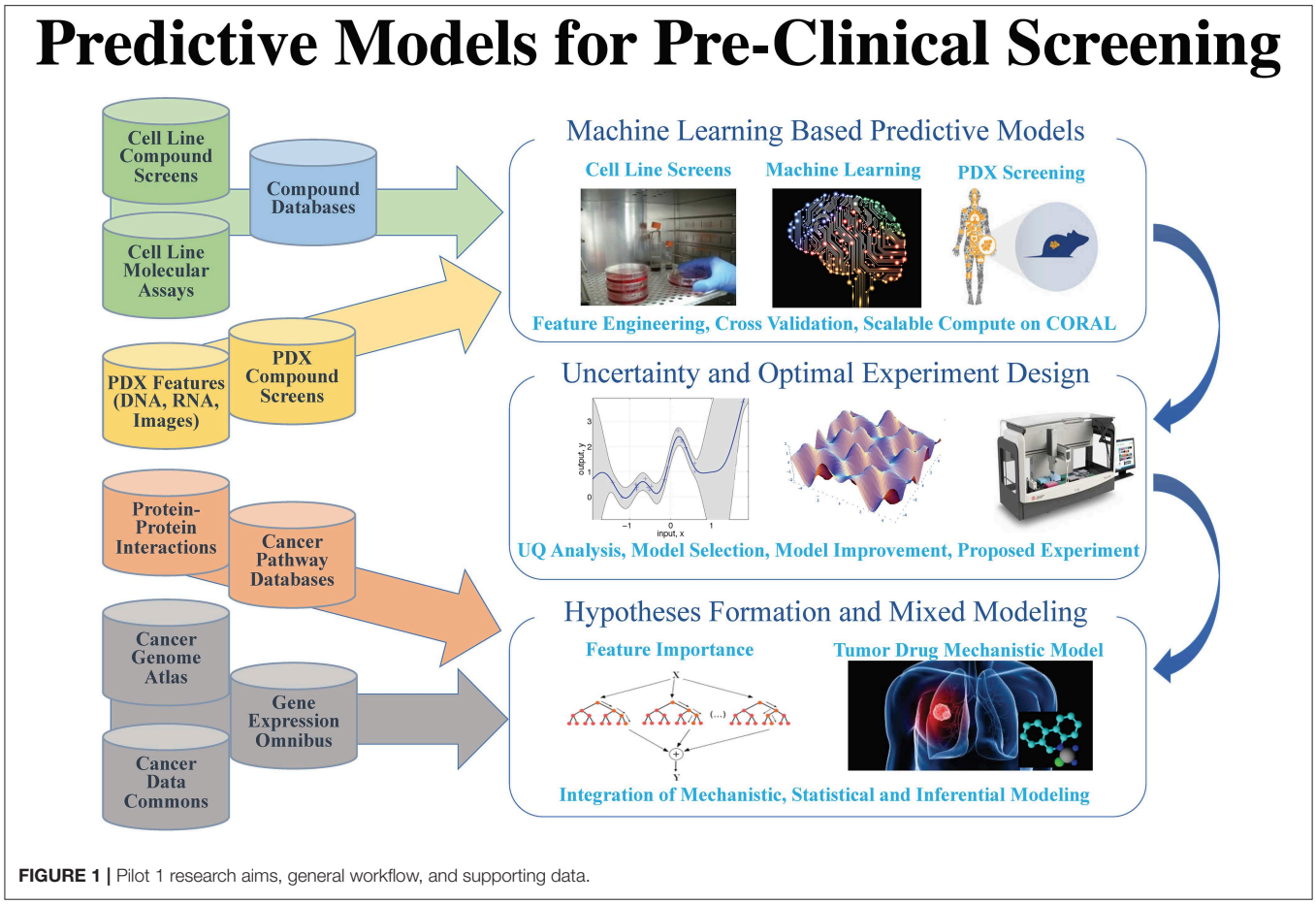


FIGURE 1 | Pilot 1 research aims, general workflow, and supporting data.

for target Patient Derived Xenografts (PDXs), where tumors are grown in mouse host animals, ongoing transfer learning work holds promise for learning on cell lines as a source for the target PDX model predictions. Pilot One is first working on generating models that generalize across cell line studies, a precursor to transfer learning from cell lines to PDXs.

Using data from the NCI-ALMANAC (4), NCI-60 (10), GDSC (11), CTRP (12), gCSI (13), and CCLE (14), models can be constructed that generalize across cell-line studies. Using multi-task networks which combines additional learning of three different classification tasks—tumor/normal, cancer type, and cancer site—with learning of the drug response task, it could be possible to capture more of the total variance and improve precision and recall when training on CTRP and predicting on CCLE for example. Demonstrating cross-study model capability will provide additional confidence that general models can be developed for prediction tasks on cell lines and PDXs and organoids.

Answering questions of how much data and what methods are suitable is a critical part of Pilot One. Although it is generally held that deep learning methods outperform traditional machine learning methods when large data sets are used, this has not yet been explored in the context of drug response prediction problem. Early efforts underway in Pilot One are exploring the relationship among sample size, deep learning methods, and traditional machine learning methods to better characterize the dependencies on predictive performance. This information of sample size together with model accuracy metrics will be of critical importance to future experimental designs for those who wish to pursue deep learning approaches to the drug response prediction problem.

Such extensive deep learning and machine learning investigations require significant computational resources, such as those available at DOE Leadership Computing Facilities (LCF) employed by Pilot 1. A recent experiment searched 23,200 deep neural network models using COXEN (15) selected features and Bayesian optimization ideas (16) to find the best model hyperparameters (hyperparameters generally define the choice of functions and relationship among functions in a given deep learning model). This produced the best cross-study validation results to-date, underscoring the critical need for feature selection and hyperparameter optimization when building predictive models. Further, uncertainty quantification (explained in more depth later) adds a new level of computing demand. Uncertainty quantification experiments involving over 30 billion predictions from 450 of the best models generated on the DOE Summit LCF system are ongoing to understand the relationship to between best model

uncertainty and the model that performs best in cross-study validation experiments.

Reflecting on insights from Pilot 1 activities and current gaps in available literature, future work will focus on exploring new predictive models to better utilize, ground, and enrich biological knowledge. Efforts to improve drug representations for response prediction are expected to benefit from research involving training semi-supervised networks on millions of compounds. In efforts to improve understanding of trained models, mechanistic information is being incorporated into more interpretable deep learning models. Active learning in response prediction—which balances uncertainty, accuracy, and lead discovery—will be used to guide the acquisition of experimental data for animal models in a cost-effective and timely manner. And finally, a necessary step toward precision models is gaining a fine-grained understanding of prediction error, an insight enabled by the demonstrated capability in large-scale model sweeps.

AI AT THE FOREFRONT OF RAS RELATED CANCERS

Oncogenic mutations in RAS genes are associated with more than 30% of cancers and are particularly prevalent in those of the lung, colon and pancreas. Though RAS mutations have been studied for decades, there are currently no RAS inhibitors and a detailed molecular mechanism for how RAS engages and activates proximal signaling proteins (RAF) remains elusive (17). RAS signaling takes place at and is dependent on cellular membranes, a complex cellular environment that is difficult to recapitulate using current experimental technologies.

Pilot Two*, Improving Outcomes for RAS-related Cancer,* is focused on delivering a validated multiscale model of RAS biology on a cell membrane by combining the experimental capabilities at the Frederick National Laboratory for Cancer Research with the computational resources of the National Nuclear Security Administration (NNSA), a semi-autonomous agency of the DOE. The principal challenge in modeling this system is the diverse length and time scales involved. Lipid membranes evolve over a macroscopic scale (micrometers and milliseconds). Capturing this evolution is critical, as changes in lipid concentration define the local environment in which RAS operates. The RAS protein itself, however, binds over time and length scales which are microscopic (nanometers and microseconds). In order to elucidate the behavior of RAS proteins in the context of a realistic membrane, our modeling effort must span the multiple orders of magnitude between microscopic and macroscopic behavior. The Pilot Two team has built such a framework, developing a macroscopic model that captures the evolution of the lipid environment and which is consistent with an optimized microscopic model that captures protein-protein and protein-lipid interactions at the molecular scale. Macroscopic model components (lipid environment, lipid-lipid interactions, protein behavior and protein-lipid interactions) were characterized through close collaboration between the experimentalists at Frederick National Laboratory and the computational scientists from the DOE/NNSA. The microscopic model is based on standard Martini force fields for Coarse-Grained Molecular Dynamics (CGMD), modified to correctly capture certain details of lipid phase behavior (18–21). A snapshot from a typical micro-scale simulation run, showing two RAS proteins on a 30 nm × 30 nm patch of lipid membrane (containing ∼150,000 particles) is shown in **Figure 2**.

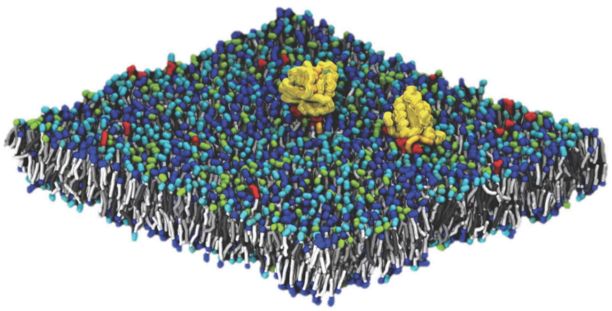


FIGURE 2 | CGMD simulation captures the molecular details of RAS in complex lipid membranes.

In order to bring the two scales together, the team devised a novel workflow whereby microscopic subsections of a running macroscopic model are scored for uniqueness using a machine learning algorithm operating in a reduced order space that has been trained on previous simulations. The most unique subsections in the macroscopic simulation are identified and re-created as CGMD simulations, which explore the microscopic behavior. Information from the (many thousands of) microscopic simulations is then fed back into the macroscopic model, so that it is continually improving even as the simulations are running (22).

This modeling infrastructure was designed to exploit the Sierra supercomputer at Lawrence Livermore National Laboratory. The scale and heterogeneous architecture of Sierra make it ideal for such a workflow that combines AI technology with predictive simulation. Running on the entire machine, the team was able to simulate at the macroscopic level a 1 by 1μm, 14-lipid membrane with 300 RAS proteins, generating over 100,000 microscopic simulations capturing over 200 ms of protein behavior. This unprecedented achievement represents an almost two orders of magnitude improvement on the previously state of the art. That being said, the space of all possible lipid mixtures is huge, requiring tens of thousands of samples for any meaningful coverage. This type of MUltiple Metrics Modeling Infrastructure (MuMMI) simulations will always be limited by the available High Performance Computing (HPC) resources. With an exascale machine we can substantially increase the dimensionality of the input space and its coverage, significantly improving the applicability of future campaigns.

In the coming years, the team will exploit this capability to explore RAS behavior on lipid membranes and extend the model in three important directions. First, both the macro and micro models will be modified to incorporate the RAF kinase, which binds to RAS as the first step in the MAPK pathway that leads to growth signaling. Second, we will extend the

infrastructure to include fully atomistic resolution, creating a three-level (macro/micro/atomistic) multiscale model. Third, we will incorporate membrane curvature into the dynamics of the membrane, which is currently constrained to remain flat. The improved infrastructure will allow the largest and most accurate computational exploration of RAS biology to date.

ADVANCING CANCER SURVEILLANCE USING AI AND HIGH-PERFORMANCE COMPUTING

The Surveillance, Epidemiology, and End Results (SEER) program funded by the NCI was established in 1973 for the advancement of public health and for reducing the cancer burden in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering ∼34.6% of the U.S. population. The curated, population-level SEER data provide a rich information source for data-driven discovery to understand drivers of cancer outcomes in the real world.

An outstanding challenge of the SEER program is how to achieve near real-time cancer surveillance. Information abstraction is a critical step to facilitate data-driven explorations. However, the process is fully manual to ensure high quality data. As the SEER program increases the breadth of information captured, the manual process is no longer scalable. By partnering computational and data scientists from DOE with NCI SEER domain experts, Pilot Three, *Population Information Integration, Analysis, and Modeling for Precision Surveillance,* aims to leverage high-performance computing and artificial intelligence to meet the emerging needs of cancer surveillance. Moreover, Pilot Three envisions a fully integrated data-driven modeling and simulation framework to enable meaningful translation of big SEER data. By collecting and linking additional patient data, we can generate profiles for patients with cancer that include information about healthcare delivery system parameters and continuity of care. Such rich data will facilitate data-driven modeling and simulation of patient-specific health trajectories to support precision oncology research at the population level.

To date, Pilot Three has mainly focused on the development, scaling, and deployment of cutting-edge AI tools to semi-automate information abstraction from unstructured pathology text reports, the main source of information of cancer registries. In partnership with the Louisiana Tumor Registry and the Kentucky Cancer Registry, several AI-based Natural Language Processing (NLP) tools have been developed and benchmarked for abstraction of fundamental cancer data elements such as cancer site, laterality, behavior, histology, and grade (23–29). The NLP tools rely on the latest AI advances including multi-task learning and attention mechanisms. Scalable training and hyperparameter optimization of the tools is managed by relying on pre-exascale computing infrastructure available within the DOE laboratory complex (30). Following an iterative optimization protocol, the most computationally efficient and clinically effective tools are deployed for evaluation across participating SEER registries. Based on preliminary testing the NLP tools have been able to accurately classify all five data elements for 42.5% of cancer cases. Further refinement of this accuracy level is underway in subsequent versions as well as incorporation of an uncertainty quantification component to ease and increase user confidence.

Although the patient information currently collected across SEER registries is mainly clinical (clin-omics), increasingly other -omics type of information is expected to become part of cancer surveillance. Specifically, radiomics (i.e., biomarkers automatically extracted from histopathological and radiological images via targeted image processing algorithms) as well as genomics will provide important insight to understand the effectiveness of cancer treatment choices.

Moving forward, Pilot Three will implement the latest NLP tools into production application across participating SEER registries using Application Program Interfaces (APIs) to determine the most effective human-AI workflow integration for broad and standardized technology integration across registries. The APIs will be integrated in the registries' workflows. In addition, working collaboratively with domain experts, the team will extend the information extraction across biomarkers and capture disease progression such as metastasis and recurrence. This pilot is engaging in several partnerships with academic and commercial entities to bring in heterogeneous data sources for more effective longitudinal trajectory modeling. Efforts to understand causal inference beyond treatment (social, economic, and environmental) impact in the real world are also part of future plans.

LOOKING AHEAD: OPPORTUNITIES AND CHALLENGES

In addition to large-scale computing as a critical and necessary element to pursue the many opportunities for AI in cancer research, other areas must also develop to realize the tremendous potential. In this section, we list some of these opportunities.

First, HPC platforms provide high-speed interconnect between compute nodes that is integral in handling the communication for data or model parallel training. While cloud platforms have recently made significant investments in improving interconnect, this remains a challenge and would encourage projects like Pilot Three to limit distributed training to a single node. That said, on-demand nature of cloud platforms can allow for more efficient resource utilization of AI workflows, and the modern Linux environments and familiar hardware configurations available on cloud platforms offer superior support for AI workflow software which can increase productivity.

Second, the level of available data currently limits the potential for AI in cancer research. Developing data resources of sufficient size, quality, and coherence will be essential for AI to develop robust models within the domain of the available data resources.

Third, evaluation and validation of data-driven AI models, and quantifying the uncertainty in individual predictions, will continue to be an important aspect for the adoption of AI in cancer research, posing a challenge to the community to

concurrently develop criteria for evaluation and validation of models while delivering the necessary data and large-scale computational resources required.

In the next two subsections, we highlight two efforts within the JDAC4C collaboration to address these challenges. The first focuses on scaling the training of the deep neural network application on HPC systems, and the second quantifies the uncertainty in the trained models to build a measure of confidence and limits on how to use them in production.

CANDLE: CANcer Distributed Learning Environment

CANDLE (16), builds a single, scalable deep neural network application and is being used to address the challenges in each of the JDACS4C pilots.

The challenge problem for the CANDLE project is to enable the most challenging deep learning problems in cancer research to run on the most capable supercomputers in the DOE and NIH. Implementations of CANDLE have been tested on the DOE Titan, Cori, Theta and Summit systems, and using container technologies on the NIH Biowulf system (31). The CANDLE software builds on open source deep learning frameworks including Keras, TesnsorFlow and PyTorch. Through collaborations with DOE computing centers, HPC vendors and Exascale Computing Project (ECP) co-design and software technology projects, CANDLE is being prepared for the coming DOE exascale platforms.

Features currently supported in CANDLE include feature selection, hyperparameter optimization, model training, inferencing and UQ. Future release plans call for supporting experimental design, model acceleration, uncertainty guided inference, network architecture search, synthetic data generation and data modality conversion. These features have been used to evaluate over 20,000 models in a single run on a DOE HPC system.

The CANDLE project also features a set of deep learning benchmarks that are aimed at solving a problem associated with each of the pilots. These benchmarks embody different deep learning approaches to problems in cancer biology, and they are implemented in compliance with CANDLE standards making them amenable to large-scale model search and inferencing experiments.

Uncertainty Quantification

UQ is a critical component across all three JDACS4C pilots. It is a field of analysis that estimates accuracy under multi-modal uncertainties. UQ allows detecting unreliable model predictions (32) and provides for improved design of experiments. UQ quantifies the effects of statistical fluctuations, extrapolation, overfitting, model misspecification and sampling biases, resulting in confidence measures for individual model prediction.

Historically, results from computational modeling in the biological sciences did not incorporate UQ, but measures of certainty are essential for *actionable* predictive analytics (33). The problems are exacerbated as we start addressing problems with poorly understood causal models using large—but noisy, multimodal and incomplete—data sets. Methodological advances are allowing all three pilots to use HPC technology to simultaneously estimate the uncertainty along with the results.

In addition to providing confidence intervals, the development of new UQ technology allows assessment and improvement of data quality (34); evaluation and design of models appropriate to the data quality and quantity; and prioritization of further observations or experiments that can best improve model quality. These developments are currently being tested in the JDACS4C pilots and are likely to impact the wider application of large-data-driven modeling.

CONCLUSION

The JDACS4C collaboration continues to provide valuable insights into the future for AI in cancer research and the essential role that extreme-scale computing will have in shaping and informing that future. Concepts have been transformed into preliminary practice in a short period of time, as a result of multi-disciplinary teamwork and access to advanced computing resources. AI is being used to guide experimental design to make more effective use of valuable laboratory resources, to develop new capabilities for molecular simulation, and to streamline and improve efficiencies in the acquisition of clinical data.

The JDACS4C collaboration established a foundation for team science and is enabling innovation at the intersection of advanced computing technologies and cancer research. The opportunities for extreme-scale computing in AI and cancer research extend well beyond these pilots.

AUTHOR CONTRIBUTIONS

GZ made the paper plan, wrote the abstract, and assembled the manuscript. ES, EG, AG, TH, and CL contributed to the introduction. RS, JD, and YE made substantial contribution to the conception and design of the work in Pilot One. TBr and FX provided acquisition, analysis, and interpretation of data in Pilot One. FS and DN made substantial contributions to the conception, design, and execution of Pilot Two. GT and LP made substantial contribution in designing, developing, and writing the section about Pilot Three. RS made substantial contribution to the conception and design of the work in CANDLE. RS, TBr, TBh, and FX developed and optimized predictive models with UQ in Pilots One, Three, and CANDLE. GZ contributed to extending the application of CANDLE at NIH. ES and EG contributed to the opportunities and challenges. TBh wrote and developed the section related to UQ. EG, AG, TH, and CL contributed to the conclusion.

FUNDING

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National

Laboratory under Contract DE-AC02-06-CH11357. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-773355. This work was performed under the auspices of the U.S. Department of Energy by Los Alamos National Laboratory under Contract DE-AC52-06NA25396. Computing support for this work came in part from the Lawrence Livermore National Laboratory Institutional Computing Grand Challenge program. This project was funded in part with federal funds from the NCI, NIH, under contract no. HHSN261200800001E. This research was supported in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This research used resources of the Argonne Leadership Computing Facility and the Oak Ridge Leadership Computing Facility, which are DOE Office of Science User Facilities. This research used resources of the Lawrence Livermore Computing Facility and the Los Alamos National Laboratory supported by the DOE National Nuclear Security Administration's Advanced Simulation and Computing (ASC) Program. This manuscript has been authored in part by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paidup, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

REFERENCES

1. PMI (2019). Retrieved from https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative (accessed September 20, 2019).
2. Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans Comput Biol Bioinform.* (2018) 16:1–1. doi: 10.1109/TCBB.2018.2879673
3. Xia F, Shukla M, Brettin T, Garcia-Cardona C, Cohn J, Allen JE, et al. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics.* (2018) 19:486. doi: 10.1186/s12859-018-2509-3
4. Holbeck SL, Camalier R, Crowell JA, Govindharajulu JP, Hollingshead M, Anderson LW, et al. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* (2017) 77:3564–76. doi: 10.1158/0008-5472.CAN-17-0489
5. Dragon. *Software for Molecular Descriptor Calculation*. (2019). Retrieved from https://chm.kode-solutions.net (accessed April 30, 2019)
6. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell.* (2017) 171:1437–52 e17. doi: 10.1016/j.cell.2017.10.049
7. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. *Bioinformatics.* (2016) 32:1832–9. doi: 10.1093/bioinformatics/btw074
8. Xie L, He S, Song X, Bo X, Zhang Z. Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC Genomics.* (2018) 19:667. doi: 10.1186/s12864-018-5031-0
9. Torrey L, Shavlik J. Chapter 11: Transfer learning. In: Olivas ES, Guerrero JDM, Martinez-Sober M, Magdalena-Benedito JR, López AJS, editors. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques.* Vol. 2. (2009). p. 242–64. doi: 10.4018/978-1-60566-766-9
10. Developmental Therapeutics Program (2019). Retrieved from: http://dtp.nci.nih.gov/ (accessed September 20, 2019).
11. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* (2013) 41:D955–61. doi: 10.1093/nar/gks1111
12. Basu A, Bodycombe N, Cheah J, Price E, Liu K, Schaefer G, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell.* (2013) 154:1151–61. doi: 10.1016/j.cell.2013.08.003
13. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol.* (2015) 33:306–12. doi: 10.1038/nbt.3080
14. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* (2012) 483:603–7. doi: 10.1038/nature11003
15. Smith SC, Baras AS, Lee JK, Theodorescu D. The COXEN principle: translating signatures of *in vitro* chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer. *Cancer Res.* (2010) 70:1753–8. doi: 10.1158/0008-5472.CAN-09-3562
16. Wozniak J, Jain R, Balaprakash P, Ozik J, Collier N, Bauer J, et al. CANDLE/supervisor: a workflow framework for machine learning applied to cancer research. *BMC Bioinformatics.* (2018) 19(Suppl. 18):491. doi: 10.1186/s12859-018-2508-4
17. Simanshu DK, Nissley DV, McCormick F. RAS proteins and their regulators in human disease. *Cell.* (2017) 170:17–33. doi: 10.1016/j.cell.2017.06.009
18. Carpenter TS, López CA, Neale C, Montour C, Ingólfsson HI, Di Natale F, et al. Capturing phase behavior of ternary lipid mixtures with a refined martini coarse-grained force field. *J Chem Theory Comput.* (2018) 14:6050–62. doi: 10.1021/acs.jctc.8b00496
19. Neale AC, Garcia EA. Methionine 170 is an environmentally sensitive membrane anchor in the disordered HVR of K-Ras4B. *J Phys Chem B.* (2018) 122:10086–96. doi: 10.1021/acs.jpcb.8b07919
20. Ingólfsson HI, Carpenter TS, Bhatia H, Bremer P-T, Marrink SJ, Lightstone FC. Computational lipidomics of the neuronal plasma membrane. *Biophys J.* (2017) 113:2271–80. doi: 10.1016/j.bpj.2017.10.017
21. Travers T, López C, Van Q, Neale C, Tonelli M, Stephen A, et al. Molecular recognition of RAS/RAF complex at the membrane: role of RAF cysteine-rich domain. *Sci Rep.* (2018) 8:8461. doi: 10.1038/s41598-018-26832-4
22. Natale FD, Bhatia H, Carpenter TS, Neale C, Schumacher SK, Oppelstrup T, et al. A massively parallel infrastructure for adaptive multiscale simulations: modeling RAS initiation pathway for cancer. In: *To Appear in Supercomputing'19: The International Conference for High Performance Computing, Networking, Storage, and Analysis*. Denver, CO (2019).
23. Qiu JX, Yoon H-J, Srivastava K, Watson TP, Christian JB, Ramanathan A, et al. Scalable deep text comprehension for cancer surveillance on high-performance computing. *BMC Bioinformatics.* (2018) 19(Suppl. 18):488. doi: 10.1186/s12859-018-2511-9
24. Gao S, Young MT, Qiu JX, Yoon H-J, Christian JB, Fearn PA, et al. Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Informatics Assoc.* (2017) 25:321–30. doi: 10.1093/jamia/ocx131

25. Qiu JX, Yoon H-J, Fearn PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE J Biomed Health Informatics.* (2018) 22:244–51. doi: 10.1109/JBHI.2017.2700722
26. Alawad M, Hinkle J, Schaefferkoetter N, Christian J, Fearn P, Wu X-C, et al. DeepAbstractor: a scalable deep learning framework for automated information extraction from free-text pathology reports. In: *AACR Special Conference on Convergence: Artificial Intelligence, Big Data, and Prediction in Cancer*. Newport, RI (2018).
27. Alawad M, Hasan SM, Christian JB, Tourassi GD. Retrofitting word embeddings with the UMLS metathesaurus for clinical information extraction. In: *2018 IEEE International Conference on Big Data (Big Data).* Seatttle, WA (2018). p. 2838–46.
28. Alawad M, Yoon H-J, Tourassi GD. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. In: *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI).* (2018). p. 218–21.
29. Yoon H-J, Robinson S, Christian JB, Qiu JX, Tourassi GD. Filter pruning of convolutional neural networks for text classification: a case study of cancer pathology report comprehension. In: *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI).* Las Vegas, NV (2018). p. 345–8.
30. Yoon H-J, Alawad M, Christian JB, Hinkle J, Ramanathan A, Tourassi G. HPC-based hyperparameter search of MT-CNN for information extraction from cancer pathology reports. In: *Computational Approaches for Cancer Workshop*. Dallas, TX (2018).
31. Zaki GF, Wozniak J, Ozik J, Collier N, Brettin T, Stevens R. Portable and reusable deep learning infrastructure with containers to accelerate cancer studies. In: Fourth *International IEEE Workshop on Extreme Scale Programming Models and Middleware*. Dallas, TX (2018). p. 54–61.
32. Hengartner N, Cuellar L, Wu X-C, Tourassi G, John Q, Christian B, et al. CAT: computer aided triage improving upon the bayes risk through ε-refusal triage rules. *BMC Bioinformatics.* (2018) 19(Suppl. 18):485. doi: 10.1186/s12859-018-2503-9
33. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell.* (2019) 1:20–3. doi: 10.1038/s42256-018-0004-1
34. Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G, Mohd-Yusof J. Combating label noise in deep learning using abstention. In: *36th International Conference on Machine Learning.* Long Beach, CA (2019).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 11 October 2019
doi: 10.3389/fonc.2019.01030

Novel Biomarkers Associated With Progression and Prognosis of Bladder Cancer Identified by Co-expression Analysis

Yejinpeng Wang[1†], Liang Chen[1†], Lingao Ju[2,3,4], Kaiyu Qian[2,3,4], Xuefeng Liu[5], Xinghuan Wang[1,6] and Yu Xiao[1,2,3,4,7*]*

[1] Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, China, [2] Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan, China, [3] Human Genetics Resource Preservation Center of Hubei Province, Wuhan, China, [4] Human Genetics Resource Preservation Center of Wuhan University, Wuhan, China, [5] Department of Pathology, Lombardi Comprehensive Cancer Center, Georgetown University Medical School, Washington, DC, United States, [6] Laboratory of Urology, Medical Research Institute, Wuhan University, Wuhan, China, [7] Laboratory of Precision Medicine, Zhongnan Hospital of Wuhan University, Wuhan, China

OPEN ACCESS

Edited by:
Barbara Karen Dunn,
National Institutes of Health (NIH),
United States

Reviewed by:
Hiroshi Miyamoto,
University of Rochester, United States
Shan Yan,
University of North Carolina at Charlotte, United States

***Correspondence:**
Xinghuan Wang
wangxinghuan@whu.edu.cn
Yu Xiao
yu.xiao@whu.edu.cn

[†]These authors have contributed equally to this work as co-first authors

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 07 March 2019
Accepted: 23 September 2019
Published: 11 October 2019

Citation:
Wang Y, Chen L, Ju L, Qian K, Liu X, Wang X and Xiao Y (2019) Novel Biomarkers Associated With Progression and Prognosis of Bladder Cancer Identified by Co-expression Analysis. Front. Oncol. 9:1030. doi: 10.3389/fonc.2019.01030

Our study's goal was to screen novel biomarkers that could accurately predict the progression and prognosis of bladder cancer (BC). Firstly, we used the Gene Expression Omnibus (GEO) dataset GSE37815 to screen differentially expressed genes (DEGs). Secondly, we used the DEGs to construct a co-expression network by weighted gene co-expression network analysis (WGCNA) in GSE71576. We then screened the brown module, which was significantly correlated with the histologic grade ($r = 0.85$, $p =$ 1e-12) of BC. We conducted functional annotation on all genes of the brown module and found that the genes of the brown module were mainly significantly enriched in "cell cycle" correlation pathways. Next, we screened out two real hub genes (ANLN, HMMR) by combining WGCNA, protein-protein interaction (PPI) network and survival analysis. Finally, we combined the GEO datasets (GSE13507, GSE37815, GSE31684, GSE71576). Oncomine, Human Protein Atlas (HPA), and The Cancer Genome Atlas (TCGA) dataset to confirm the predict value of the real hub genes for BC progression and prognosis. A gene-set enrichment analysis (GSEA) revealed that the real hub genes were mainly enriched in "bladder cancer" and "cell cycle" pathways. A survival analysis showed that they were of great significance in predicting the prognosis of BC. In summary, our study screened and confirmed that two biomarkers could accurately predict the progression and prognosis of BC, which is of great significance for both stratification therapy and the mechanism study of BC.

Keywords: bladder cancer (BC), gene-set enrichment analysis (GSEA), protein-protein interaction (PPI), weighted co-expression network analysis (WGCNA), The Cancer Genome Atlas (TCGA) dataset

INTRODUCTION

BC is one of the most common malignancies of the urinary tract (1), and is a complex disease with high morbidity and mortality if not diagnosed timely and treated optimally (2). It is estimated that there are 429,000 new cases and 165,000 deaths worldwide each year (3). The most common symptom of BC is painless hematuria, which is seen in more than 80% of patients. At present, BC

can be divided into two major categories according to tumor stage: non-muscle invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC) (4, 5). NMIBC is characterized by the co-activation of FGFR3 mutations, high recurrence rate (50–70%), and the 5-year survival rate > 90% (6). However, MIBC is characterized by frequent TP53 mutations, high metastasis and a 5-year survival rate < 50% (7). 70–80% of BC patients had non-muscle-invasive bladder cancer (NMIBC) (8), and 20–30% of these patients will progress to MIBC (9). Once BC progression is detected, the patient's prognosis decreases (10, 11); currently, there is a lack of effective biomarkers that can accurately predict the progress and prognosis of BC, so such biomarkers need to be discovered urgently.

With the rapid development of microarray and high-throughput sequencing technology, bioinformatics plays an important role in various fields (12–15). In the medical field, the most commonly used means of bioinformatics is to find biomarkers (16–18). However, at present, many studies only consider the differences in gene expression between different samples, and only look for biomarkers with differential expression as the limiting condition, while ignoring the underlying connection of each gene (19, 20).

Here, we constructed WGCNA co-expression network and incorporated genes with similar expression patterns into the same modules. After all the modules were related to the calculation of clinical phenotype data, the modules most related to the progression of BC were obtained. Finally, after a series of screening tests, we found the real hub genes (ANLN, HMMR) that could truly predict the progression and prognosis of BC. Our study fully considered the internal relationship between genes, rather than only considering differential expression genes. The GSEA analysis and functional annotation showed that the real hub genes played their role in BC through signaling pathways such as "bladder cancer" and "cell cycle." We combined a large number of databases (GEO, TCGA, Oncomine, HPA, String, GEPIA, GSCALite) to verify the ability of real hub genes to predict the progression and prognosis of BC, ensuring the stability and reliability of the results.

MATERIALS AND METHODS

Data Collection and Study Design

The microarray dataset GSE13507, GSE31684, GSE37815, GSE71576 and the corresponding clinical information data of these microarray datasets were downloaded from the Gene Expression Omnibus (GEO) database of the NCBI database (https://www.ncbi.nlm.nih.gov/). The datasets GSE37815 and GSE13507 both performed on the Illumina human-6 v2.0 platform, the former was used to screen for different expression genes (DEGs), the latter was used to verify the hub genes. The dataset GSE71576, which performed on the Affymetrix Human Gene 1.0 ST platform, was used to perform weighted co-expression network analysis. The dataset GSE31684, which performed on the Affymetrix Human Genome U133 Plus 2.0 platform, was also used to verify the hub genes. The level three RNA-seq data (Illumina RNASeqV2) and corresponding clinical information about BC were downloaded from The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/). The dataset, which included 408 BC samples and 19 normal bladder samples, was used to verify the hub genes, perform GSEA, correlation analysis and survival analysis. The inclusion cohort was defined as a cohort containing microarray or RNA-seq data and clinical phenotypes and follow-up data. By consulting the literature, we took the cohorts without performed WGCNA as training sets and internal validation sets, and the cohorts that have undergone WGCNA research as external validation sets. Dataset GSE37815 contained 18 BC and 6 normal bladder samples, so we chose it for DEGs analysis. Furthermore, we chose datasets GSE37815 and GSE71576 as training and internal validation datasets, whereas the datasets GSE13507, GSE31684, and TCGA were set as external validation datasets. The detailed information of these datasets was listed in **Table 1**, and the flow chart of our entire experiment is presented in **Figure 1**.

Data Preprocessing and DEGs Screening

All the raw expression data were subject to quality control, background correction, normalization, logarithmic conversion and remove batch effects processing, using the R packages "affy" (21) or "limma" (22). After that, samples without clinical data were filtered out, and the resulting data were subsequently analyzed. The RNA-seq data of the TCGA dataset were normalized using the "DESeq2" (23) R package. The "limma" R package was used to screen the DEGs between eighteen BC and six normal bladder samples in dataset GSE37815. The false discovery rate (FDR) <0.05 and $|\log2FC| \geq 1$ were set as the threshold for screening DEGs.

Establishment of Weighted Co-expression Network

The DEGs were used to construct a weighted co-expression network by the R package "WGCNA" (24). Firstly, we used the function "goodSamplesGenes" in the "WGCNA" package checked to see if the input genes (DEGs) and input samples were good genes and good samples. Secondly, Pearson's correlation analysis of all pairs of genes was used to construct an adjacency matrix. After that, the adjacency matrix was used to construct a scale-free co-expression network based on a soft-thresholding parameter β (β was a soft-thresholding parameter that could enhance strong correlations between genes and penalize weak correlations) (25). The adjacency matrix was then turned into a topological overlap matrix (TOM). TOM could measure the network connectivity of a gene, which was defined as the sum of its adjacency with all other genes, and was used for network generation (26). At the same time, in order to classify genes with similar expression patterns into gene modules, average linkage hierarchical clustering was conducted according to the TOM-based dissimilarity measure with a minimum size (gene group) of 50 for the genes dendrogram.

Identify Significant Relevant Module and Module Functional Annotation

To investigate the biological function of the brown module, which significantly related to the histologic grade of BC, we uploaded the list of all genes in the brown module to

TABLE 1 | Information of datasets used in this study.

Datasets	GSE37815	GSE71576	GSE13507	GSE31684	TCGA
	Training validation datasets		External validation datasets		
Platform	Illumina human-6 v2.0	Affymetrix human gene 1.0 ST	Illumina human-6 v2.0	Affymetrix human genome U133 plus 2.0	Illumina RNASeqV2
SAMPLE NUMBER					
Total	18	44	256	93	427
Bladder cancer	6	44	165	93	408
Normal bladder	–	0	68	0	19
Recurrent bladder cancer	–	–	23	–	–
pStage I	–	–	–	10	2
pStageII	–	–	–	17	130
pStage III	–	–	–	42	140
pStage IV	–	–	–	19	135
Unknown stage	–	–	0	5	1
Grade I	–	14	105	–	–
GradeII	–	11	60	–	–
Grade III	–	17	0	–	–
High grade	–	–	–	87	385
Low grade	–	–	–	6	22
Unknown grade	–	2	–	–	1
Ta	–	27	24	–	–
T1	–	6	80	–	–
T2	–	3	31	–	–
T3	–	2	19	–	–
T4	–	4	11	–	–
Unknown T stage	–	2	–	–	–

the DAVID website (https://david.ncifcrf.gov) for functional annotation analysis. The threshold was the $p < 0.05$.

Real Hub Genes Identification by WGCNA, PPI, and Survival Analysis

By calculating the correlation between modules and clinical phenotypes by the module-trait relationship of WGCNA, we could screen the module most relevant to the clinical phenotype we were interested in. In our study, histologic grade ($r = 0.85$, $p = 1e\text{-}12$) was selected as interested clinical phenotype for subsequent analysis.

After the interesting module was chosen, same as in the past (27, 28), we defined the cor.geneModuleMembership > 0.8 (the correlation between the gene and a certain clinical phenotype) and cor.geneTraitSignificance > 0.2 (the correlation between the module eigengene and the gene expression profile) as the threshold for screening hub genes in a module.

To further target and screen more meaningful hub genes, we uploaded the list of 49 hub genes to the STRING database (https://string-db.org/) to construct a protein-protein interaction (PPI) network (29). The minimum interaction score of these genes was >0.4 and were defined as the threshold of the hub genes of the PPI network. The Cytoscape software (30) was used to visualize network diagrams for PPI analysis. Finally, we used the Gene Expression Profiling Interactive Analysis (GEPIA) database (31) (http://gepia.cancer-pku.cn/) to test the prognostic value of hub genes, and the hub genes with the ability to predict prognosis were the real hub genes. To verify the value of predicting prognosis of hub genes, a survival analysis of real hub genes was performed using the GSE13507 dataset from GEO datasets.

Gene Set Enrichment Analysis of Real Hub Genes

The GSEA software was downloaded from http://software.broadinstitute.org/gsea/index.jsp. The GSEA analysis was conducted with a small cohort GSE37815 and a large cohort TCGA dataset, respectively. We divided the samples into two groups according to the median expression of hub genes, and chose the C2 (c2.cp.kegg.v6.1.symbols.gmt) sub-collection downloaded from the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb/index.jsp) as the reference gene sets to perform GSEA analysis.

Verify the Expression Pattern and the Prognostic Value of Real Hub Genes

The datasets GSE37815 and GSE71576 were selected as internal validation datasets, the datasets GSE31684, GSE13507, and TCGA were set as external validation datasets. All of them were used to verify the real hub genes' mRNA expression pattern in different histologic grades or pathologic stage of BC. In addition, we used the Oncomine database

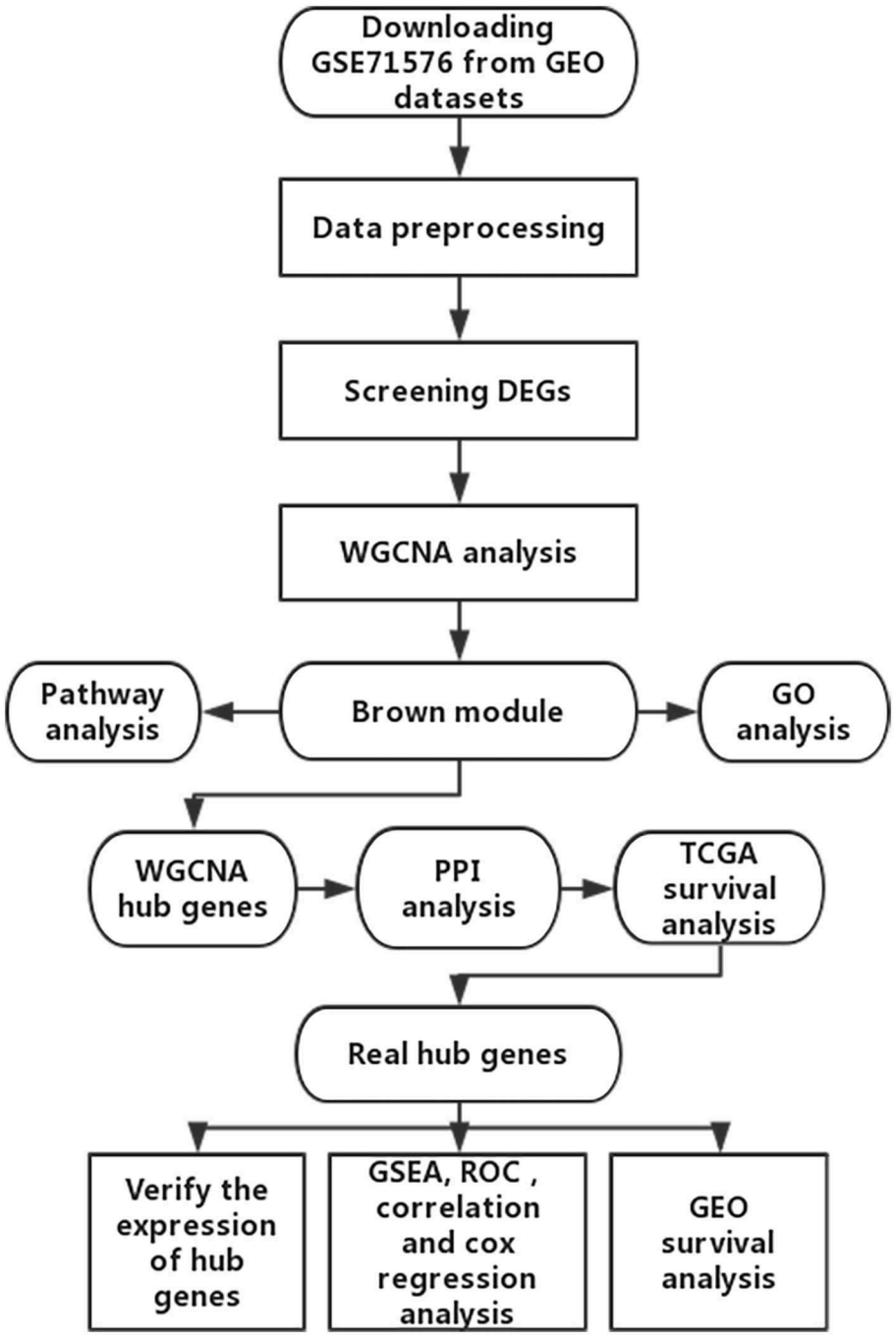


FIGURE 1 | Flow diagram of the study.

(https://www.oncomine.org/resource/main.html) and the above dataset to verify the expression of real hub genes between BC tissues and adjacent tissues. We used the one-way analysis of variance (ANOVA) or Student's *t*-test to measure the statistical significance of the calculated results. After that, we performed a Kaplan-Meier survival analysis of hub genes in each cohort using the "survival" R package.

RESULTS

Screening of Differentially Expressed Genes

The R package "limma" was used to screen DEGs between BC and normal bladder samples in GSE37815, where a total of 792 DEGs were screened (240 up-regulated and 552 down-regulated) under the threshold of FDR < 0.05 and logFC (fold change) ≥ 1. The

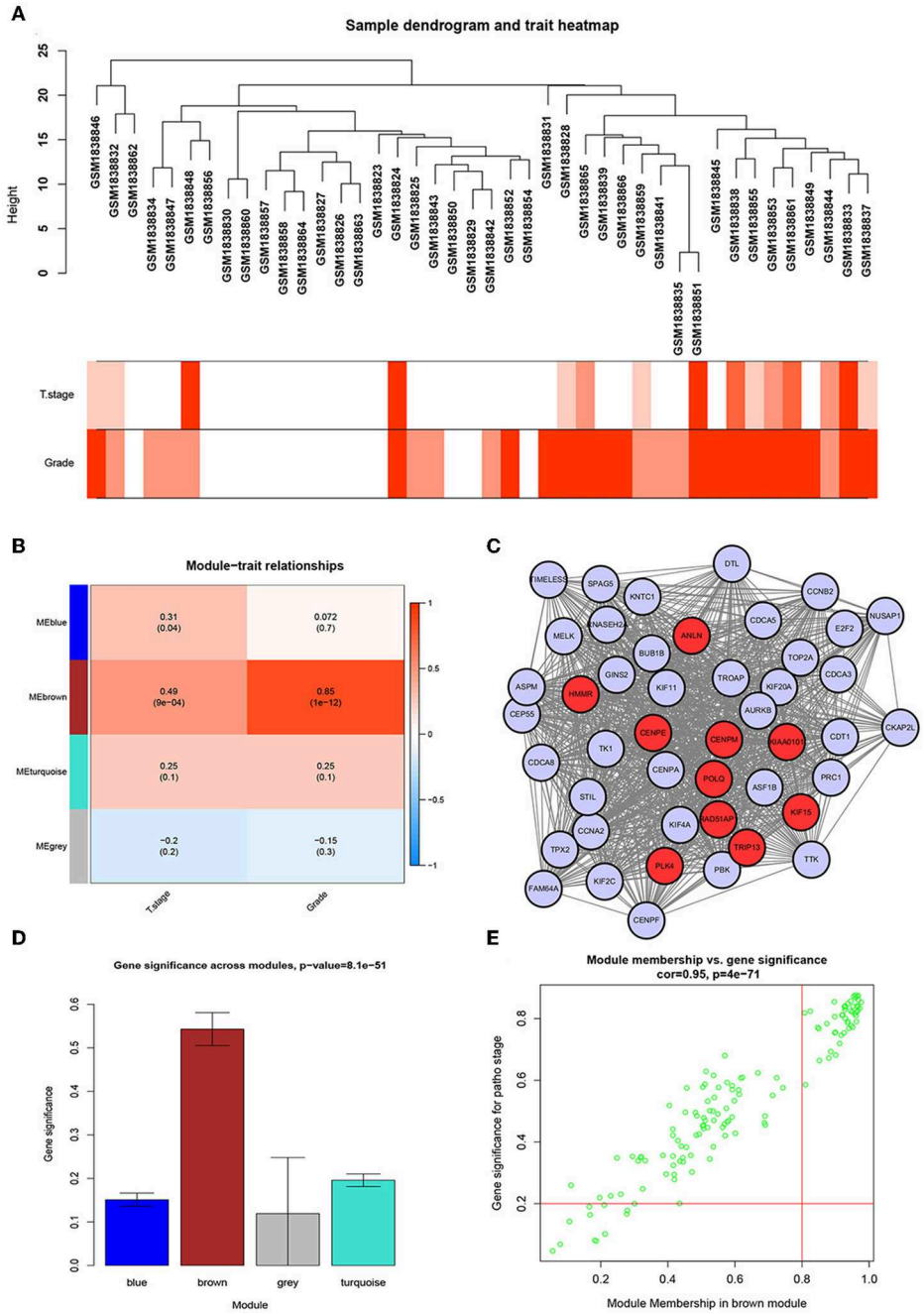


FIGURE 2 | WGCNA and PPI network analysis. **(A)** Sample dendrogram and trait indicator. The clustering was a visual result of calculations based on Pearson correlation coefficients between samples. The color intensity was proportional to T stage and histologic grade of BC. **(B)** Identification of modules associated with the clinical traits of BC. **(C)** PPI network of WGCNA hub genes, the red nodes represent the hub genes in the PPI network. **(D)** Distribution of average gene significance and errors in the modules associated with histologic grade of BC. **(E)** Scatter plot of module eigengenes related to histologic grade in the brown module.

heatmap of DEGs is shown in **Supplementary Figure S1**, and all DEGs are listed in **Supplementary Table S1**.

Establishment of Co-expression Network

We used the R package of "WGCNA" to construct the weighted co-expression network. No outlier samples were found by Pearson correlation analysis (**Figure 2A**). We put 792 DEGs with similar expression patterns into modules by cluster analysis. In this study, the power of $\beta = 6$ (scale-free $R^2 = 0.95$) was chosen for the soft-thresholding to ensure a scale-free network (**Supplementary Figures S2A–D**), and we got four modules for the next analysis (**Supplementary Figure S2E**).

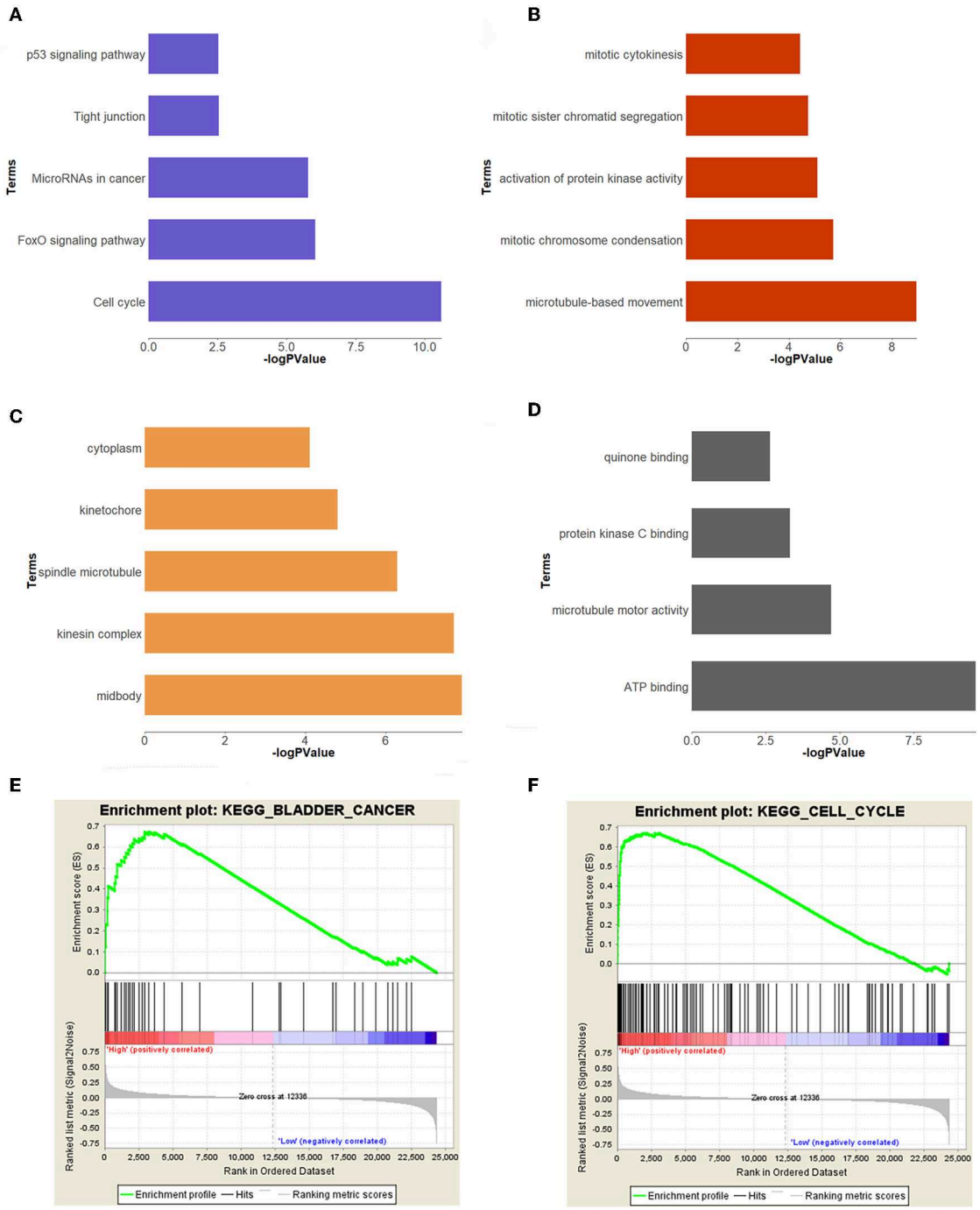


FIGURE 3 | Functional annotation and GSEA analysis for brown module. **(A)** The signaling pathways, **(B)** biological process, **(C)** cellular components, **(D)** molecular composition of the brown module. **(E,F)** GSEA analysis revealed that the genes of brown module were mainly enriched in bladder cancer and cell cycle related pathways.

Identification of the Most Significant Modules

To identify genes associated with the progression of BC, we analyzed the association between modules and clinical phenotypes. The modules most significantly associated with tumor grade and T stage are of great value in predicting BC progression. Histologic grade ($r = 0.85$, $p = 1e\text{-}12$) and T stage ($r = 0.49$, $p = 9e\text{-}04$, **Figure 2B**) were significantly associated with brown module by Module-feature relationship analysis. Besides, the brown module had the highest gene significance in relation

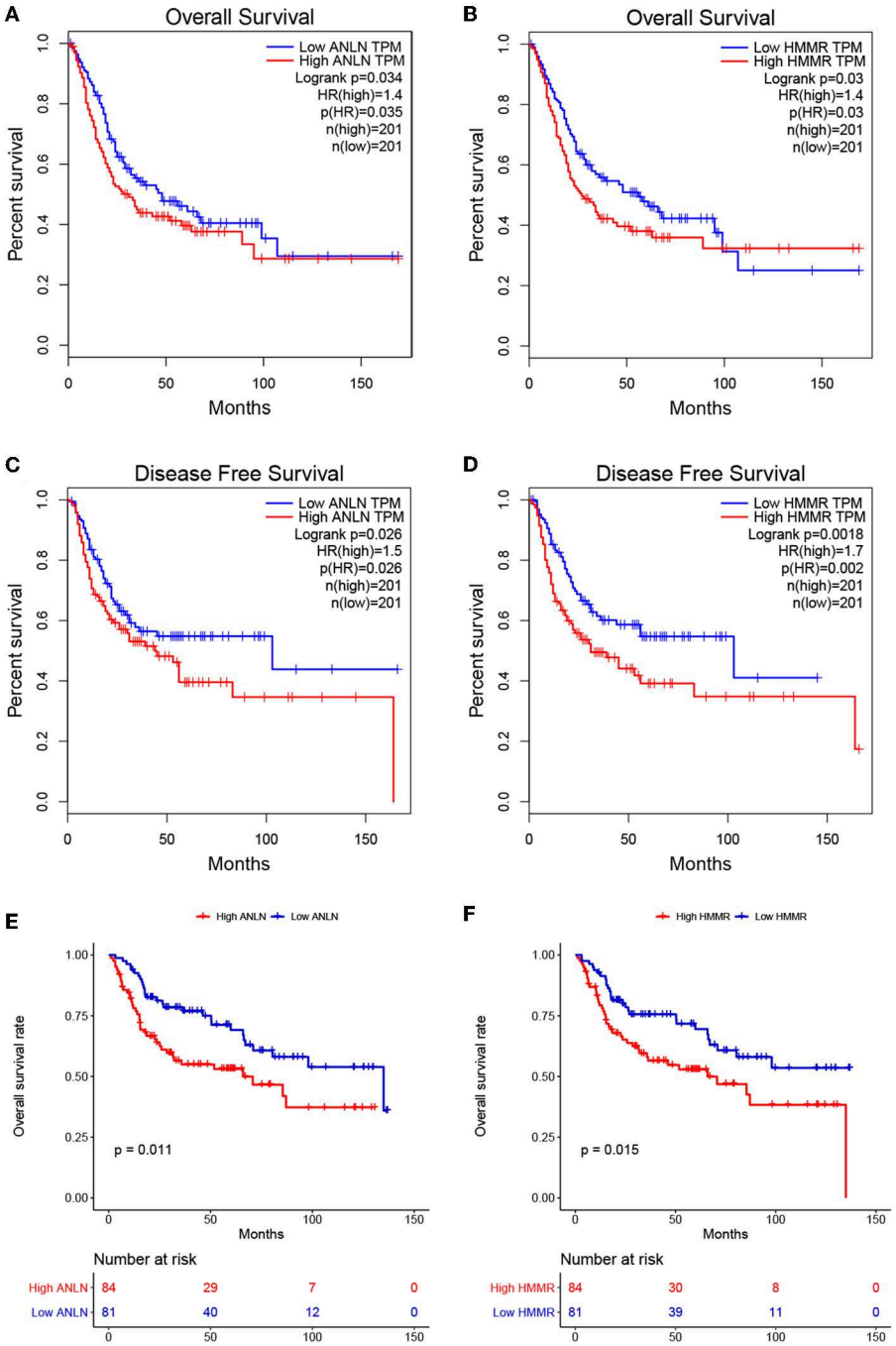


FIGURE 4 | Survival analyses on real hub genes in the TCGA and GEO database. **(A,B)** Overall survival analysis related to ANLN **(A)** or HMMR **(B)** expression levels in the TCGA database. **(C,D)** Disease-free survival analyses related to ANLN **(C)** or HMMR **(D)** expression levels in the TCGA database. **(E,F)** Overall survival analysis related to ANLN **(E)** or HMMR **(F)** in the GEO database (GSE13507).

to histologic grade (**Figure 2D**). Therefore, we chose the brown module for further analysis.

Brown Module Functional Annotation

In order to study the function of the brown module, we uploaded the list of all genes in the brown module to the DAVID (https://david.ncifcrf.gov) website for a functional annotation analysis. The KEGG analysis revealed that the "cell cycle," "FoxO signaling pathway," "Tight junction," "MicroRNAs in cancer," and "p53 signaling pathway" were mainly enriched in the brown module (**Figure 3A**). The biological process of the brown module was mainly related to "microtubule-based

TABLE 2 | Results of GSEA analysis based on the expression level of hub genes.

Group	Term	Enrichment score	NOM *p*-val
High ANLN/HMMR	KEGG_BLADDER_CANCER	0.674	0.004
	KEGG_CELL_CYCLE	0.671	0.010
	KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.431	0.012
	KEGG_HOMOLOGOUS_RECOMBINATION	0.717	0.014
	KEGG_RNA_DEGRADATION	0.485	0.022
	KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.495	0.026
	KEGG_BASE_EXCISION_REPAIR	0.643	0.032
	KEGG_MISMATCH_REPAIR	0.730	0.040
	KEGG_NUCLEOTIDE_EXCISION_REPAIR	0.613	0.045
Low ANLN/HMMR	KEGG_TYPE_II_DIABETES_MELLITUS	−0.475	0.004
	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	−0.520	0.014
	KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	−0.446	0.016
	KEGG_DILATED_CARDIOMYOPATHY	−0.618	0.018
	KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM	−0.606	0.026
	KEGG_HEMATOPOIETIC_CELL_LINEAGE	−0.647	0.027
	KEGG_HISTIDINE_METABOLISM	−0.675	0.030
	KEGG_TRYPTOPHAN_METABOLISM	−0.643	0.034
	KEGG_FOCAL_ADHESION	−0.580	0.047

movement," "mitotic chromosome condensation," "activation of protein kinase activity," and so on (**Figure 3B**). The cell component of brown module was mainly enriched in "midbody," "kinesin complex," "spindle microtubule," etc. (**Figure 3C**). And the molecular function was mainly enriched in "ATP binding," "microtubule motor activity," "protein kinase C binding," etc. (**Figure 3D**). The threshold was the $p < 0.05$. The information of functional annotation is listed in **Supplementary Table S2**.

Identification of Real Hub Genes

To further screen for the most significant hub genes, we combined three methods (WGCNA, PPI, and survival analysis) to screen real hub genes together. First, 49 hub genes with high connectivity were screened out from the brown module (**Figure 2E**). Secondly, we uploaded these 49 hub genes to the STRING database for a PPI network analysis. Under the threshold of a minimum required interaction score > 0.4, 10 hub PPI genes were screened (**Figure 2C**, **Supplementary Table S3**). Finally, we used the GEPIA database for the survival analysis of these 10 hub genes, and the hub genes with the ability to predict prognosis were real hub genes (ANLN, HMMR, **Supplementary Table S4**). The results showed that both real hub genes were predictive of overall survival and disease-free survival in BC (**Figures 4A–D**, **Supplementary Table S3**). Meanwhile, the external validation dataset GSE13507 was used to confirm the prognostic value of real hub genes (**Figures 4E,F**).

GSEA Analysis of Real Hub Genes

In order to explore the functions and pathways of these two hub genes, we conducted GSEA on these hub genes, respectively. The GSEA analysis of two hub genes in the GSE37815 dataset revealed that the samples of highly expressed real hub genes were mainly enriched in "bladder cancer," "cell cycle," and "ubiquitin mediated proteolysis" related pathways (**Figures 3E,F**, **Table 2**). Subsequently, our GSEA analysis in the TCGA database produced similar results (**Supplementary Tables S7–S9**).

Verification of the Expression Pattern of Real Hub Genes

Since these real hub genes were screened out by DEGs, we first verified the expression pattern of real hub genes between BC and paracancerous. The results showed that the expression of real hub genes was up-regulated in BC (**Supplementary Figure S3**), and the results were consistent in multiple datasets (Oncomine dataset, GSE13507, GSE37815, and TCGA dataset). Secondly, since the real hub genes belong to the brown module, which was significantly related to the histological grade and pathological stage of BC, the expression pattern of ANLN (**Figure 6**) and HMMR (**Figure 7**) in different histological grade and pathological stage were verified in internal validation datasets (GSE71576) and external validation datasets (GSE13507, GSE31684, and TCGA dataset). The one-way analysis of variance (ANOVA) or Student's *t*-test was used to measure the statistical significance of the calculated results. The results of receiver operating characteristic curve (ROC) analysis showed that real hub genes could well distinguish cancer and paracancer, different grades, different stages, NMIBC and MIBC (**Supplementary Table S5**). In addition, we verified the expression patterns of the protein levels of ANLN and HMMR in tissues in the HPA database, and found that the higher the grade of BC, the higher the protein levels of these two genes were (**Supplementary Figure S6**).

Validation of Prognostic Value of Real Hub Genes

To further explore the prognostic value of hub genes in BC, we conducted a subgroup survival analysis of these two genes

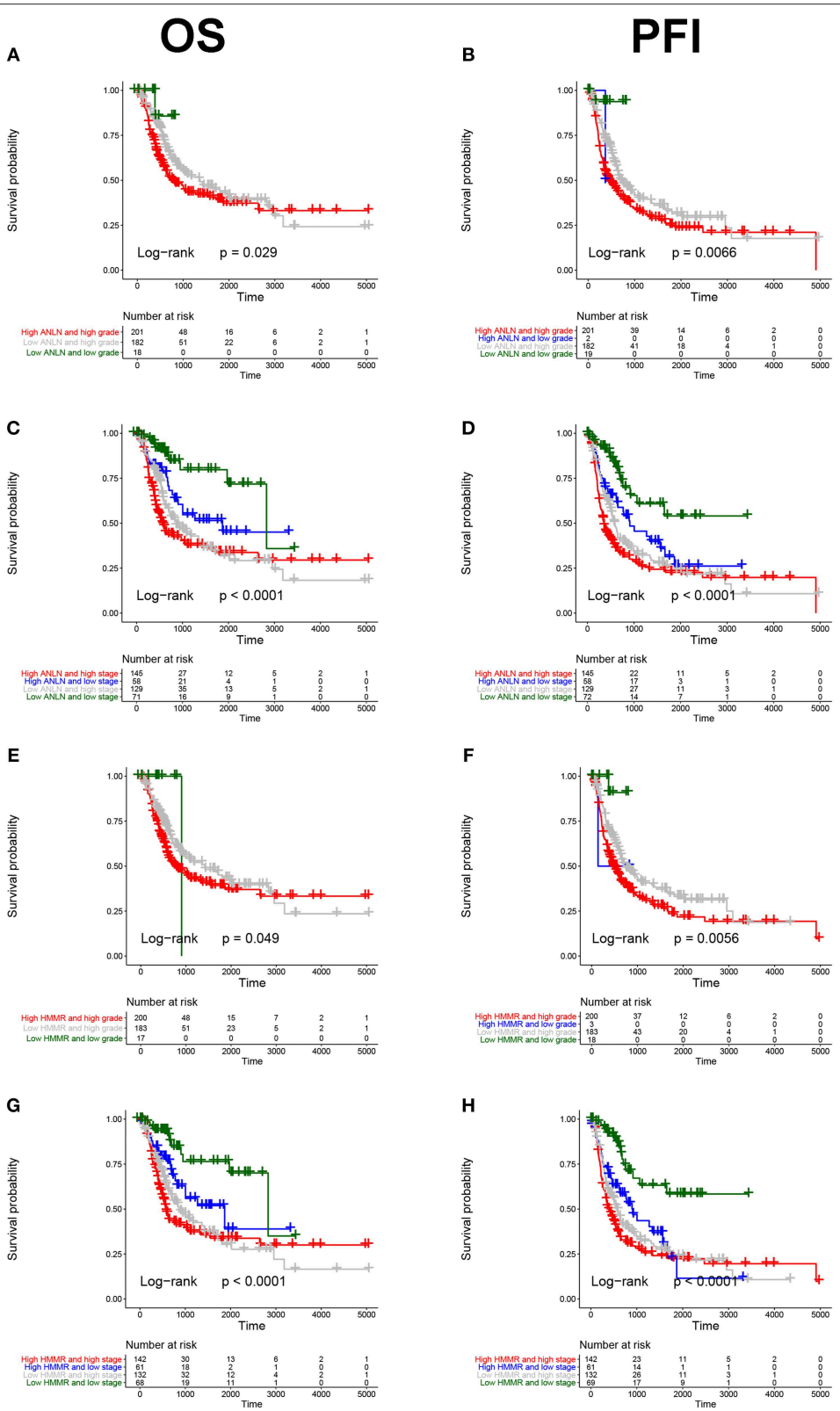


FIGURE 5 | Subgroup survival analysis of hub genes. Overall survival (left panels) and Progression-free interval analyses (right panels) of ANLN were performed in different tumor grade subtypes **(A,B)**; Again, we did the same analysis in tumor stage subtypes **(C,D)**; Corresponding to the previous analysis, we did the same analysis for HMMR **(E–H)**.

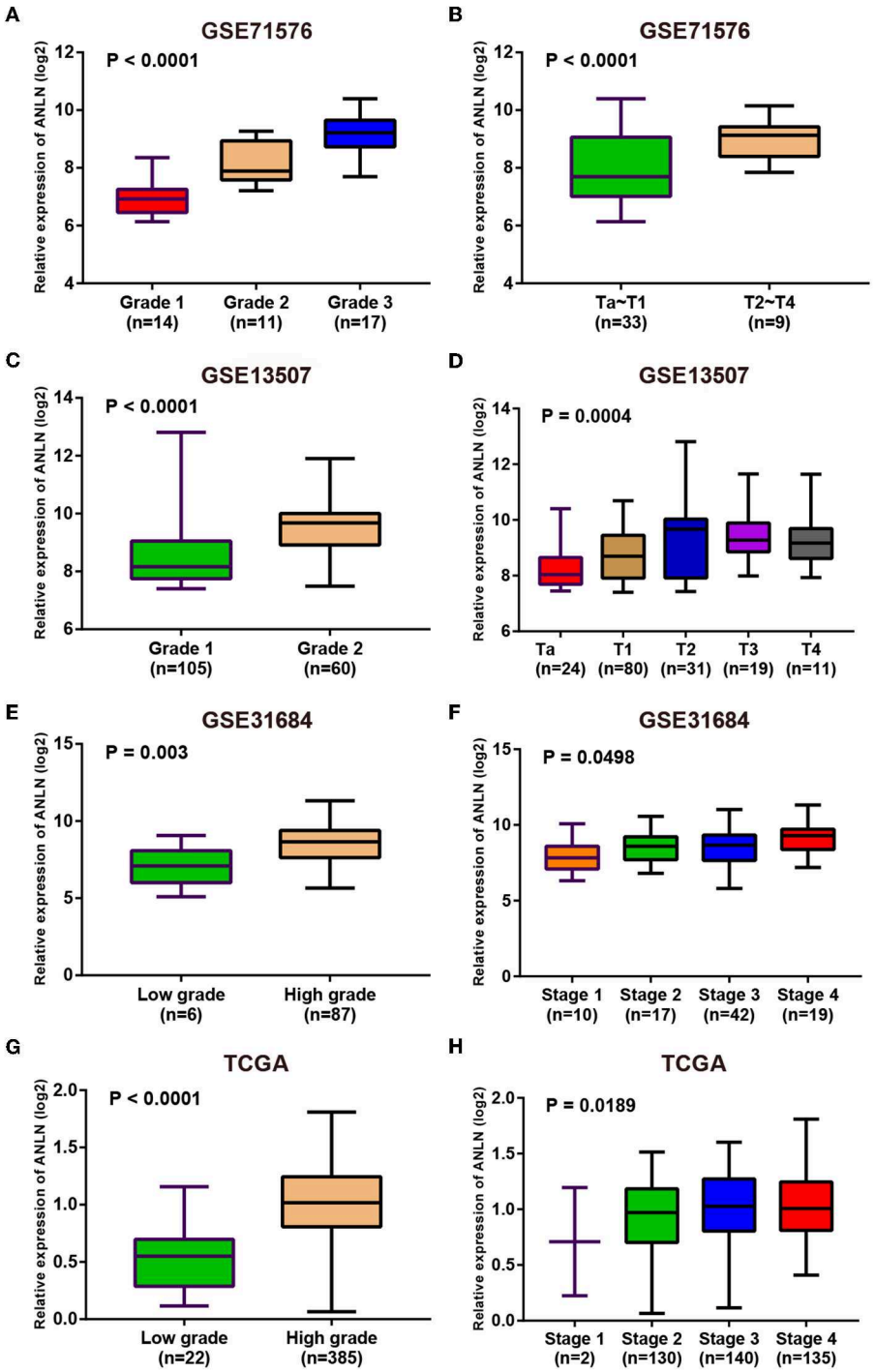


FIGURE 6 | Expression pattern validation for ANLN. ANLN in GSE71576 **(A,B)** GSE13507 **(C,D)** GSE31684 **(E,F)**, TCGA database **(G,H)** of different grade and stage of expressing validation. Statistical differences in these data were calculated using One-way analysis of variance (ANOVA) or Student's *t*-test.

in the TCGA dataset. The results showed that these two genes showed significant prognostic value in different stages and grades, which could not only accurately predict the overall survival rate of BC, but also predict its progression-free interval (PFI) event (**Figure 5**).

Drug Sensitivity of Real Hub Genes

GSCALite (http://bioinfo.life.hust.edu.cn/web/GSCALite/) is a web-based analysis platform for gene set cancer analysis (32). We used this database to analyze drug sensitivity of real hub genes, which provides

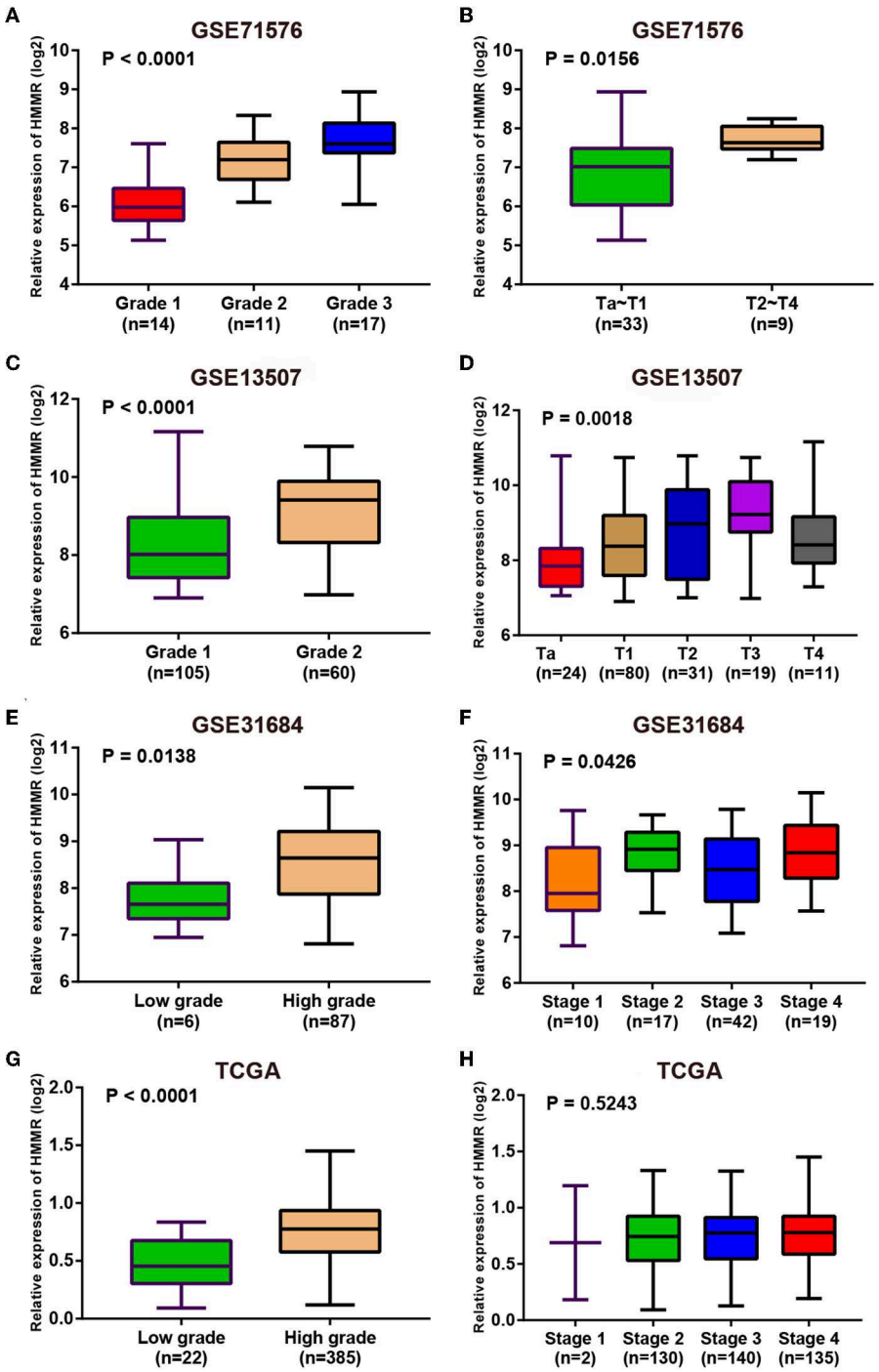


FIGURE 7 | Expression pattern validation for HMMR. HMMR in GSE71576 **(A,B)** GSE13507 **(C,D)** GSE31684 **(E,F)**, TCGA database **(G,H)** of different grade and stage of expressing validation. Statistical differences in these data were calculated using One-way analysis of variance (ANOVA) or Student's *t*-test.

support for drug selection of real hub genes targeted therapy.

Finally, we explored the drug sensitivity of real hub genes using the GSCALite database, and the results were shown in **Supplementary Figure S5**, which provides support for drug targeted therapy of real hub genes.

DISCUSSION

BC is one of the most common tumors of the urinary system. Currently, radical cystectomy is the most effective treatment for BC, but in most cases, this treatment will greatly reduce the quality of life of patients (33). Therefore, it is urgent to

find biomarkers that can accurately predict the progression and prognosis of BC.

Through a series of rigorous screening, two real hub genes (ANLN, HMMR) that could accurately predict the progression and prognosis of BC were found. Similar studies have focused mostly on one clinical phenotype (34–36). Our study conducted correlation analysis of T staging and grading as both clinical phenotypes and modules are of interest to us, and the results revealed that the brown module was highly correlated with both T staging ($r = 0.49$, $p = 9e\text{-}04$) and grading ($r = 0.85$, $p = 1e\text{-}12$). We then used a lot of datasets to verify this, and it turned out that the real hub genes were actually significantly correlated with BC T stage, pathological stage, and histological grade. Moreover, we also hoped to find biomarkers that could accurately predict the prognosis of BC, so we used survival analysis as a screening condition, which was neglected in some similar studies (37, 38). In the following two hub genes, we analyzed the survival analysis of the two hub genes and analyzed them in different subgroups (stage, grade), and found that these two genes had a high prognostic value for BC.

The excessive proliferation of tumors is often accompanied by cell cycle disorders. We used GSEA analysis to explore the function of real hub genes, and we found that both ANLN and HMMR were significantly enriched in functions and pathways related to "cell cycle." Correlation analysis also supports this result. These two genes were also enriched in the pathway related to "bladder cancer," and we speculate that these two genes may play a key role in the pathogenesis of BC.

ANLN (Anillin) is an actin-binding protein and has reportedly been shown to be significantly upregulated in the BC, knockdown of ANLN results in G2/M phase block and reduces expression of cyclin B1 and D1, and it was also demonstrated that ANLN can promote the progression, migration, and invasion of BC (39). Other studies have found that ANLN could promote the progression of pancreatic cancer by inducing the up-regulation of EZH2 by mediating the mir-218-5p /LASP1 signaling axis (40). ANLN has also been found to play a key role in the development of human lung cancer (41). All these suggest that ANLN plays a very important role in the development and progression of tumors. We found a high correlation between ANLN and CIRBP (**Supplementary Figure S4B**, **Supplementary Table S6**), a gene that we studied before (42); therefore, we can further explore the interaction between ANLN and CIRBP in the pathogenesis of BC. We also found a strong correlation between ANLN and KIF23 (**Supplementary Figure S4A**, **Supplementary Table S6**), an independent prognostic target for glioma (43).

HMMR (Hyaluronan Mediated Motility Receptor) is widely expressed in many types of tumors, including prostate and breast cancer, and various forms of leukemia (44–46). Previously reported overexpression of HMMR is associated with the development of metastatic prostate cancer (PCa) and castration-resistant PCa (46). But HMMR has never been studied in human BC, so our study found a new potential biomarker for BC. We found a strong correlation between HMMR and KIF20A (**Supplementary Figure S4C**, **Supplementary Table S6**), and a recent study found that KIF20A affects the prognosis of BC by promoting the proliferation and metastasis of BC (47). These studies are very helpful for our future research on the pathogenesis of HMMR in BC.

Taken together, through the integrated analysis of multiple databases and the establishment of the co-expression network by WGCNA analysis, two hub genes that can accurately predict the progression and prognosis of BC were screened out layer by layer, providing potential targets for the pathogenesis and treatment selection of BC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/.

AUTHOR CONTRIBUTIONS

YW, XW, and YX conceived and designed the study. YW and LC performed the analysis procedures. YW, LC, LJ, KQ, XL, and YX analyzed the results. YW, LC, and YX contributed analysis tools. YW, LC, XW, and YX contributed to the writing of the manuscript. All authors reviewed the manuscript.

FUNDING

This study was supported in part by the Zhongnan Hospital of Wuhan University Science, Technology and Innovation Seed Fund (grant number: cxpy2017028 and cxpy2017045), and Wuhan Clinical Cancer Research Center of Urology and Male Reproduction (grant number 303-230100055). The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The excellent technical assistance of Shanshan Zhang and Danni Shan is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.01030/full#supplementary-material

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. (2010) 127:2893–917. doi: 10.1002/ijc.25516
2. Kamat AM, Hahn NM, Efstathiou JA, Lerner SP, Malmstrom PU, Choi W, et al. Bladder cancer. *Lancet*. (2016) 388:2796–810. doi: 10.1016/S0140-6736(16)30512-8
3. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and

major patterns in GLOBOCAN 2012. *Int J Cancer*. (2015) 136:E359–86. doi: 10.1002/ijc.29210
4. Youssef RF, Lotan Y. Predictors of outcome of non-muscle-invasive and muscle-invasive bladder cancer. *ScientificWorldJournal*. (2011) 11:369–81. doi: 10.1100/tsw.2011.28
5. Humphrey PA, Moch H, Cubilla AL, Ulbright TM, Reuter VE. The 2016 WHO classification of tumours of the urinary system and male genital organs-part B: prostate and bladder tumours. *Eur Urol.* (2016) 70:106–19. doi: 10.1016/j.eururo.2016.02.028
6. Knowles MA, Hurst CD. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nat Rev Cancer*. (2015) 15:25–41. doi: 10.1038/nrc3817
7. Alfred Witjes J, Lebret T, Comperat EM, Cowan NC, De Santis M, Bruins HM, et al. Updated 2016 EAU guidelines on muscle-invasive and metastatic bladder cancer. *Eur Urol.* (2017) 71:462–75. doi: 10.1016/j.eururo.2016.06.020
8. Lodewijk I, Duenas M, Rubio C, Munera-Maravilla E, Segovia C, Bernardini A, et al. Liquid biopsy biomarkers in bladder cancer: a current need for patient diagnosis and monitoring. *Int J Mol Sci.* (2018) 19:E2514. doi: 10.3390/ijms19092514
9. Chamie K, Litwin MS, Bassett JC, Daskivich TJ, Lai J, Hanley JM, et al. Recurrence of high-risk bladder cancer: a population-based analysis. *Cancer*. (2013) 119:3219–27. doi: 10.1002/cncr.28147
10. Wolff EM, Liang G, Jones PA. Mechanisms of disease: genetic and epigenetic alterations that drive bladder cancer. *Nat Clin Pract Urol.* (2005) 2:502–10. doi: 10.1038/ncpuro0318
11. Burger M, Catto JW, Dalbagni G, Grossman HB, Herr H, Karakiewicz P, et al. Epidemiology and risk factors of urothelial bladder cancer. *Eur Urol.* (2013) 63:234–41. doi: 10.1016/j.eururo.2012.07.033
12. Gu P, Chen H. Modern bioinformatics meets traditional Chinese medicine. *Brief Bioinformatics*. (2014) 15:984–1003. doi: 10.1093/bib/bbt063
13. Li Q, Eichten SR, Hermanson PJ, Zaunbrecher VM, Song J, Wendt J, et al. Genetic perturbation of the maize methylome. *Plant Cell*. (2014) 26:4602–16. doi: 10.1105/tpc.114.133140
14. Huang MD, Huang AH. Bioinformatics reveal five lineages of oleosins and the mechanism of lineage evolution related to structure/function from green algae to seed plants. *Plant Physiol.* (2015) 169:453–70. doi: 10.1104/pp.15.00634
15. Turei D, Foldvari-Nagy L, Fazekas D, Modos D, Kubisch J, Kadlecsik T, et al. Autophagy regulatory network - a systems-level bioinformatics resource for studying the mechanism and regulation of autophagy. *Autophagy*. (2015) 11:155–65. doi: 10.4161/15548627.2014.994346
16. Omura S, Kawai E, Sato F, Martinez NE, Chaitanya GV, Rollyson PA, et al. Bioinformatics multivariate analysis determined a set of phase-specific biomarker candidates in a novel mouse model for viral myocarditis. *Circ Cardiovasc Genet*. (2014) 7:444–54. doi: 10.1161/CIRCGENETICS.114.000505
17. Ren J, Zhao G, Sun X, Liu H, Jiang P, Chen J, et al. Identification of plasma biomarkers for distinguishing bipolar depression from major depressive disorder by iTRAQ-coupled LC-MS/MS and bioinformatics analysis. *Psychoneuroendocrinology*. (2017) 86:17–24. doi: 10.1016/j.psyneuen.2017.09.005
18. Rong L, Huang W, Tian S, Chi X, Zhao P, Liu F. COL1A2 is a novel biomarker to improve clinical prediction in human gastric cancer: integrating bioinformatics and meta-analysis. *Pathol Oncol Res.* (2018) 24:129–34. doi: 10.1007/s12253-017-0223-5
19. Song E, Song W, Ren M, Xing L, Ni W, Li Y, et al. Identification of potential crucial genes associated with carcinogenesis of clear cell renal cell carcinoma. *J Cell Biochem.* (2018) 119:5163–74. doi: 10.1002/jcb.26543
20. Zhu QN, Renaud H, Guo Y. Bioinformatics-based identification of miR-542-5p as a predictive biomarker in breast cancer therapy. *Hereditas*. (2018) 155:17. doi: 10.1186/s41065-018-0055-7
21. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. (2004) 20:307–15. doi: 10.1093/bioinformatics/btg405
22. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* (2014) 15:550. doi: 10.1186/s13059-014-0550-8
24. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. (2008) 9:559. doi: 10.1186/1471-2105-9-559
25. Chen L, Yuan L, Qian K, Qian G, Zhu Y, Wu CL, et al. Identification of biomarkers associated with pathological stage and prognosis of clear cell renal cell carcinoma by co-expression network analysis. *Front Physiol.* (2018) 9:399. doi: 10.3389/fphys.2018.00399
26. Botia JA, Vandrovcova J, Forabosco P, Guelfi S, D'Sa K, United Kingdom Brain Expression Consortium, et al. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Syst Biol.* (2017) 11:47. doi: 10.1186/s12918-017-0420-6
27. Chen L, Yuan L, Wang Y, Wang G, Zhu Y, Cao R, et al. Co-expression network analysis identified FCER1G in association with progression and prognosis in human clear cell renal cell carcinoma. *Int J Biol Sci.* (2017) 13:1361–72. doi: 10.7150/ijbs.21657
28. Yuan L, Zeng G, Chen L, Wang G, Wang X, Cao X, et al. Identification of key genes and pathways in human clear cell renal cell carcinoma (ccRCC) by co-expression analysis. *Int J Biol Sci.* (2018) 14:266–79. doi: 10.7150/ijbs.23574
29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* (2015) 43:D447–52. doi: 10.1093/nar/gku1003
30. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics.* (2014) 47, 11–24. doi: 10.1002/0471250953.bi0813s47
31. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* (2017) 45:W98–102. doi: 10.1093/nar/gkx247
32. Liu CJ, Hu FF, Xia M, Han L, Zhang Q, Guo AY. GSCALite: a web server for gene set cancer analysis. *Bioinformatics*. (2018) 34:3771–2. doi: 10.1093/bioinformatics/bty411
33. Stein JP, Lieskovsky G, Cote R, Groshen S, Feng AC, Boyd S, et al. Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1,054 patients. *J Clin Oncol.* (2001) 19:666–75. doi: 10.1200/JCO.2001.19.3.666
34. He Z, Sun M, Ke Y, Lin R, Xiao Y, Zhou S, et al. Identifying biomarkers of papillary renal cell carcinoma associated with pathological stage by weighted gene co-expression network analysis. *Oncotarget*. (2017) 8:27904–14. doi: 10.18632/oncotarget.15842
35. Huang H, Zhang Q, Ye C, Lv JM, Liu X, Chen L, et al. Identification of prognostic markers of high grade prostate cancer through an integrated bioinformatics approach. *J Cancer Res Clin Oncol.* (2017) 143:2571–9. doi: 10.1007/s00432-017-2497-0
36. Tian H, Guan D, Li J. Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (WGCNA). *Medicine.* (2018) 97:e10781. doi: 10.1097/MD.0000000000010781
37. Yuan L, Shu B, Chen L, Qian K, Wang Y, Qian G, et al. Overexpression of COL3A1 confers a poor prognosis in human bladder cancer identified by co-expression analysis. *Oncotarget*. (2017) 8:70508–20. doi: 10.18632/oncotarget.19733
38. Zhou XG, Huang XL, Liang SY, Tang SM, Wu SK, Huang TT, et al. Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis. *Onco Targets Ther.* (2018) 11:2815–30. doi: 10.2147/OTT.S163891
39. Zeng S, Yu X, Ma C, Song R, Zhang Z, Zi X, et al. Transcriptome sequencing identifies ANLN as a promising prognostic biomarker in bladder urothelial carcinoma. *Sci Rep.* (2017) 7:3151. doi: 10.1038/s41598-017-02990-9
40. Wang A, Dai H, Gong Y, Zhang C, Shu J, Luo Y, et al. ANLN-induced EZH2 upregulation promotes pancreatic cancer progression by mediating miR-218-5p/LASP1 signaling axis. *J Exp Clin Cancer Res.* (2019) 38:347. doi: 10.1186/s13046-019-1340-7
41. Suzuki C, Daigo Y, Ishikawa N, Kato T, Hayama S, Ito T, et al. ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway. *Cancer Res.* (2005) 65:11314–25. doi: 10.1158/0008-5472.CAN-05-1507

42. Lu M, Ge Q, Wang G, Luo Y, Wang X, Jiang W, et al. CIRBP is a novel oncogene in human bladder cancer inducing expression of HIF-1alpha. *Cell Death Dis*. (2018) 9:1046. doi: 10.1038/s41419-018-1109-5
43. Sun L, Zhang C, Yang Z, Wu Y, Wang H, Bao Z, et al. KIF23 is an independent prognostic biomarker in glioma, transcriptionally regulated by TCF-4. *Oncotarget*. (2016) 7:24646–55. doi: 10.18632/oncotarget.8261
44. Greiner J, Schmitt M, Li L, Giannopoulos K, Bosch K, Schmitt A, et al. Expression of tumor-associated antigens in acute myeloid leukemia: implications for specific immunotherapeutic approaches. *Blood*. (2006) 108:4109–17. doi: 10.1182/blood-2006-01-023127
45. Kalmyrzaev B, Pharoah PD, Easton DF, Ponder BA, Dunning AM, Team S. Hyaluronan-mediated motility receptor gene single nucleotide polymorphisms and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev*. (2008) 17:3618–20. doi: 10.1158/1055-9965.EPI-08-0216
46. Gust KM, Hofer MD, Perner SR, Kim R, Chinnaiyan AM, Varambally S, et al. RHAMM (CD168) is overexpressed at the protein level and may constitute an immunogenic antigen in advanced prostate cancer disease. *Neoplasia*. (2009) 11:956–63. doi: 10.1593/neo.09694
47. Shen T, Yang L, Zhang Z, Yu J, Dai L, Gao M, et al. KIF20A Affects the prognosis of bladder cancer by promoting the proliferation and metastasis of bladder cancer cells. *Dis Markers*. (2019) 2019:4863182. doi: 10.1155/2019/4863182

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

CASE REPORT
published: 22 October 2019
doi: 10.3389/fgene.2019.00952

A Pediatric Case of Glioblastoma Multiforme Associated With a Novel Germline p.His112CysfsTer9 Mutation in the *MLH1* Gene Accompanied by a p.Arg283Cys Mutation in the *TP53* Gene: A Case Report

*Aleksandra Stajkovska[1†], Sanja Mehandziska[2†], Rodney Rosalia[2], Margarita Stavrevska[2], Marija Janevska[1], Martina Markovska[1], Ivan Kungulovski[1], Zan Mitrev[2] and Goran Kungulovski[1]**

[1] Sector of Genetics, Bio Engineering LLC, Skopje, Macedonia,[2] Laboratory of Genetics and Personalized Medicine, Zan Mitrev Clinic, Skopje, Macedonia

OPEN ACCESS

Edited by:
Barbara Karen Dunn,
National Institutes of Health (NIH),
United States

Reviewed by:
Elena Tosti,
Albert Einstein College of Medicine,
United States
Jun Zhong,
National Cancer Institute (NCI),
United States

***Correspondence:**
Goran Kungulovski
goran@bioengineering.mk;
goran.kungulovski@zmc.mk

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: 20 February 2019
Accepted: 05 September 2019
Published: 22 October 2019

Citation:
Stajkovska A, Mehandziska S, Rosalia R, Stavrevska M, Janevska M, Markovska M, Kungulovski I, Mitrev Z and Kungulovski G (2019) A Pediatric Case of Glioblastoma Multiforme Associated With a Novel Germline p.His112CysfsTer9 Mutation in the MLH1 Gene Accompanied by a p.Arg283Cys Mutation in the TP53 Gene: A Case Report. Front. Genet. 10:952. doi: 10.3389/fgene.2019.00952

Targeted gene panel testing has the power to interrogate hundreds of genes and evaluate the genetic risk for many types of hereditary cancers simultaneously. We screened a 13-year-old male patient diagnosed with glioblastoma multiforme with the aim to get further insights into the biology of his condition. Herein, we applied gene panel sequencing and identified a heterozygous frameshift mutation c.333_334delTC; p.His112CysfsTer9 in the *MLH1* gene in blood and tumor tissue accompanied by a known heterozygous missense variant of unknown significance c.847C > T; p.Arg283Cys in the *TP53* gene. Parental screening revealed the presence of the same *TP53* variant in the father and the same *MLH1* variant in the mother, who was in fact undergoing treatment for early-stage breast cancer at the time of her son's unfortunate diagnosis. This case reports for the first time the co-occurrence of a genetic mutation in the *MLH1* gene of the mismatch repair pathway, commonly associated with the Lynch syndrome, accompanied by a rare variant in the *TP53* gene. This report underlines the need for broad panel gene testing in lieu of single-gene or syndrome-focused gene screening and evaluation of the effects of multiple pathogenic or modifier variants on the phenotypic spectrum of the disease.

Keywords: next-generation sequencing, North Macedonia, hereditary cancer syndromes, Lynch syndrome, *TP53*, *MLH1*, Li-Fraumeni, case report

BACKGROUND

MMR-Dependent Hereditary Cancer Syndromes

The DNA mismatch repair (MMR) machinery is a highly conserved cell-intrinsic fail-safe system that recognises and repairs mismatched bases emanating from spurious DNA replication, recombination, or chemical/physical insults (Richman, 2015). A malfunction of the MMR machinery may lead to microsatellite instability, which in turn increases the rate of mutations.

By virtue of this, germline genetic mutations in the *MLH1*, *MSH2 (*or through *EPCAM*), *MSH6*, and *PMS2* genes lead to an inborn functional deficiency of the MMR pathway, thereby significantly increasing the risk of cancer. Mutations in the MMR pathway are associated with hereditary cancer syndromes such as Lynch syndrome, Turcot syndrome, and Muir–Torre syndrome.

Turcot syndrome (OMIM 276300) is a disease that manifests *via* multiple adenomatous colon polyps; patients have an increased risk of colorectal cancer and brain cancers, namely glioblastoma. Turcot syndrome typically follows an autosomal dominant inheritance pattern. It is closely associated with other rare hereditary cancers, such as familial adenomatous polyposis or Lynch syndrome (Hegde et al., 2014; Khattab and Monga, 2019).

Genetic mutations in *APC* gene associated with familial adenomatous polyposis, or a mutation in one of the MMR genes, the *MLH* gene in particular associated with Lynch syndrome, form the molecular basis for most cases of Turcot syndrome (Carethers and Stoffel, 2015). There is a dichotomous trend observed in regard to the etiology and clinical presentation of the hereditary brain cancers; *APC* mutations typically trigger an oncogenic pathway leading to medulloblastoma. In contrast, mutations in the MMR machinery usually lead to glioblastoma multiforme (GBM) (Alifieris and Trafalis, 2015), a devastating brain cancer; diagnosis of GBM is associated with a dire clinical outcome in the majority of cases. Despite aggressive combinatorial therapy, survival of (adults) ranges between 8 and 18 months, depending on the extent of the disease (Kohlmann and Gruber, 1993; Sehgal et al., 2014; Stepanenko and Chekhonin, 2018).

TP53-Dependent Hereditary Cancer Syndrome

TP53 is a tumor-suppressor gene, encoding the p53 protein, which has a crucial role in the regulation of cell proliferation (Wawryk-Gawda et al., 2014). In particular, p53 regulates apoptosis, genomic stability, and angiogenesis (Pentimalli, 2018). Li-Fraumeni syndrome (LFS) (OMIM 151623) is a rare disorder, inherited in an autosomal dominant manner, caused by germline mutations in the *TP53* gene. Mutations that lead to suboptimal function or total loss of function of the p53 lead to compromised tumor suppression and cell proliferation. Consequently, individuals with dysfunctional p53 are highly susceptible to a broad range of cancers (Olivier et al., 2010).

The tumors most closely associated with LFS are so-called "core" cancers; brain cancers form part of this group of LFS-associated malignancies (Malkin, 2011; Sorrell et al., 2013; McBride et al., 2014; Kratz et al., 2017). The Chompret criteria have been proposed for the screening of patients suspected for LFS (Tinat et al., 2009).

Individuals with LFS are eligible for treatment; a personalized approach is required according to the intrinsic properties of the tumor. In addition, caution is warranted due to the known adverse effects of conventional radiotherapy. Several (pre-) clinical studies have shown an increased risk for radiation-induced cancers in LFS patients.

CASE PRESENTATION

The proband was a 13-year-old boy, who was referred for genetic testing due to a suspected hereditary cancer syndrome, following a diagnosis of GBM WHO grade IV. He previously underwent surgical resection combined with adjuvant temozolomide chemotherapy. The father reported no family history of cancer; however, the mother was diagnosed with breast cancer at the age of 56 and underwent a bilateral mastectomy. Furthermore, the maternal grandmother was also diagnosed with breast cancer at old age. Upon taking written and signed informed consent from the proband's legal guardians, gene panel sequencing revealed a novel heterozygous frameshift mutation c.333_334delTC; p.His112CysfsTer9 in the *MLH1* gene and another known, but rare, heterozygous missense VUS c.847C > T; p.Arg283Cys in the *TP53* gene **(Table 1; Figure 1).**

The mutation in the *MLH1* gene results in a truncated protein and most likely leads to loss of function, predisposing carriers to hereditary malignant syndromes, for example Turcot syndrome or the related Lynch syndrome. The discovered variant in the *TP53* gene meets PM1, PP3, and PP5 ACMG pathogenicity criteria (Richards et al., 2015); in the ClinVar database, it is annotated as a *variant of uncertain significance*—albeit with probable functional relevance. The presence of the *MLH1* mutation was validated independently in blood and tumor tissue with next-generation sequencing (NGS) (130 brain tumor-relevant genes) and Sanger sequencing. In addition, DNA methylation analysis of the tumor tissue in comparison with a reference database of 2,800 tumors, categorized the tumor in the methylation class glioblastoma, isocitrate dehydrogenase wild type, subtype RTK III. These data indicated *MGMT* promoter methylation and potential loss of *CDKN2A*.

Moreover, we evaluated the NGS result and inheritance pattern by Sanger sequencing—we concluded that the *TP53* mutation was inherited from the paternal side, while the

TABLE 1 | Properties of detected variants, symptoms and medical history.

Method	*Gene*	Nucleotide change	Protein change	rsID	MAF	Clinvar sig	ACMG	Novelty	Inheritance	Symptoms and family history
TruSight Cancer; MMR sequencing; Sanger sequencing	*MLH1*	c.333_334delTC	p.His112CysfsTer9	/	/	/	PVS1, PM1, PM2	Unknown	Dominant	**Proband**: Glioblastoma multiforme, age 12 **Mother**: breast cancer, age 56, and bilateral mastectomy **Grandmother**: breast cancer
	TP53	c.847C > T	p.Arg283Cys	rs149633775	0.00008	Uncertain significance	PM1, PP3, PP5	Known VUS	Dominant	

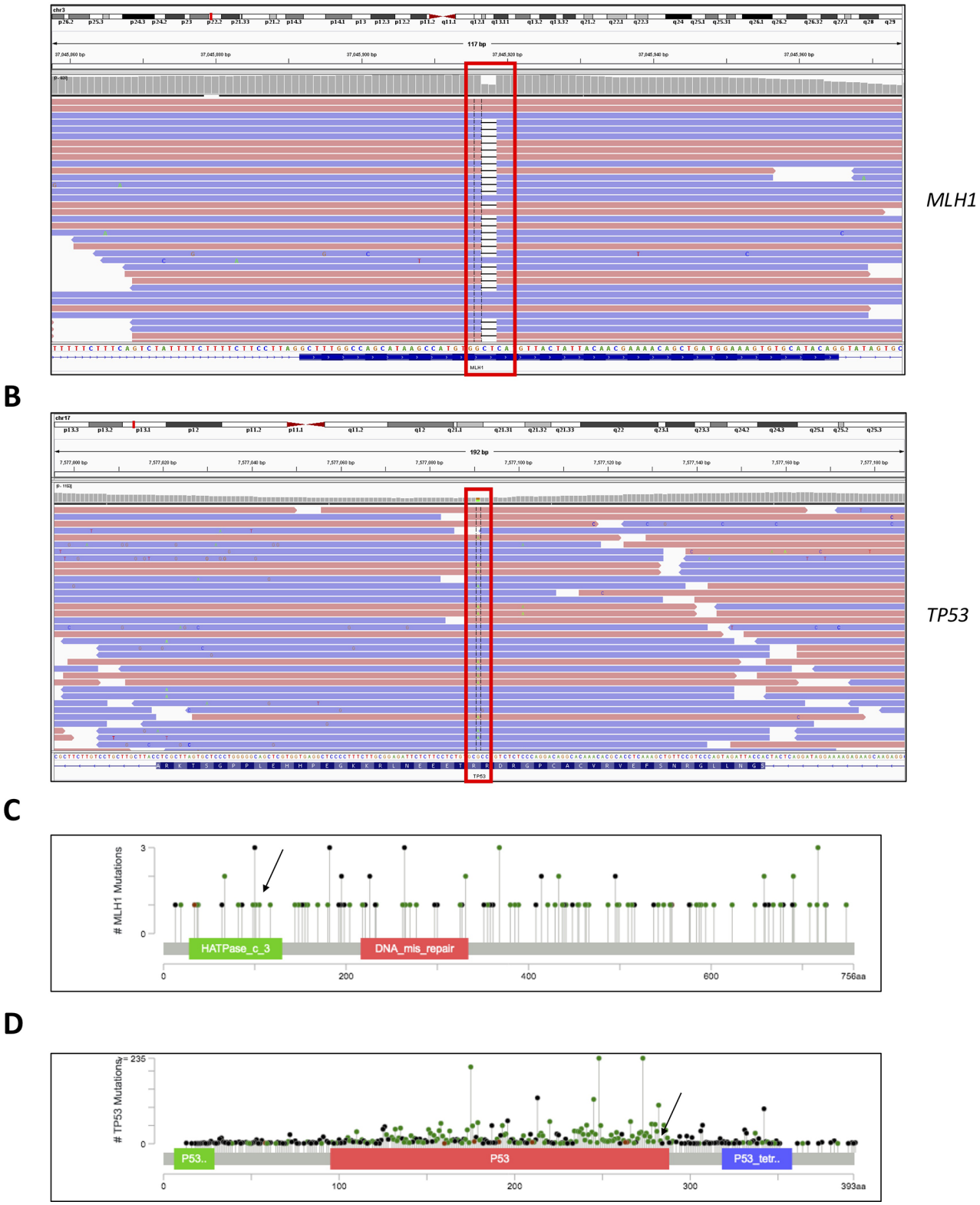


FIGURE 1 | Genome browser views of the **(A)** *MLH1* gene showing sequence alignment around the region of the detected c.333_334delTC mutation and **(B)** *TP53* gene showing sequence alignment around the region of the detected c.847C > T mutation. **(C)** Lollipop scheme of the *MLH1* gene, pinpointing the region of the detected mutation. **(D)** Lollipop scheme of the *TP53* gene, pinpointing the region of the detected mutation.

MLH1 mutation was inherited from the maternal side. Sanger sequencing of a maternal cousin failed to detect the *TP53* and *MLH1* mutations **(Figure 2)**.

Laboratory Tests

The patient and his family underwent gene panel sequencing (TruSight Cancer, Illumina) or Sanger sequencing, respectively. In brief, DNA was extracted from 400 µl of whole blood on a SaMag-12 automatic nucleic acid extraction system (Sacace Biotechnologies, Como, Italy). Libraries were prepared following the manufacturer's recommendations, and raw sequences were obtained from the NextSeq machine (Illumina, San Diego, USA). Sequence quality control, single nucleotide polymorphism, and insertion/deletion calling, together with advanced variant annotation were done with proprietary technologies such as the Sophia DDM platform (Sophia Genetics, Saint-Sulpice, Switzerland). The NGS panel of brain-tumor-relevant genes, the DNA methylation analysis with the 850K Illumina array, and methylation classification (internal classifier V11b2) were carried out at the University Clinic in Heidelberg.

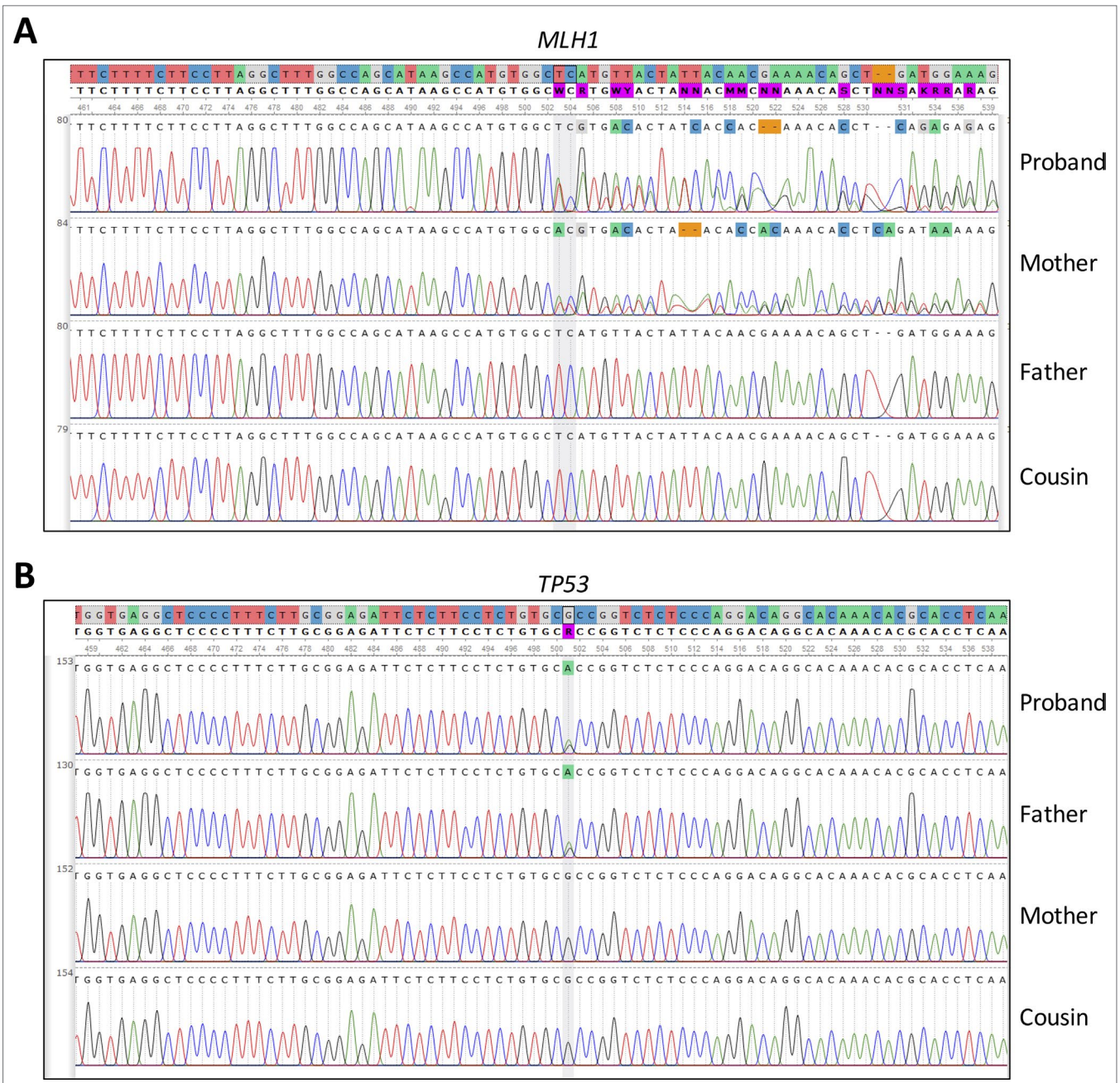


FIGURE 2 | Sanger sequencing of **(A)** *MLH1* gene, pinpointing the presence or absence of the c.333_334delTC mutation in the patient and family members and **(B)** *TP53* gene, pinpointing the presence or absence of the c.847C > T mutation detected by in the patient and family members.

DISCUSSION AND CONCLUDING REMARKS

In this case report, we used gene panel sequencing to evaluate the hereditary cancer risk of a pediatric GBM patient and his family and to get an insight into the biology of the tumor.

Through this approach, we identified two heterozygous highly probable disease-causing mutations: c.333_334delTC; p.His112CysfsTer9 in the *MLH1* gene, and c.847C > T; p.Arg283Cys in the *TP53* gene. The proband inherited the pathogenic variant in the *MLH1* gene from his mother. We hypothesize that the frameshift mutation in the *MLH1* gene is most likely the primary oncogenic driver and the main culprit causing cancer proclivity in this pediatric case of GBM.

Given the rare clinical presentation and absence of abdominal symptoms, the patient was never suspected of Turcot syndrome or Lynch syndrome. Hence, colonoscopy screening was never performed. We are unable to exclude the presence of other primary (pre-)malignant lesions at distal sites, which could have strengthened the diagnosis. However, the hereditary genetic profile strongly points to the aforementioned cancer syndromes manifesting as GBM.

GBM (Alifieris and Trafalis, 2015) is an epithelial tumor of the central nervous system with frequent genetic and epigenetic alterations (Heiland et al., 2017) and a worldwide incidence of <10% that commonly manifests as a solitary lesion; multiple GBM lesions are rare. GBM manifests in adults between the age of 45 and 70 years old (Zhang et al., 2016). Conversely, our patient developed aggressive intracranial malignancy at a very young age, prompting the suspicion of multiple oncogenic or modifier mutations. Indeed, further analysis uncovered the paternally inherited mutation, c.847C > T; p.Arg283Cys, in the *TP53* gene.

We speculate that the p.Arg283Cys variant in *TP53* served as an additional oncogenic driver or modifier, resulting in the unusually early onset of GBM. There are several lines of evidence supporting this claim. First, Monti et al. showed that the p.Arg283Cys variant, among other *TP53* germline variants, showed severe deficiency to transactivate *MDM2*, *BAX*, and *PUMA,* but not *CDKN1A*, in a luciferase-based quantitative assay in yeast, when compared to the wild-type allele (Monti et al., 2011). Similarly, another *in vitro* functional study indicated that the germ-line p.Arg283Cys variant could still transactivate the *CDKN1A* but not the *BAX* gene and thus retained the ability to induce growth arrest of human glioblastoma cells (Fulci et al., 2002). Furthermore, it has been shown that the p.Arg283Cys p53 protein is cold sensitive and unable to activate p53-RE placed upstream of the *ADE2* reporter in yeast (Jagosova et al., 2012). These studies illustrate that the c.847C > T; p.Arg283Cys mutation, unlike other highly pathogenic mutations in the *TP53* gene (Pavletich et al., 1993; Muller and Vousden, 2014; Shajani-Yi et al., 2018), causes a partial loss of function with unclear clinical repercussions.

Second, in genetic studies, the p.Arg283Cys mutation was identified together with a nonsense variant in *BRCA2* in a patient with metachronous breast cancers and a subsequent leiomyosarcoma, with a family history of ovarian cancer, breast and ovarian cancer, and glioblastoma (Manoukian et al., 2007). In addition, the c.847C > T; p.Arg283Cys mutation was the only variant detected in a *CDH1* negative gastric cancer patient, with a family history of gastric cancer, leukemia, and liver cancer (Yurgelun et al., 2015).

Collectively, our current observations and published reports provide circumstantial evidence for the functional relevance of the p.Arg283Cys, *TP53* variant, but further studies are necessary to substantiate this claim. Moreover, functional studies are warranted to evaluate the ramifications of the co-occurrence of *MLH1* loss-of-function mutations and the p.Arg283Cys mutation in *TP53*.

Targeted gene panel testing of known cancer-associated genes is a cost-effective diagnostic tool to simultaneously evaluate patients and their relatives suspected of hereditary malignant syndromes. As genetic testing is getting readily available to an increasing number of institutions, we anticipate that the number of similar cases will increase. This is a cautionary tale for clinicians, medical geneticists, and genetic counselors to take into account the possibility of a patient having two or more disease-causing or disease-modifying variants, which might influence the severity, tissue specificity, and onset of the disease (Cohen et al., 2016)

DATA AVAILABILITY STATEMENT

FASTQ data have been deposited to the NCBI under the accession number PRJNA516553 (https://www.ncbi.nlm.nih.gov/sra/PRJNA516553).

ETHICS STATEMENT

Written and signed informed consent was obtained from all subjects or their legal guardians for participation and publication of this case study.

AUTHOR CONTRIBUTIONS

GK conceived and designed this case study. GK processed, analyzed, and interpreted the sequencing data with the help of AS, SM, MS, MM, and MJ. GK, AS, and SM contributed to genetic counseling. RR, IK, and ZM contributed to the recruitment of patients in the hospital and contributed intellectually. GK and RR wrote the manuscript. AS and SM contributed equally to the manuscript. All authors contributed to the improvement of the manuscript and read the final version of the manuscript.

FUNDING

The authors declare that this study has not received any funding. It was carried out as part of the routine clinical work at the Zan Mitrev Clinic.

ACKNOWLEDGMENTS

We thank the family for participating in this study. We are also grateful to Steffen Hirsch from the University Clinic in Heidelberg for providing additional molecular information of the patient.

REFERENCES

Alifieris, C., and Trafalis, D. T. (2015). Glioblastoma multiforme: pathogenesis and treatment. *Pharmacol. Ther.* 152, 63–82. doi: 10.1016/j.pharmthera.2015.05.005

Carethers, J. M., and Stoffel, E. M. (2015). Lynch syndrome and Lynch syndrome mimics: the growing complex landscape of hereditary colon cancer. *World J. Gastroenterol.* 21, 9253–9261. doi: 10.3748/wjg.v21.i31.9253

Cohen, S. A., Tan, C. A., and Bisson, R. (2016). An individual with both MUTYH-associated polyposis and Lynch syndrome identified by multi-gene hereditary cancer panel testing: a case report. *Front. Genet.* 7, 36. doi: 10.3389/fgene.2016.00036

Fulci, G., Ishii, N., Maurici, D., Gernert, K. M., Hainaut, P., Kaur, B., et al. (2002). Initiation of human astrocytoma by clonal evolution of cells with progressive loss of p53 functions in a patient with a 283H TP53 germ-line mutation: evidence for a precursor lesion. *Cancer Res.* 62, 2897–2905.

Hegde, M., Ferber, M., Mao, R., Samowitz, W., Ganguly, A., and Working Group of the American College of Medical Genetics and Genomics (ACMG) Laboratory Quality Assurance Committee. (2014). ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *Genet Med.* 16, 101–116. doi: 10.1038/gim.2013.166

Heiland, D. H., Haaker, G., Delev, D., Mercas, B., Masalha, W., Heynckes, S., et al. (2017). Comprehensive analysis of PD-L1 expression in glioblastoma multiforme. *Oncotarget* 8, 42214–42225. doi: 10.18632/oncotarget.15031

Jagosova, J., Pitrova, L., Slovackova, J., Ravcukova, B., Smarda, J., and Smardova, J. (2012). Transactivation and reactivation capabilities of temperature-dependent p53 mutants in yeast and human cells. *Int. J. Oncol.* 41, 1157–1163. doi: 10.3892/ijo.2012.1520

Khattab, A., and Monga, D. K. (2019). *Turcot syndrome*. Treasure Island (FL): StatPearls Publishing.

Kohlmann, W., and Gruber, S. B. (1993). "Lynch syndrome," in *GeneReviews((R))*. Eds. M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, K. Stephens, and A. Amemiya. Seattle (WA): University of Washington. https://www.ncbi.nlm.nih.gov/books/NBK1211/

Kratz, C. P., Achatz, M. I., Brugieres, L., Frebourg, T., Garber, J. E., Greer, M. C. et al. (2017). Cancer screening recommendations for individuals with Li-Fraumeni syndrome. *Clin. Cancer Res.* 23, e38–e45. doi: 10.1158/1078-0432.CCR-17-0408

Malkin, D. (2011). Li-Fraumeni syndrome. *Genes Cancer* 2, 475–484. doi: 10.1177/1947601911413466

Manoukian, S., Peissel, B., Pensotti, V., Barile, M., Cortesi, L., Stacchiotti, S, et al. (2007). Germline mutations of TP53 and BRCA2 genes in breast cancer/sarcoma families. *Eur. J. Cancer* 43, 601–606. doi: 10.1016/j.ejca.2006.09.024

McBride, K. A., Ballinger, M. L., Killick, E., Kirk, J., Tattersall, M. H., Eeles, R. A., et al. (2014). Li-Fraumeni syndrome: cancer risk assessment and clinical management. *Nat. Rev. Clin. Oncol.* 11, 260–271. doi: 10.1038/nrclinonc.2014.41

Monti, P., Perfumo, C., Bisio, A., Ciribilli, Y., Menichini, P., Russo, D., et al. (2011). Dominant-negative features of mutant TP53 in germline carriers have limited impact on cancer outcomes. *Mol. Cancer Res.* 9, 271–279. doi: 10.1158/1541-7786.MCR-10-0496

Muller, P. A., and Vousden, K. H. (2014). Mutant p53 in cancer: new functions and therapeutic opportunities. *Cancer Cell* 25, 304–317. doi: 10.1016/j.ccr.2014.01.021

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2, 1–17. doi: 10.1101/cshperspect.a001008

Pavletich, N. P., Chambers, K. A., and Pabo, C. O. (1993). The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes Dev.* 7, 2556–2564. doi: 10.1101/gad.7.12b.2556

Pentimalli, F. (2018). Updates from the TP53 universe. *Cell Death Differ.* 25, 10–12. doi: 10.1038/cdd.2017.190

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30

Richman, S. (2015). Deficient mismatch repair: read all about it (Review). *Int. J. Oncol.* 47, 1189–1202. doi: 10.3892/ijo.2015.3119

Sehgal, R., Sheahan, K., O'Connell, P. R., Hanly, A. M., Martin, S. T., and Winter, D. C. (2014). Lynch syndrome: an updated review. *Genes (Basel)* 5, 497–507. doi: 10.3390/genes5030497

Shajani-Yi, Z., de Abreu, F. B., Peterson, J. D., and Tsongalis, G. J. (2018). Frequency of somatic TP53 mutations in combination with known pathogenic mutations in colon adenocarcinoma, non-small cell lung carcinoma, and gliomas as identified by next-generation sequencing. *Neoplasia* 20, 256–262. doi: 10.1016/j.neo.2017.12.005

Sorrell, A. D., Espenschied, C. R., Culver, J. O., and Weitzel, J. N. (2013). Tumor protein p53 (TP53) testing and Li-Fraumeni syndrome: current status of clinical applications and future directions. *Mol. Diagn. Ther.* 17, 31–47. doi: 10.1007/s40291-013-0020-0

Stepanenko, A. A., and Chekhonin, V. P. (2018). Recent advances in oncolytic virotherapy and immunotherapy for glioblastoma: a glimmer of hope in the search for an effective therapy? *Cancers (Basel)* 10, 1–24. doi: 10.3390/cancers10120492

Tinat, J., Bougeard, G., Baert-Desurmont, S., Vasseur, S., Martin, C., Bouvignies, E., et al. (2009). 2009 version of the Chompret criteria for Li Fraumeni syndrome. *J. Clin. Oncol.* 27, e108–e109; author reply e110. doi: 10.1200/JCO.2009.22.7967

Wawryk-Gawda, E., Chylińska-Wrzos, P., Lis-Sochocka, M., Chłapek, K., Bulak, K., Jędrych, M., et al. (2014). P53 protein in proliferation, repair and apoptosis of cells. *Protoplasma* 251, 525–533. doi: 10.1007/s00709-013-0548-1

Yurgelun, M. B., Masciari, S., Joshi, V. A., Mercado, R. C., Lindor, N. M., Gallinger, S., et al. (2015). Germline TP53 mutations in patients with early-onset colorectal cancer in the colon cancer family registry. *JAMA Oncol.* 1, 214–221. doi: 10.1001/jamaoncol.2015.0197

Zhang, R. Q., Shi, Z., Chen, H., Chung, N. Y., Yin, Z., Li, K. K., et al. (2016). Biomarker-based prognostic stratification of young adult glioblastoma. *Oncotarget* 7, 5030–5041. doi: 10.18632/oncotarget.5456

Conflict of Interest: Author AS, MJ,MM,IK AND GK were employed by company Bio Engineering LLC. Author SM, RR, MS, ZM were employed by company Zan Mitrev Clinic.

frontiers in Oncology

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/fonc.2019.01172

Telomere Length Maintenance and Its Transcriptional Regulation in Lynch Syndrome and Sporadic Colorectal Carcinoma

Lilit Nersisyan[1], Lydia Hopp[2], Henry Loeffler-Wirth[2], Jörg Galle[2], Markus Loeffler[2,3], Arsen Arakelyan[1] and Hans Binder[2]**

[1] Group of Bioinformatics, Institute of Molecular Biology, National Academy of Sciences, Yerevan, Armenia, [2] Interdisciplinary Centre for Bioinformatics, Leipzig University, Leipzig, Germany, [3] Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig, Germany

OPEN ACCESS

Edited by:
Daoud Meerzaman,
George Washington University,
United States

Reviewed by:
Alan Meeker,
Johns Hopkins Medicine,
United States
Elena Tosti,
Albert Einstein College of Medicine,
United States

***Correspondence:**
Lilit Nersisyan
l_nersisyan@mb.sci.am
Hans Binder
binder@izbi.uni-leipzig.de

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 04 May 2019
Accepted: 18 October 2019
Published: 05 November 2019

Citation:
Nersisyan L, Hopp L, Loeffler-Wirth H, Galle J, Loeffler M, Arakelyan A and Binder H (2019) Telomere Length Maintenance and Its Transcriptional Regulation in Lynch Syndrome and Sporadic Colorectal Carcinoma. Front. Oncol. 9:1172. doi: 10.3389/fonc.2019.01172

Background: Activation of telomere maintenance mechanisms (TMMs) is a hallmark of most cancers, and is required to prevent genome instability and to establish cellular immortality through reconstitution of capping of chromosome ends. TMM depends on the cancer type. Comparative studies linking tumor biology and TMM have potential impact for evaluating cancer onset and development.

Methods: We have studied alterations of telomere length, their sequence composition and transcriptional regulation in mismatch repair deficient colorectal cancers arising in Lynch syndrome (LS-CRC) and microsatellite instable (MSI) sporadic CRC (MSI s-CRC), and for comparison, in microsatellite stable (MSS) s-CRC and in benign colon mucosa. Our study applied bioinformatics analysis of whole genome DNA and RNA sequencing data and a pathway model to study telomere length alterations and the potential effect of the "classical" telomerase (TEL-) and alternative (ALT-) TMM using transcriptomic signatures.

Results: We have found progressive decrease of mean telomere length in all cancer subtypes compared with reference systems. Our results support the view that telomere attrition is an early event in tumorigenesis. TMM gets activated in all tumors studied due to concerted overexpression of a large fraction of genes with direct relation to telomere function, where only a very small fraction of them showed recurrent mutations. TEL-related transcriptional state was dominating in all CRC subtypes, showing, however, subtype-specific activation patterns; while contribution of the ALT-TMM was slightly more prominent in the hypermutated MSI s-CRC and LS-CRC. TEL-TMM is mainly activated by over-expression of DKC1 and/or TERT genes and their interaction partners, where DKC1 is more prominent in MSS than in MSI s-CRC and can serve as a transcriptomic marker of TMM activity.

Conclusions: Our results suggest that transcriptional patterns are indicative for TMM pathway activation with subtle differences between TEL and ALT mechanisms in a CRC subtype-specific fashion. Sequencing data potentially provide a suited measure to study alterations of telomere length and of underlying transcriptional regulation. Further studies are needed to improve this method.

Keywords: telomere attrition, colorectal cancer, mismatch repair, telomerase and alternative telomere maintenance, pathway models, DNAseq and RNAseq data analysis, telomere length, telomere repeat variants

INTRODUCTION

The view on telomeres has progressed from simple caps that conceal chromosome ends from DNA repair machinery (1, 2) to complex structures involving hundreds of proteins that have an active role in organizing the genome (3, 4). Telomeres are shortened with each cell division and finally trigger a DNA-damage response resulting in senescence (5). Tumors avoid this by adding newly synthesized telomeric DNA to the chromosome ends via a telomere length maintenance mechanism (TMM), which counteracts telomere shortening and saves the tumor cells from the onset of telomeric crisis thus essentially contributing to cancer progression (6). In most tumors, TMM gets activated via the telomerase pathway (TEL) which utilizes the telomerase ribonucleoprotein containing an RNA template for telomeric DNA synthesis (7). The TEL-TMM is typically active in germline, and to a less degree, in stem cells, but not in somatic cells, due to transcriptional silencing of the TERT-encoded catalytic subunit of telomerase (7, 8). A lower proportion of tumors activates an alternative lengthening of telomeres (ALT) pathway that relies on homologous recombination events between telomeric strands of sister chromatids, distant chromosomes, or extrachromosomal telomeric repeat sequences (9, 10). Usually ALT is associated with altered chromatin environment at telomeres, frequent mutations in ATRX and DAXX genes, the presence of extra-chromosomal telomeric repeat sequences and ALT-associated promyelocytic leukemia bodies (APB) (11, 12).

Most of the tumors (70–90%) are usually assumed to utilize TEL-TMM, while the rest are thought to refer to ALT-TMM (10). Several studies in the last years suggest a more diverse picture where tumors seem to characterized not by just one TEL or ALT TMM phenotype. A recent PanCancer study cross 31 tumor types demonstrated that 73% of the analyzed samples expressed TEL, 5% was associated with ALT, while the remaining 22% of tumors neither expressed clear TERT nor harbored ALT-associated alterations (13). This result is supported by reports that in a so-called ever-shorter telomeres phenotype neither of the two TMMs get activated (14). In addition to such "neither ALT nor TEL" situations, also "TEL and ALT coexistence" *in vitro* and in cancer and "TEL-to-ALT switching" situations were discussed [see (12) and references cited therein]. Mutations of ATRX and of TERT are not sufficient as possible indications for ALT- and TEL-TMM because loss of ATRX coexist with TEL-TMM in some cell lines (15) and melanomas, which can show ATRX and TERT mutations in parallel (16), while they are mutually exclusive in in glioma (17). On the other hand, TERT promoter mutations are not enough to cause activation of telomerase (18). Despite emerging conceptual models, e.g., to explain TEL-to-ALT switching in epithelial tissues (12), it remains largely unclear as to why TEL and/or ALT become activated in specific cancer subsets and what is the molecular mechanism (19).

TEL-positive tumors are typically identified by mutated and/or activated TERT where however about 20% of CRC do not show this characteristics (20). ALT-positive tumors are often deduced from the presence of telomere length maintenance in the absence of TERT activity and/or by assays based on genetic or phenotypic markers, such as the presence of C-circles and/or APBs, but these assays are potentially not definitive for several reasons (21). For example, existence of APBs does not yet ensure telomere synthesis (22). On the other hand, C-circles may be missing in cells with otherwise high ALT activity (22).

Whole genome DNA and RNA sequencing data open novel perspectives for studying telomere length dynamics and TMM in cancer. Here we have applied a bioinformatics approach of telomere length and of sequence variant computation based on DNA-seq data, where, at least the former application represents a robust and accurate alternative to experimental techniques (23–25). This structural information about telomeres is combined with a thorough expression analysis of genes contributing to TEL and ALT activation to shed light into aspects of the underlying transcriptional regulation of TMM. *Omics* data are frequently available in many molecular cancer studies and data repositories, such as The Cancer Genome Atlas (https://www.cancer.gov/about- nci/organization/ccg/ research/ structural-genomics/tcga). They offer an alternative and independent option for studying telomere biology of cancer based on omics data and judging the telomere status as a potential marker of disease development. Understanding the mechanisms regulating telomere length is of importance for development of telomere-targeted cancer therapies (26, 27) and also for identification of markers suited for characterization of early and later stages of cancer development.

TMM may vary from cancer to cancer, and even among cancer subtypes. Consequently, the study of TMM requires a tumor-type specific approach. For example, dysregulation of telomere length is a hallmark of colorectal cancer (CRC), but reports of telomere lengths and their ascribed cancer risks have been discordant, with both very short and very long telomeres implicated (28–30, 30–33). While most studies have addressed telomere length alterations in CRC (30, 32, 34), the mechanisms of telomere length maintenance regulation and, particularly, the

role of mismatch repair deficiency in TMM are still not fully characterized. Here, we focus on CRC showing microsatellite instability (MSI) arising from dysfunctional mismatch repair (MMR) mechanisms in Lynch syndrome (LS-) CRC and in sporadic (s)-CRC as well. LS is one of the most frequently inherited cancer predisposition syndromes contributing to about 3% of all CRC cases (35, 36). It is defined by an autosomal dominant heterozygous constitutional mutation in one of the four key MMR genes MLH1 (about 60%), MSH2 (about 30%), MSH6 or PMS2 (37, 38) all leading to MSI. In contrast, MSI in s-CRC most frequently results from promoter hyper-methylation of the MLH1 gene giving rise to about 20% of all CRC cases (39, 40). The MMR machinery not only has a role in mismatch repair, but also in cell cycle checkpoint activation and DNA damage induced cell cycle regulation. Proteins involved in the MMR pathway, such as PCNA, RPA, and DNA polymerase δ, are also important players in ALT-TMM (30, 41). It has been reported that MSH2 deficiency can accelerate telomere shortening (42). Additionally, it has been shown that MSH6-MMR deficiency leads to a hyper-recombinant phenotype, increased survival of tumor cells in response to telomerase inhibition and shows some evidence of telomeric sister chromatid exchange that are possible signs of ALT (43). Another study has observed a trend of lower expression of TERT and high levels of APBs in MMR-deficient gastric cancer (44). However, possible activation of the ALT TMM in response to MMR-deficiency in CRC still has to be investigated.

With this aim our study addresses TMM of MSI cancers in LS-CRC and in s-CRC, and also in benign colon mucosa and in MS stable (MSS) s-CRC for comparison, which overall constitutes about 60% of all CRC cases. Our study is based on whole genome DNA and RNA sequencing data of patient matched tumor and tumor-distant mucosa samples generated recently by us (45) and of s-CRC data taken from the TCGA repository (40). An interesting aspect results from the fact that cancerogenesis of LS-CRC is driven by immune escape from inflamed non-cancerogenous mucosa (36, 46) with possible impact on telomere biology. The publication is organized as follows: in the first part we analyze alterations of telomere length and of the abundance of canonical and non-canonical telomere repeat variants in the different tumor subtypes and in the reference mucosa systems. In the second part we study how TEL and ALT TMM are regulated at transcriptional level, thus forming different TMM phenotypes.

MATERIALS AND METHODS

DNA- and RNA-seq Data

We made use of whole-genome DNA-seq and RNA-seq data of Lynch Syndrome (LS) referring to paired patient-matched fresh frozen tissue specimens of tumor and tumor-distant non-neoplastic mucosa (reference samples), which were collected from 11 LS-CRC patients, as described and characterized in Binder et al. (45). Tumor samples split into adenoma ($N = 3$) and cancer ($N = 9$) specimen with only one patient-matched adenoma-cancer pair (samples were assigned by patient no. and "reference," "adenoma" or "cancer" sample types). DNA- and RNA-seq data refer to the same mucosa and tumor samples. The data are available at the dbGaP database (www.ncbi.nlm.nih.gov/gap) under accession number phs001407). According to our previous analysis, the LS cases split into two genetically distinct groups named G1 (six patients) and G2 (five patients). G1 tumors showed higher load of somatic mutations (108.000 vs. 34.000 per tumor), a higher number of MLH1 constitutional mutations (5x MLH1 and 1x MSH2 vs. 1x MLH1, 2x MSH2 and 1x MSH6) and higher microsatellite slippage rate, compared to G2 (45). For comparison, we included sequencing data of microsatellite stable (MSS) and instable (MSI) sporadic CRC (s-CRC) cases and of healthy (normal) colonic mucosa taken from the TCGA repository as described in Binder et al. (45). DNA-seq data were taken from patient matched pairs of s-CRC tumors and normal mucosa (5 MSS cases and 8 MSI cases). RNA-seq data refer to unmatched cases of reference mucosa (20 samples), MSS s-CRC (21), MSI-low s-CRC (24), and MSI-high s-CRC (20). In accordance with previous studies (47) the MSS and MSI-low samples were subsumed into one combined MSS group. In support of this, transcriptome patterns along the chromosomes show clearly a common chromosome instability phenotype for MSS and MSI-low s-CRC in contrast to MSI-high s-CRC samples (48), which were assigned the CpG hypermethylation phenotype (CIMP, **Supplementary Figure 1**). MSI-high cases were annotated as MSI throughout the paper. TCGA-accession numbers of all cases studied were listed in Supplementary Table 2 in Binder et al. (45).

Telomere Length and Telomeric Repeat Variants

Mean telomere lengths (MTL) were calculated using the whole genome DNA-seq data and the program Computel (v1.2, accessible at: https://github.com/lilit-nersisyan/computel) using default parameter settings (25). This program detects reads originating from telomeres by alignment to a reference sequence that consists of telomeric repeat patterns (25). It then computes MTL across the chromosomes in units of base pairs (bp), by comparing the coverage at the telomeric reference to the total sequencing depth and normalizing to the number of chromosomes. All LS-tumors, and all s-CRC tumors, except for one, were diploid [see Supplementary Table 1 in Binder et al. (45) which also provided detailed sample characteristics in terms of constitutional mutations, microsatellite status, tumor cell content and patient characteristics, and (49) for s-CRC]. Among s-CRC MTLs were computed for all the runs per sample, and the median MTL was taken for subsequent analysis. Computel also estimates the composition of telomeric repeat variants (TRVs), providing the amount of canonical ("TTAGGG") and non-canoncial TRVs. In contrast to pattern matching algorithms, Computel is not restricted to predefined non-canonical variants, but can capture any variation, be it substitution, insertion or deletion.

Gene Expression Analysis

Identification of differentially expressed genes (DEGs) was performed based on read count data using Wald test implemented in DESeq2 package (50). For functional interpretation of gene expression data we applied gene set analysis in terms of gene set enrichment z-score (GSZ) profiles

(51). Gene sets were taken from the GSEA-repository and from literature for different functional categories (52).

Pathway and Network Analysis of Telomere Maintenance Mechanisms

The genes and pathways involved in TEL and ALT TMM were taken from a literature search and pathway reconstruction approach using reference gene expression data in TEL- and ALT-positive cell systems [see (53, 54) and **Supplementary Methods** for details]. A list of TMM genes is provided in **Supplementary Table 1** together with two independent verifications by means of enrichment analysis in gene ontology categories (**Supplementary Table 2**) and their characteristics as provided by TELNet telomere knowledge base (**Supplementary Table 3**). The activity of the TMM-pathways was estimated by means of the pathway signal flow (PSF) algorithm (55) using the TMM app for Cytoscape. It estimates the transcriptional activity of each pathway node in terms of PSF-scores making use of the local pathway topology and of gene expression fold changes compared to average expression as described in Nersisyan et al. (55, 56). The impact and specifics of PSF-pathway analyses compared with gene set approaches were demonstrated recently in a series of applications to characterize aberrant pathway activation in the context of different diseases (45, 57–59).

We performed TMM-based computations for each of the LS- and s-CRC groups separately. The PSF scores of the different TEL and ALT pathway branches and of the final sink nodes were used to characterize the different tumor subtypes. To estimate the effect, which a selected gene exerts on a certain node of the pathway, we have calculated the partial influence (PI)-score. It is defined as the node's differential PSF-score upon neutralizing the affecting gene by setting its expression fold change to unity. We used the PI-score to select the genes that exert strongest effect on the PSF-scores of the major TMM-branches, either as activators (PI $>$ 0) or as inhibitors (PI $<$ 0), with respect to mean pathway activity of the respective group of samples (see also **Supplementary Figure 2**).

The correlation networks of gene expression and PSF values of the TMM network nodes were constructed using a Pearson correlation significance threshold of $p < 0.05$ for edge selection. Visualization and betweenness centrality (BC) analysis were performed with NetworkAnalyzer in Cytoscape 3.6 (60).

RESULTS

Telomeres Predominantly Shorten in CRC as an Early Event in Tumor Development

In order to explore telomere length changes during malignant transformations, we have analyzed mean telomere length (MTL) in LS-CRC and in s-CRC from whole genome sequencing data using Computel software (25). MTL systematically shortens in all tumor tissues of types G1 and G2 LS-CRC and in MSI and MSS subtypes of s-CRC compared to the respective reference mucosa samples (**Figures 1A,B**), which is in agreement with prior knowledge (56). On average, MTL decreases by 2.7 and 2.3 kbp in G1 and G2 LS-CRC, by 2.7 kbp in MSI s-CRC and only by 1 kbp MSS s-CRC (see also **Supplementary Table 4A**). The larger differences in LS-CRC and MSI s-CRC are in agreement with previous observations that link MSI and (sporadic) defects in MMR with higher telomere shortening rates (31). The MTL-differences between the cancer subtypes and the respective reference mucosa can be eventually attributed to different mean ages of the respective patients (44 $\pm$ 9 vs. 53 $\pm$ 15 years for G1 and G2 LS-CRC patients, respectively; and 63 $\pm$ 12 vs. 75 $\pm$ 12 years for MSI and MSS s-CRC, respectively) and the overall age-related shortening of telomeres in healthy colon mucosa (30, 61), and eventually also CRC (62), which suggests shorter telomeres in the mucosa of older patients (see also **Supplementary Figure 3** for detailed analysis). Overall, we find a broad decrease of mean telomere length in all cancer subtypes.

Telomeric Repeat Variants Suggests Accumulation Near Proximal Regions Without Substantial Changes of Their Composition

Telomeres are not merely composed of canonical TTAGGG repeats, but can also incorporate several types of repeat variants (TRV), such as TCAGGG, TGAGGG, and GTAGGG, particularly in the proximal telomeric and subtelomeric regions (63–65). In order to estimate whether novel TRVs are generated during malignant transformations or as a result of dysfunctional mismatch repair machinery, we have computed the TRV content in our samples. **Figures 1C,D** schematically depicts the average changes in TRV content (mean length in units of bp) in LS-CRC and s-CRC cancers and in reference mucosa. All the samples showed similar TRV distributions (**Supplementary Figures 4–6**). In LS-CRC and s-CRC, the most abundant non-canonical repeat variants all terminated with "GGG," in agreement with the notion of strong selective pressure of this sequence (63). The top TRVs were the G- and A-insertion variants TTAGGGG and TTAAGGG, the (TG)-substitution variant TGAGGG and the T- and A-deletion variants TAGGG and TTGGG, respectively (**Supplementary Figures 4–6**). The mean cumulative length of the TRV was within the range of 20–60 bp per chromosome end, which, in total, comprises $<$1% of the overall MTL. The shortening rate of canonical TTAGGG repeats (35% in LS-CRC and s-CRC) was slightly higher compared to non-canonical TRVs (26% in LS-CRC and 32% in s-CRC). This difference can be explained by a biased placement of non-canonical TRVs toward the proximal (centromeric) regions of telomeres (**Figures 1C,D**). Further differences are noted when comparing TRV in MSI vs. MSS s-CRC. The mean length of TRVs was larger in MSI, consistent with longer telomeres in this subtype (**Figure 1**). Concomitantly, the percentage of most TRVs was lower in MSI tumors, as well as in reference samples compared to MSS (**Supplementary Figure 6**). Relative lower proportion of TRVs were previously reported in ALT positive vs. ALT negative cancers, also attributed to longer telomeres in the former (66). Interestingly, selected TRVs such as the C-substitution variants TTCGGG and TCAGGG are found to show largest differential lengths in our data

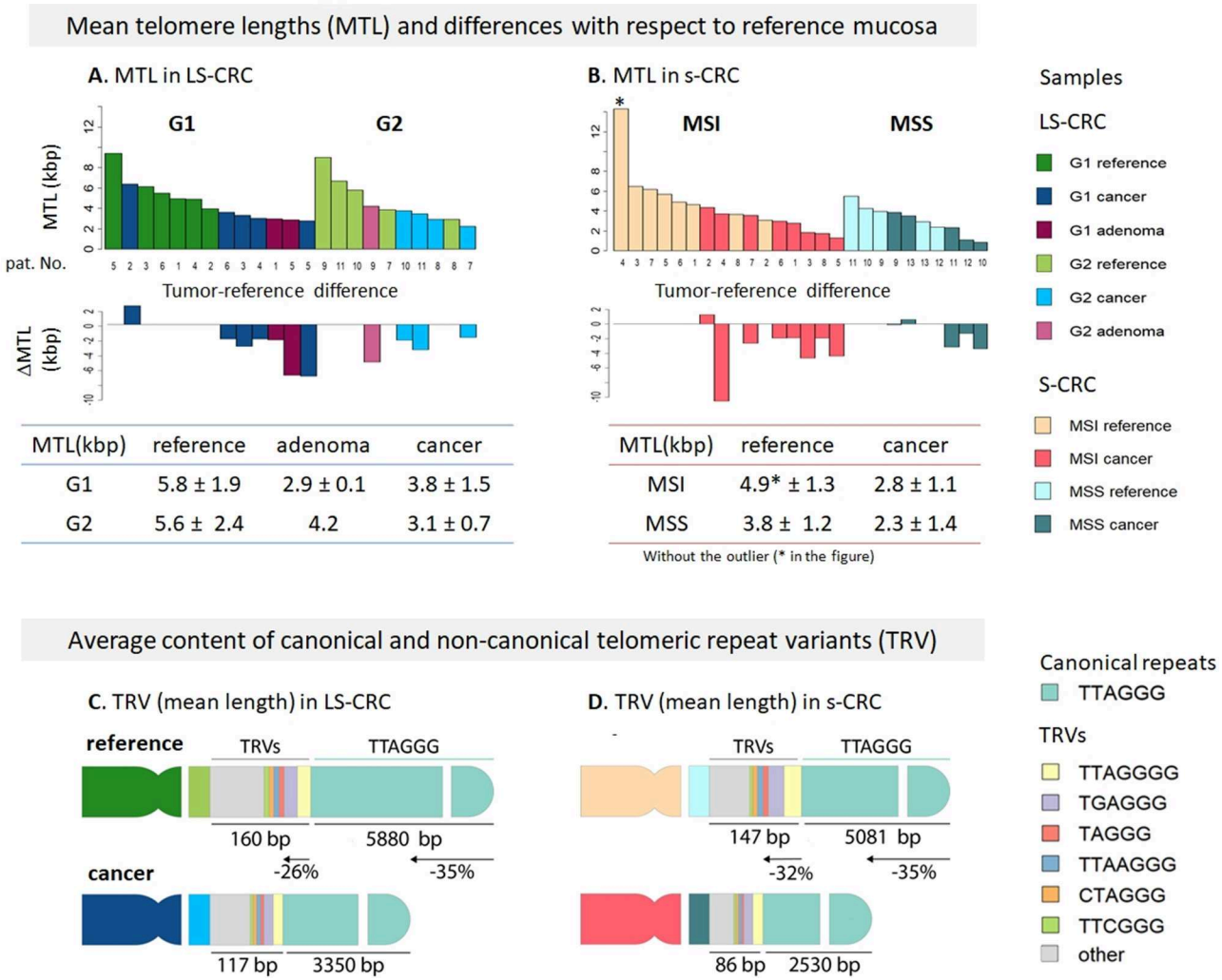


FIGURE 1 | Mean telomere length (MTL) and telomere repeat variant (TRV) analysis in Lynch syndrome and sporadic colorectal cancer. MTL and its differences in tumors with respect to paired reference mucosa samples for LS-CRC **(A)** and s-CRC **(B)** indicate that telomeres broadly get shorter in all tumor types on the average (see **Supplementary Table 4A** for details). Average TRV content in reference and tumor samples of LS-CRC **(C)** and s-CRC **(D)** showed that non-canonical repeats get shorter at slightly lower rates (26–32%) compared to canonical repeats (35%) which suggests their accumulation in the sub-telomeric region as indicated schematically in the figure. The TRVs comprise only 1–2% of the telomere length on the average. The TRV shortening showed a consistent trend in all samples (see **Supplementary Figures S1**, **S2**).

(**Supplementary Table 3B**). TRV analyses largely suggests a small effect size and their likely accumulation in proximal telomeric regions, with selected TRVs (e.g., TTCGGG), showing different trends compared to the rest of the TRVs (66).

All in all, the effects we have observed are small in amplitude and mechanistically not fully understood. Additionally, we also find similar differences in the reference system of MSI and MSS s-CRC. Therefore, TRV dynamics require further, more systematic studies.

TMMs Compensate for Proliferative Telomere Attrition

We next proceeded with gene set analysis to identify biological processes associated with telomere length regulation. We considered two Reactome gene sets for telomerase-based elongation of telomeres ("extension of telomeres" and "telomere maintenance") and one gene set related to alternative lengthening mechanism collecting genes involved in ALT obtained from literature (67). Since activation of TMM usually accompanies the processes of apoptosis and DNA damage-response in most cancer cells, we have also analyzed cellular programs related to cell division, namely, KEGG "mismatch repair", Reactome "regulation of apoptosis" and "cell cycle" taken from Whitfield et al. (68) (**Figure 2A**). They are clearly at lower activity levels in G2 LS-CRC compared to G1, even though MTL shortening is comparable in both subtypes (**Figure 2A**). Possible reasons of this difference between G1 and G2 are addressed below. The comparison of TMM gene sets between tumor and reference

tissue of each LS-CRC subtype showed that the telomerase based TMM is markedly activated both in G1 and in G2 cancers, while the ALT-TMM shows, if at all, only weak activation in tumors. Similar to LS-CRC, the TMM- and the cell division-related gene sets show transcriptional activation in MSS and MSI s-CRC compared to normal mucosa. We also observe activation of the ALT gene set in the MSI and, to a slightly smaller degree, in the MSS s-CRC subtypes.

Plots combining the GSZ-scores of the gene sets with that of cell-cycle activity show marked correlation in all cases, which suggests a high degree of mutual co-regulation, particularly between cell cycle on one hand and TMM, apoptosis and MMR on the other hand (**Figure 2B**). In other words, high cell cycle rates obviously require also high rates of MMR and of TMM to compensate for replication errors and telomere attrition, respectively, which, in turn, relate to increased apoptosis rates (69) that require feedback toward increased cell cycle activity for net survival of the cells. On one hand, TMM, especially TEL, represses apoptosis via telomere maintenance and probably also by extra-telomeric functions of *TERT*, e.g., via modulation of oxidative stress in mitochondria and interactions with apoptotic pathways [see (70) and references cited therein]. On the other hand, only a part of cells acquires immortality at telomere crisis and proceeds to cancerogenesis while the other part becomes apoptotic (71). Our transcriptomics data thus suggest a direct relation between cell cycle, TMM and apoptotic regulation rates. Note also that the data points of MSI s-CRC are systematically shifted toward smaller values for "extension of telomeres" and "mismatch repair" compared with MSS s-CRC, which reflects lower activity of these processes in MSI s-CRC at the same proliferation rate. This kind of feedback is also observed in reference mucosa, which means that the feedback mechanism is obviously not restricted to tumors, but is also present in pre-neoplastic reference mucosa. Hence, TMM seems to follow rather a continuous than a stepwise activation beyond a certain threshold. This hypothesis is further supported by the plot of the MTL of the LS samples as a function of cell cycle activity. It demonstrates that MTL decays non-linearly with increased proliferation rate and levels off into a lower critical value in tumors (**Figure 2B**, part top-left). In other words, telomere attrition due to increased cell cycle activity in tumors gets compensated by TMM resulting in a low, "steady state" critical

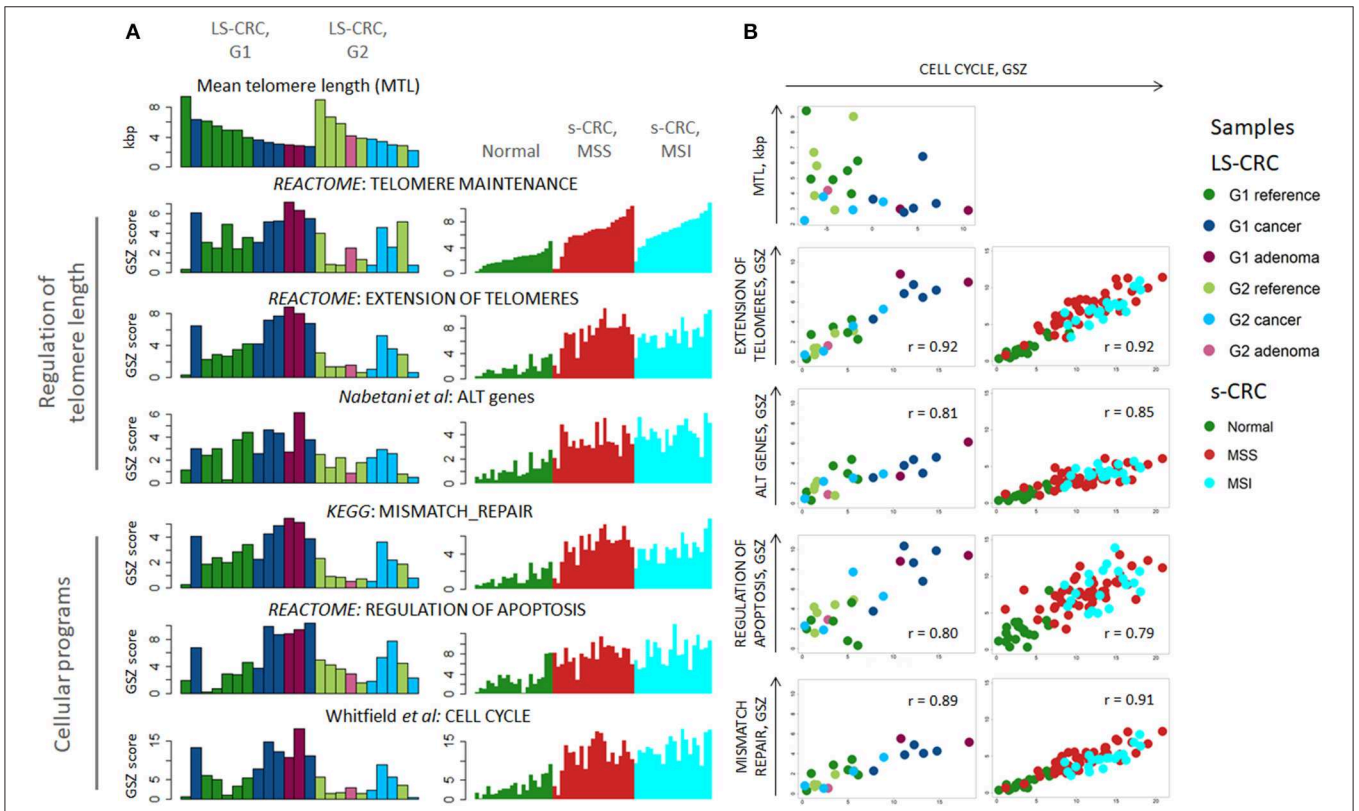


FIGURE 2 | Transcriptome analysis of cellular programs associated with regulation of telomere lengths: **(A)** The gene-set Z-score (GSZ) profiles reflect activation of cellular programs ensuring lengthening of telomeres, cell division, apoptosis and DNA mismatch repair in LS-tumors and s-CRC compared with reference mucosa. LS-CRC samples are sorted with decreasing telomere length in each sample group (**A**, top-left), while s-CRC samples are ranked with increasing GSZ-score of the gene set "telomere maintenance" because of lack of MTL-information. **(B)** Biplots of the GSZ-scores of the gene sets "extension of telomeres," "ALT genes," "mismatch repair" and "regulation of apoptosis" as a function of cell cycle activity suggest a high degree of co-regulation. Note that there is virtually no gene overlap between the gene sets. The plot of MTL as a function of cell cycle activity indicates that telomere lengths asymptotically levels off toward a lower critical MTL-limit in the tumors with increasing cell cycle activity. This trend reflects the fact that replicative telomere loss in tumors is compensated by upregulation of "telomere length maintenance" and "extension of telomeres" mechanisms.

MTL-value. Overall, we find that a whole battery of cellular processes must get up-regulated in concert with cell division rates in order to maintain proper cell functionality, and particularly, a minimum critical telomere length required for cell survival.

Concerted activation of TMM, mismatch-repair, cell cycle and apoptosis related gene sets in cells with high proliferative activity inherently imply that unsupervised analyses of gene expression, e.g., based on correlation with MTL, usually reveal not only canonical TMM genes, but also a large number of genes involved in other cellular programs. To avoid these interferences of mostly unknown background, we focus on a set of genes involved in TMM pathways which have previously been selected based on literature reports and reference gene expression data (53).

Telomerase (TEL) and Alternative (ALT) TMM Pathways in LS-CRC and s-CRC

For detailed supervised analysis on telomere maintenance mechanisms, we make use of previously constructed TMM pathways describing (i) the "classical" TMM that is governed by the catalytic action of the telomerase enzyme (TEL), and (ii) the alternative TMM (ALT) which is realized through homologous recombination events [**Supplementary Figure 7**, (53) and references cited therein]. These pathways decompose into sub-processes that concertedly affect the activity of the TEL- or ALT-TMMs (**Figures 3A,B**). Particularly, the final sink of the TEL-pathway collects activities from the three pathway branches related to telomerase complex components hTERT, hTR, and dyskerin, encoded by *TERT*, *TERC,* and *DKC1*, respectively, and processes leading to their activation, such as nuclear localization and complex assembly (**Figure 3A**). The ALT pathway gets activated via homologous recombination (HR) events involved in break induced repair (BIR) at telomeres, as well as by chromatin decompaction near the telomeres, accumulation of other proteins involved in ALT associated promyelocytic leukemia body (APB) formation and by TERRA induction and telomeric instability. Verification of pathway genes selected using independent knowledge information confirms enrichment of genes with direct involvement in telomere biology (see **Supplementary Tables 2**, **3** for details).

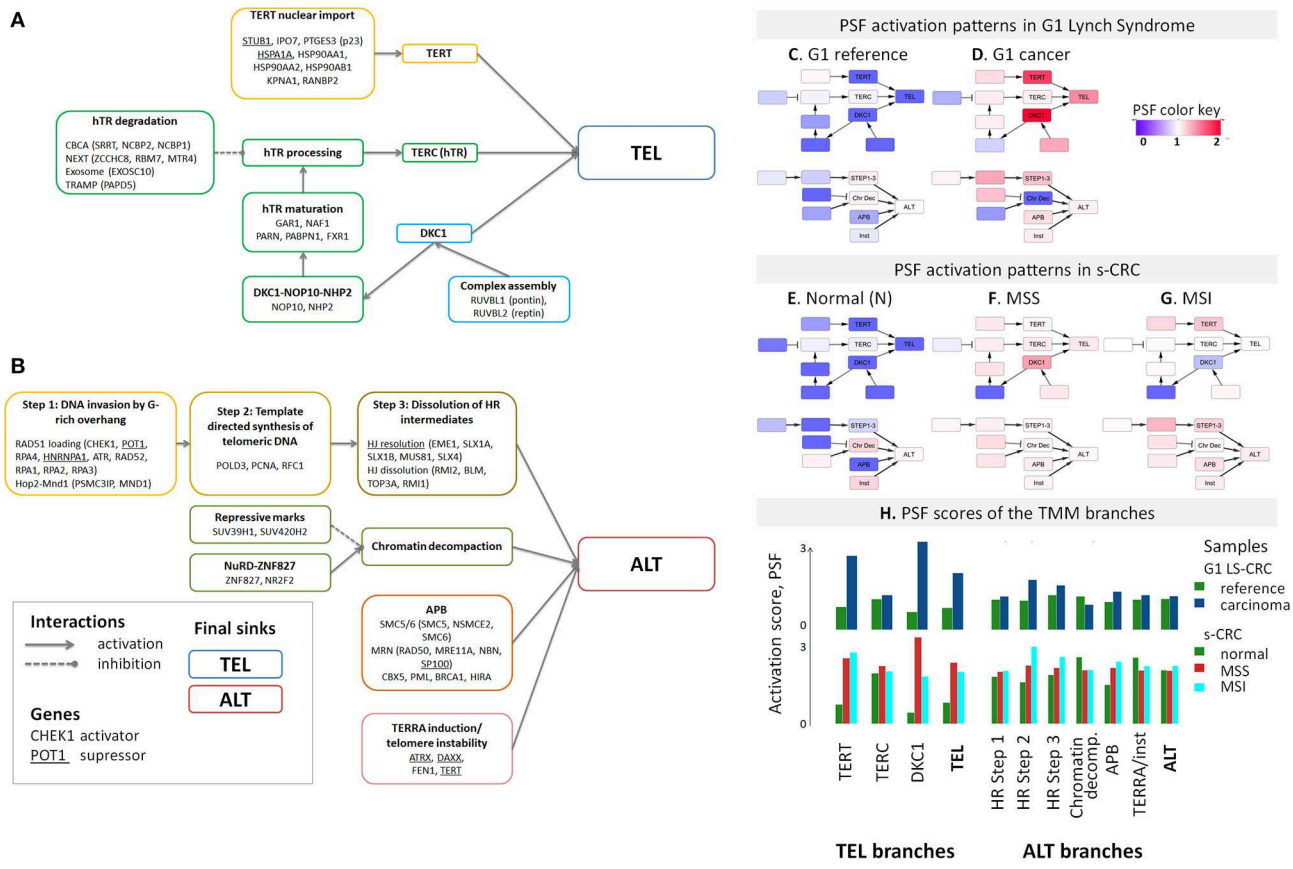


FIGURE 3 | Schematic representation of the TEL **(A)** and ALT **(B)** TMM pathways and their mean PSF-activation patterns averaged over the tumors of each CRC subtype **(C–H)**. The most relevant genes acting either as activators or suppressors are listed in each of the nodes [see (53) for details]. The color of the nodes in part **(A,B)** codes the respective genes and processes throughout the paper. The TEL and ALT-TMM get activated in all CRC subtypes compared with reference mucosa. **(H)** The barplot of the PSF scores of the major TMM-pathway branches reveal that TEL pathway activation in G1 LS-CRC occurs mainly through TERT and DKC1 branches. In s-CRC the TEL pathway is activated either through the DKC1 and TERT branches (MSS) or merely the TERT branch (MSI). ALT-TMM activation occurs mainly via HR- Step 2 and HR-Step 3 and APB nodes in all tumor subtypes, with pronounced activation of Step 2 in MSI s-CRC and G1 LS-CRC.

The activity of these pathways was estimated with the pathway signal flow (PSF) algorithm (53, 55, 56). The algorithm considers expression values of the genes and their mutual interactions to estimate the pathway activity in terms of PSF-scores in each the individual sample, as well as PSF-activities of each individual pathway node. We find marked activation of the TEL- and ALT- TMM pathways in G1 LS-CRC and s-CRC compared with the respective reference mucosa for each of cancer subtypes studied (**Figures 3C–H**). The PSF-scores of the final sinks of the TEL- and ALT-TMM pathways increase in patient-matched tumor samples compared with reference mucosa in G1 (**Figures 4A,C,D**), but not in G2 LS-CRC (**Figure 4B**). Further analysis showed that neither of the TMM genes is significantly differentially expressed in G2 tumors with respect to reference mucosa (**Supplementary Figure 8B**). Moreover, the G2 tumors showed relatively low cell cycle activity compared with G1 tumors (**Figure 2A**). Because of these facts we, excluded G2 data from further analysis, as their transcriptomes seem not to reflect the TMM phenotype of G2 cancer cells. One reason for this problem can be seen in the fact that stromal components in G2 LS-CRC samples (45) can dominate over more subtle expression traits inherent to cancer cells (72, 73).

TMM analysis of the s-CRC samples indicate considerable activation of the TEL pathway in MSS and MSI s-CRC compared to normal mucosa, while ALT-TMM gets activated specifically in MSI s-CRC ($p = 0.004$, Mann-Whitney U test, **Figures 4E,F**). Notably, MSI s-CRC show low variance of TEL pathway activity compared to MSS (F test $p = 0.001$), suggesting existence of a regulatory mechanism dumping variability of TEL TMM activity in these samples (vide infra). Overall, supervised TMM pathway analysis reveals pronounced activation of TEL-TMM in all cancers. Moreover, it suggests specific activation of ALT-TMM in MSI s-CRC.

Transcriptional and Mutational Patterns of TMM Genes

An expression heatmap of the TMM genes, provided in **Figure 5**, suggests their widespread activation in cancer compared to reference mucosa. Indeed, 34% (LS-CRC) and 79% (s-CRC) of all 67 TMM genes in the TEL and ALT-pathways show significant up-regulation (adjusted $p < 0.05$), while only three genes (*RBM7*, *SP100,* and *RAD52*) get significantly down-regulated in at least one of the subtypes (**Figure 6**, **Supplementary Figure 9**). Overall 19 TMM genes (32%) were commonly up-regulated in all three cancer types and another 24 (40%) in MSS and MSI s-CRC (**Figure 6A**) (see **Table 1** for top genes). No gene is found down-regulated in all three cancer subtypes at once: SP100 loses expression in LS-CRC and MSS s-CRC, while RAD52 deactivates in MSS and MSI s-CRC.

Analysis of somatic mutations of the tumors of all three types doesn't reveal high mutational recurrence of TMM genes and also no clear effect of mutations on gene expression in G1 LS-CRC (**Supplementary Figure 10**). Interestingly, we found four genes (*FXR1*, *RAD50*, *SP100*, *SMC6*) mutated in 50% of the G1 LS-cancer samples, with the latter three belonging to the APB branch of the ALT-TMM pathway. All four genes are also recurrently mutated in MSI s-CRC in more than 40% of cases what suggests eventually a mutation-driven mechanism of activation of the APB-branch in G1 LS- and MSI s-CRC as well. No recurrently mutated TMM genes were found in MSS s-CRC possibly due to smaller mutational load compared with the hypermutated subtypes LS-CRC and MSI s-CRC. Besides mutations, epi-mutations, via, e.g., alterations of DNA-methylation patterns in the promoter regions of the genes can affect their expression level. CIMP gene signatures obtained from independent MSI s-CRC and LS-CRC datasets don't show pronounced differential methylation in the promoter regions of TMM genes which makes DNA methylation, at least not a dominant factor that shapes TMM activity (**Supplementary Figure 1**).

In summary, TMM gets activated in all cancers studied due to concerted overexpression of a large fraction of the TMM genes, which seems not to be driven by mutations and/or aberrant DNA-methylation of these genes. In LS-CRC and MSI s-CRC recurrent mutations were found in a few genes of the APB branch of the ALT pathway.

TERT and *DKC1* Activate TEL-TMM

Genes of the DKC1 and TERT branches of the TEL-TMM were commonly up-regulated in all three cancer types (**Figure 6A**), which resulted in the markedly increased PSF-score along these pathway branches (**Figures 3C–H**). The TERT branch involves expression of *TERT*, the catalytic subunit of telomerase, as well as factors supporting and repressing its posttranslational activation (53, 74–76). Activating genes in this branch, first of all *TERT*, heat shock protein 90 (*HSP90AA1*, *HSP90AB1*), importin 7 (*IPO7*) and p23 (*PTGES3*) are overexpressed in cancer, while the heat shock protein 70 (*HSPA1A*) and CHIP ubiquitin ligase (*STUB1*), both acting as suppressors, are underexpressed (**Figure 5**). Note that *TERT* gets up-regulated in MSI and MSS s-CRCs as well (adjusted $p < 0.05$, **Figure 6A**). In G1 LS-CRC, it is not among the top up-regulated DEGs (adjusted $p = 0.23$), however it shows highly variant response with strong activation in two-three patients and weak activation in four out of seven tumors (**Supplementary Figures 8**, **9**, **Supplementary Table 5**).

The genes of the DKC1 branch including *DKC1*, encoding the telomerase subunit dyskerin, and the telomerase complex assembly genes Pontin (*RUVBL1*) and Reptin (*RUVBL2*) are consistently up-regulated in all cancer subtypes (**Figures 5**, **6**, **Supplementary Figure 9**). Expression of *RUVBL1* and *DKC1* progressively increases with telomere length in G1 LS- mucosa (see the plots for these genes in **Supplementary Figure 11**). Also, previous studies report overexpression of *DKC1* upon telomere shortening (77) and increased proliferation (78). The overexpression of *DKC1* and *RUVBL1* in s-CRC is more prominent in MSS, than in MSI (adjusted $p < 0.05$, **Figure 6B**), which explains the less pronounced activation of the DKC1 branch in MSI s-CRC (**Figure 3H**) and presumably also the lower variability of TEL activity in MSI compared to MSS (**Figure 4**).

Next, we evaluated the gene's partial influence (PI) on pathway and branch activity. We find that *TERT* and *DKC1* are indeed the most influential genes strongly affecting the activity of the TEL sink in all CRC subtypes (**Figure 7**). *RUVBL1* and *RUVBL2* are among the top four genes influencing the ALT

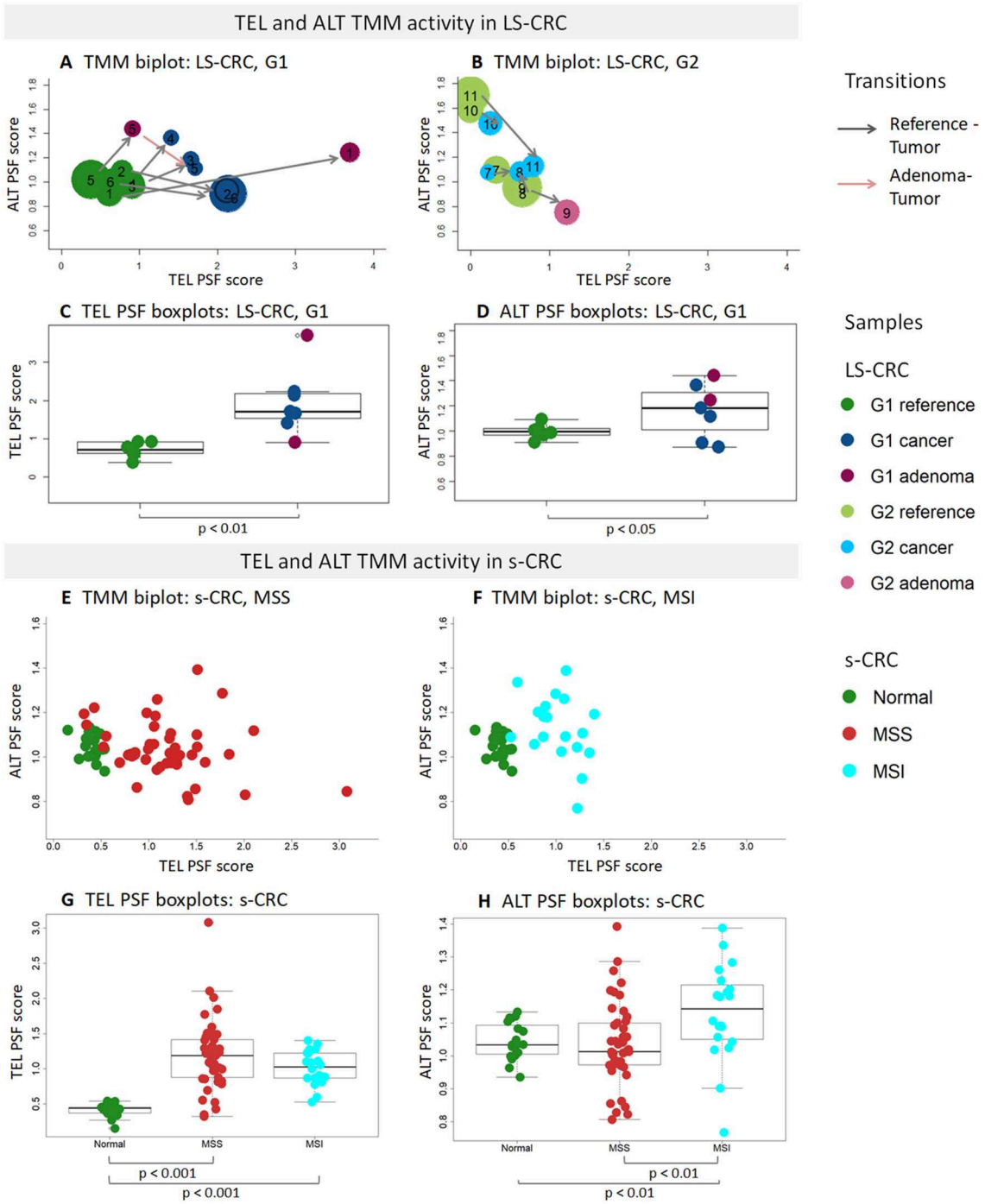


FIGURE 4 | PSF analysis of TEL and ALT TMM pathways in LS-CRC, s-CRC, and the respective reference mucosa samples. The biplots show mutual activation pattern of the ALT and TEL TMM pathways in the LS-CRC **(A,B)** and s-CRC **(E,F)** samples studied. The arrows point from the reference to the tumor for the patient-matched sample pairs. They indicate consistent activation of TMM in all G1 samples but not in G2. The point sizes in **(A,B)** scale proportional to the MTL-values in LS-CRC revealing that shortening of telomeres associates with TMM activation in G1 tumors. The boxplots of the TEL **(C,G)** and ALT PSF **(D,H)** scores show activation of TEL TMM in G1 LS-CRC **(C)** and both s-CRC subtypes **(G)** on the average while ALT-TMM specifically activates in MSI s-CRC **(H)** and to a less extend in G1 LS-CRC **(D)**. Note the larger variability of ALT-TMM PSF in MSI s-CRC compared with MSS s-CRC **(H)**.

sink, with a prominent effect of *RUVBL2* specifically in LS-CRC. At branch level, we observe subtype-specific differences (**Supplementary Figures 12**, **13**). In particular, the TERT branch is affected by p21 (*PTGES3*) in LS-CRC, while in s-CRC cancers, we observe high influence of heat shock proteins (*HSP90AB1* and *HSPA1A*). In MSI s-CRC, the TERT branch is activated by

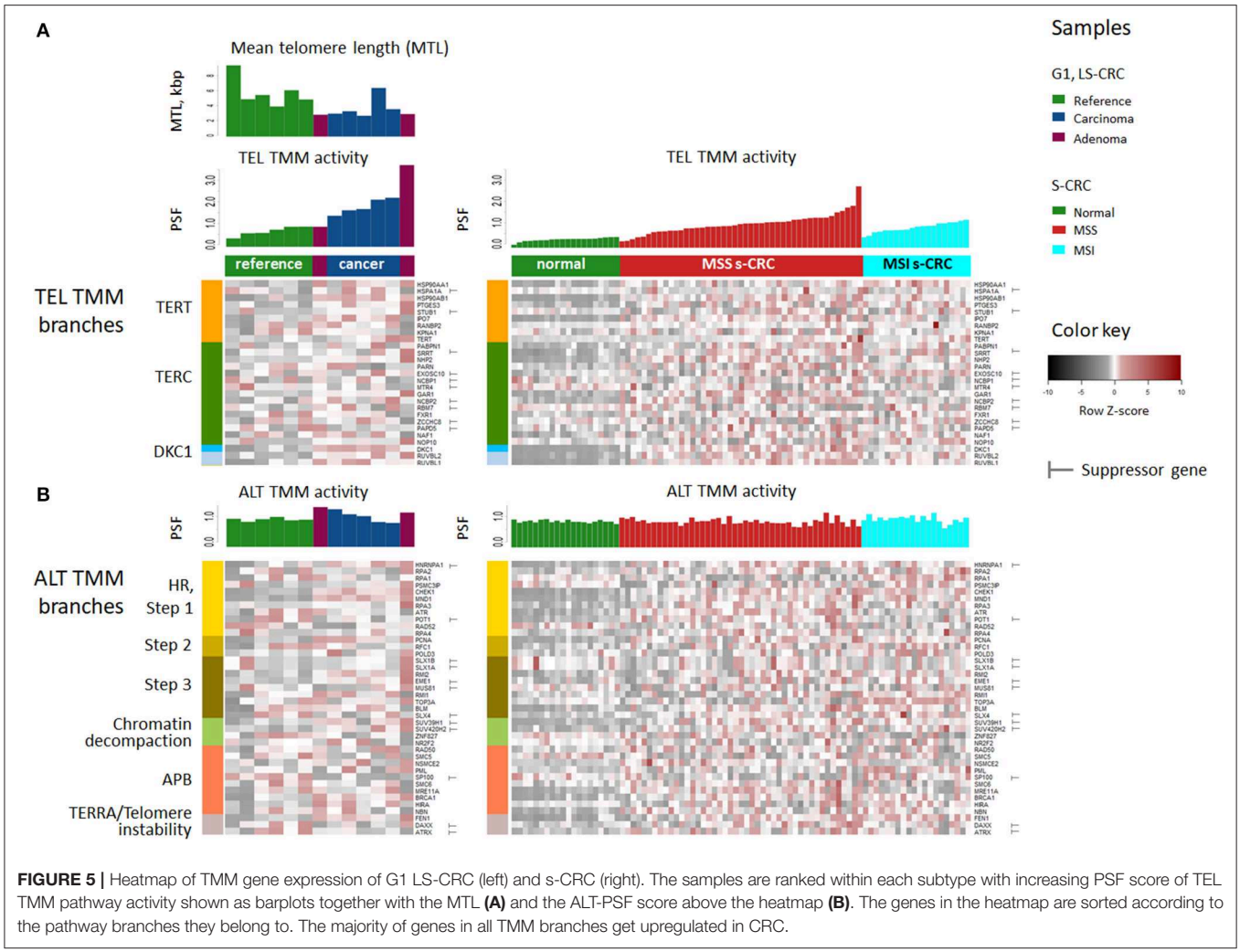


FIGURE 5 | Heatmap of TMM gene expression of G1 LS-CRC (left) and s-CRC (right). The samples are ranked within each subtype with increasing PSF score of TEL TMM pathway activity shown as barplots together with the MTL **(A)** and the ALT-PSF score above the heatmap **(B)**. The genes in the heatmap are sorted according to the pathway branches they belong to. The majority of genes in all TMM branches get upregulated in CRC.

RANBP2, a gene encoding a nuclear pore complex that together with importin 7 (*IPO7*) activates the so called alternative pathway of hTERT entry to the nucleus (75). Interestingly, *IPO7* strongly influences the TERT branch also in LS-CRC, suggesting that in LS-CRC and MSI s-CRC, nuclear import of hTERT occurs via the alternative pathway (**Supplementary Figures 12**, **13**) (53, 75).

These results, altogether, show that the TEL pathway is mainly activated through the TERT and DKC1 branches, by overexpression of *DKC1* and/or *TERT* genes in all CRC subtypes. Importantly, expression of *DKC1* is more prominent in MSS, than in MSI.

Activation of ALT-TMM

Expression of the majority of genes of the ALT-TMM increases in all cancers studied compared with the reference mucosa with a large overlap between them (**Figures 5**, **6**). On mean PSF-level, we found that the ALT pathway is activated in MSI s-CRC and partly also in G1 LS-CRC, but not in MSS s-CRC, and is paralleled by a markedly increased variability of the PSF-values of the ALT-branch compared with that of reference mucosa (**Figure 4**). In all cancer types, we noted activation of the HR branch of ALT-TMM, especially of step 2 and 3, and also of the APB branch compared to reference with larger amplitude in MSI compared with MSS s-CRC (**Figure 3**).

Activation of ALT in MSI s-CRC is mainly due to overexpression of *RAD51* (HR Step 1), *POLD3,* and *RFC1* (polymerase δ subunit, HR Step 2) which suggests activation of template directed synthesis of telomeres via the *RFC1-PCNA-POLD3* axis (41) (**Figure 6**). In addition, the APB branch component *PML* is overexpressed in MSI s-CRC. MSS s-CRC shows also specific up-regulation of a series of other APB-genes (*MRE11A, RAD50, SMC6,* and *NSMCE2*), and down-regulation of *SP100*, which has an inhibitory effect on ALT through sequestration of the MRN complex (*NBN, RAD50, MRE11A*) from APBs (**Figure 6**) (79, 80). Interestingly, *SP100* is found to be the only gene significantly down-regulated in G1 LS-CRC (**Figure 6**, **Supplementary Figures 8A**), which suggests a common function in G1 LS- and MSS s-CRC. Notably, SP100 differential gene expression between CRC tumors and reference mucosa changes in concert with transcriptional signatures of inflammation which indicates especially a marked decay in G1 LS-CRC due to immune

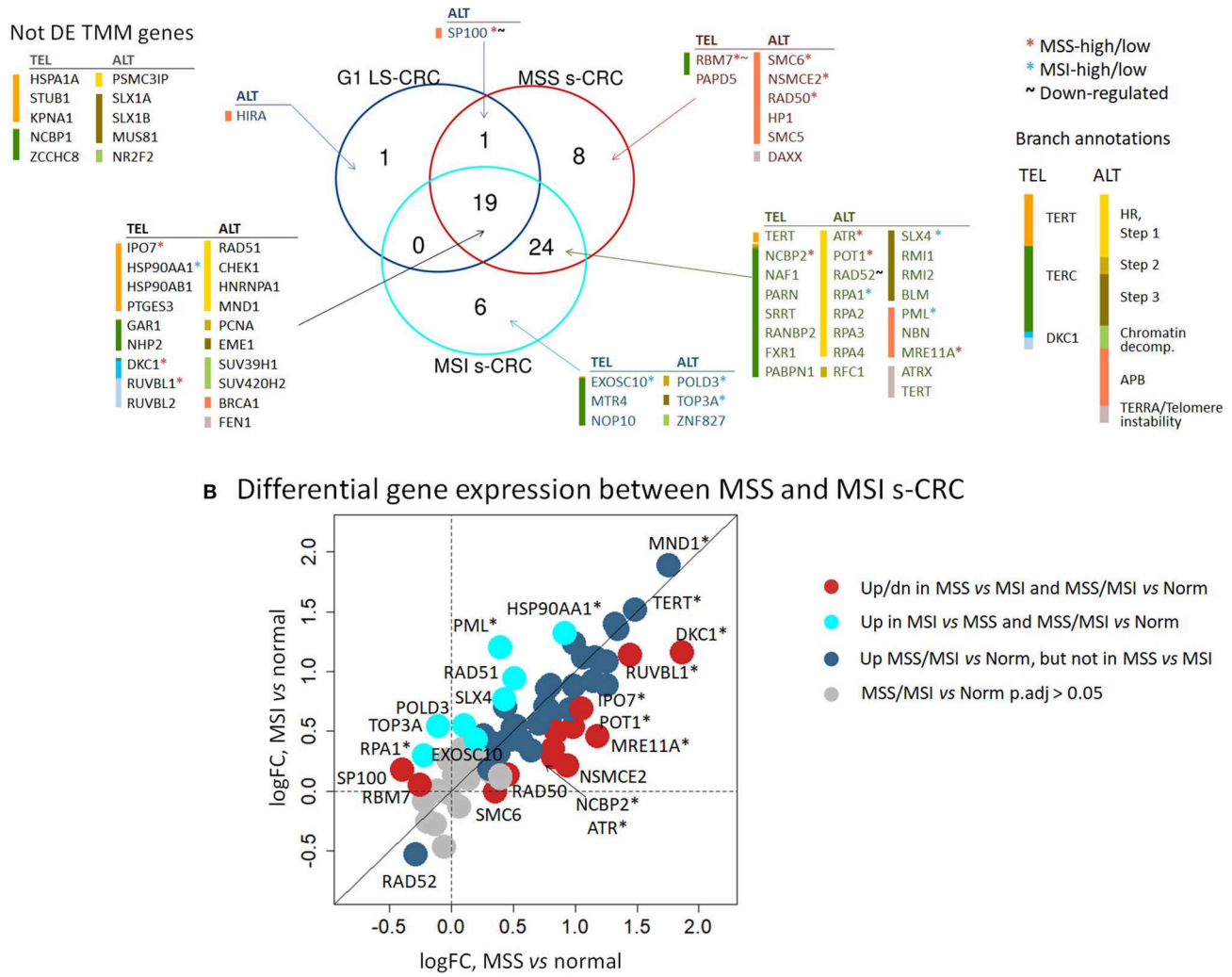


FIGURE 6 | Differential expression analysis of TMM genes of tumor-vs. -reference comparison in LS-CRC and s-CRC and of MSI-vs. -MSS comparison in s-CRC. **(A)** The Venn diagram indicates that 19 out of 59 TMM genes (32%) are commonly differentially expressed between tumors of all subtypes and reference mucosa (adjusted $p < 0.05$, Wald test) and another 24 DEGs (40%) overlap between MSI and MSS s-CRC. **(B)** The biplot of the logged fold expression changes (log FC) between the tumors and reference mucosa of MSS and MSI s-CRC provides a more focused view on gene expression differences only between s-CRC subtypes and shows two types of DEGs which are differentially expressed in tumors-vs. -reference a) but not between MSS and MSI s-CRC (blue circles), or b) also between both s-CRC subtypes in either direction (red and cyan circles). DEGs with adjusted *p* smaller than 0.05 (Wald test) are indicated by asterisks.

escape driven tumorigenesis (45). *SP100* and *PML* accomplish also extra-telomeric functions related to inflammation and immune response (81), and oxidative stress reduction (82), which presumably overlay, or even couple with their roles in TMM (83). High immune cells infiltration is a characteristics of MSI s-CRC (84).

Generally, the top PI-values of the ALT-genes are markedly smaller (range −0.05–0.05) than that of the TEL-TMM (−0.2–0.2). This difference indicates an overall smaller influence of single genes on the ALT-TMM in units of PSF. In addition, we have observed stronger inhibitory effects (PI < 0) of repressor genes in ALT, compared to TEL TMM (**Figure 7**). Among them, the chromatin modifiers *SUV39H1* and *SUV420H2* affecting chromatin decompaction (85), *ATRX* repressing ALT via the TERC/TERRA-instability branch (11), and Holiday junction resolvases *EME1* and *SLX4* that suppress telomere synthesis during ALT (19). Among the top activators of ALT-TMM are the nuclear receptor *NR2F2* and *ZNF827*, with NR2F2 promoting ZNF827-directed recruitment of the NuRD complex to telomeres (86); *BRCA1* and *PML*, genes involved in APB formation (87); and *POLD3*, encoding the catalytic subunit of DNA polymerase δ, involved in template directed telomere synthesis during ALT (41).

TABLE 1 | Top TMM genes in LS- and s-CRC according to different measures[a].

Method	TMM	G1 LS-CRC		MSS s-CRC		MSI s-CRC	
DE[b]		**Gene**	**log2 FC**	**Gene**	**log2 FC**	**Gene**	**log2 FC**
	TEL	NHP2	1.92	**DKC1**	1.86	**HSP90AB1**	1.08
		RUVBL2[e]	1.74	RUVBL1	1.44	HSP90AA1	1.32
		DKC1	1.21	**HSP90AB1**	1.26	**DKC1**	1.16
		GAR1	1.56	IPO7	1.05	RUVBL1	1.14
		HSP90AB1	0.86	**RUVBL2**	0.99	**RUVBL2**	0.88
	ALT	SP100	−1.58	MRE11A	1.18	**CHEK1**	1.40
		CHEK1	1.75	**CHEK1**	1.32	**MND1**	1.89
		EME1	2.06	ATR	0.82	FEN1	1.24
		SUV420H2	1.31	**MND1**	1.76	BRCA1	1.36
		MND1	2.48	BRCA1	1.34	PML	1.20
PI[c]		**Gene**	**Mean PI**	**Gene**	**Mean PI**	**Gene**	**Mean PI**
	TEL	**TERT**	0.20	**TERT**	0.20	**TERT**	0.20
		DKC1	0.19	**DKC1**	0.14	**DKC1**	0.13
		RUVBL2	0.16	**RUVBL1**	0.06	**RUVBL1**	0.05
		RUVBL1	0.07	**RUVBL2**	0.05	**RUVBL2**	0.05
		GAR1	0.05	HSPA1A	−0.02	HSPA1A	−0.02
	ALT	EME1	−0.04	SUV39H1	−0.03	SUV39H1	−0.03
		ATRX	−0.04	**NR2F2**	0.03	**NR2F2**	0.03
		SLX4	−0.03	SUV420H2	−0.02	SUV420H2	−0.03
		NR2F2	0.02	**ATRX**	−0.02	PML	0.02
		BRCA1	0.02	BRCA1	0.02	**ATRX**	−0.02
BC[d]		**Gene**	**BC**	**Gene**	**BC**	**Gene**	**BC**
	TEL	-	-	DKC1	197	PTGES3	252
		-	-	RUVBL2	121	**TERT**	125
		-	-	**TERT**	121	NAF1	59
		-	-	RANBP2	120	PARN	56
		-	-	SRRT	117	HSP90AA1	54
	ALT	-	-	BLM	178	EME1	352
		-	-	FEN1	111	ATRX	330
		-	-	MND1	101	HNRNPA1	194
		-	-	BRCA1	92	RPA3	193
		-	-	ATR	84	NR2F2	176

[a] *The full list of all TMM genes with differential expression values is given in* ***Supplementary Table 1****.*
[b] *Mean differential expression (DE) of genes in CRC vs. reference (log2 fold change (FC), adjusted $p < 0.001$, Wald test).*
[c] *Partial influence (PI) of genes on TEL and ALT pathways averaged over sample groups.*
[d] *Betweenness centrality (BC) of genes in the pairwise gene expression correlation network in MSS and MSI s-CRC. Not computed for LS-CRC, because of small sample size.*
[e] *Redundantly found genes were highlighted in bold font.*

Hence, ALT seems to be affected by numerous genes, especially in MSI s-CRC, which concertedly adjust the activity of this pathway by activating and inhibitory influences of relatively small amplitudes. This is in line with the notion that the regulation of ALT is more complex and involves multiple layers of processes such as epigenetic modifications and homologous recombination events (12, 88), while TEL may be regulated in a simpler way by single factors, such as induction of *TERT* or *TERC* expression. Altogether, our data indicate that TEL is the major TMM in the CRC cases studied, while the ALT pathway additionally activates mainly in MSI s-CRC due to the concerted action of a number of factors, among them the HR and APB TMM branches as the main drivers.

Gene Regulatory Networks in MSI and MSS s-CRC

To assess the degree of co-regulation between the TMM-genes, we constructed pairwise correlation networks of expression values separately for MSI and MSS s-CRC (but not for LS-CRC because of small sample size). We included also the PSF-scores of the major sink nodes of the TEL- and ALT-branches of the TMM-pathways to directly evaluate correlations

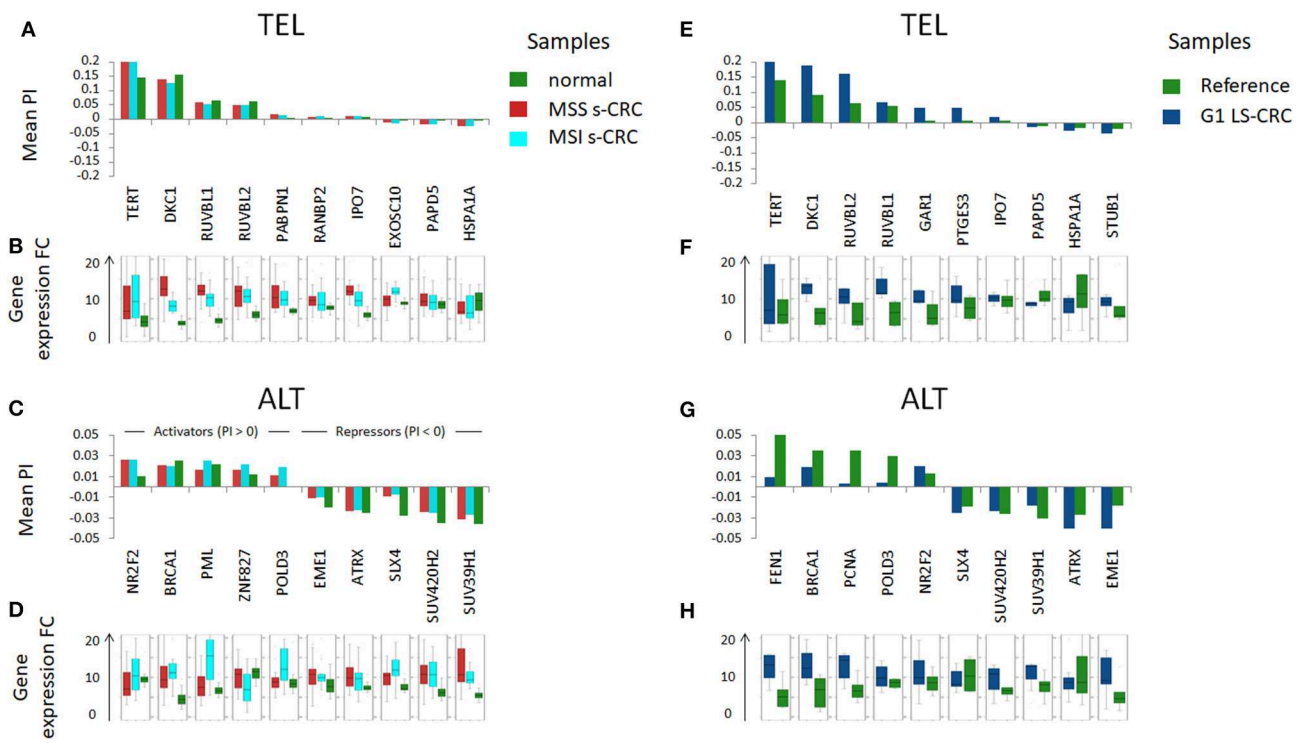


FIGURE 7 | Partial influences (PI) of TMM genes on TEL and ALT pathway activities and their expression differences in units of log FC. *DKC1* and *TERT*, as well as *RUVBL1* play strongest roles in TEL pathway activity in all cancer subtypes **(A,E)**. The ALT TMM pathway shows less pronounced PI amplitudes: it becomes activated by *NR2F2* and *BRCA1* and suppressed by chromatin modifiers (*SUV420H2, SUV39H1*), holiday junction proteins (*SLX4, EME1*) and *ATRX* **(C,G)**. These observations are mostly (with some exceptions) supported by respective gene expression changes **(B,D,F,H)**.

between branch and gene activities (**Figure 8**). The degree of interconnectivity of the nodes of the networks was then compared between the two s-CRC types using betweenness centrality (BC) as a measure (**Figure 8**). The distributions of BC-values of both s-CRC types indicate scale-free properties of the networks, which are characterized by a few highly interconnected "hub"-genes and/or -nodes accounting for most of the regulatory interactions and a large number of weakly connected genes/nodes.

More detailed inspection revealed pronounced differences between MSS and MSI s-CRC: most of the genes having high BC values in MSS s-CRC belong to the TEL pathway (e.g., *TERT* and *DKC1*), including also the TEL-sink node while the tail of the distribution showing low BC values accumulates ALT genes (**Figure 8B**). The reverse picture with highly connected ALT- and weakly connected TEL-genes and nodes is found for MSI (**Figure 8D**). In this subtype, the ALT-genes *EME1* (step 3 of HR branch) and *ATRX* (TERRA/Telomere instability branch) are strongest hub regulators according to their large BC values. This result is in line with the known fact that the chromatin re-modeler *ATRX,* being responsible for proper histone deposition at telomeres, acts as a key regulator suppressing ALT in many cancers, where however its deactivating mutation (as, e.g., in astrocytic gliomas) is not mandatory by unknown reasons. However, also a few TEL genes predominantly from the TERT- (*PTGES3, TERT, HSP90AA1*, see **Table 1**) and TERC- (*NAF1, PARN*) branches are obviously strongly involved into the network of this subtype suggesting coupling with ALT-TMM. Note also that *DKC1*, which is one of the strongest regulators in MSS s-CRC, nearly completely lacks interconnections in the MSI network, which is in line with the decreased activity of this gene in this subtype. Overall, these results suggest that the TEL pathway is more prone for activation in MSS, while ALT in MSI s-CRC according to the "guilt by association" paradigm assuming that co-regulated genes are likely to be involved in the activation of a biological process (89).

We do not observe separate clustering of TEL and ALT pathway genes in either of the subtypes, but rather a common network with a high degree of cross-connectivity suggesting mutually linked co-regulation of the two TMM processes. Interestingly, a number of anti-correlated and thus mutually repressive interactions were detected between TEL- and ALT-networks, especially in MSI s-CRC, e.g., between *ATRX* (TERRA branch) and *PTGES3* (TERT branch) both showing also highest BC-values which makes them candidates of regulatory links

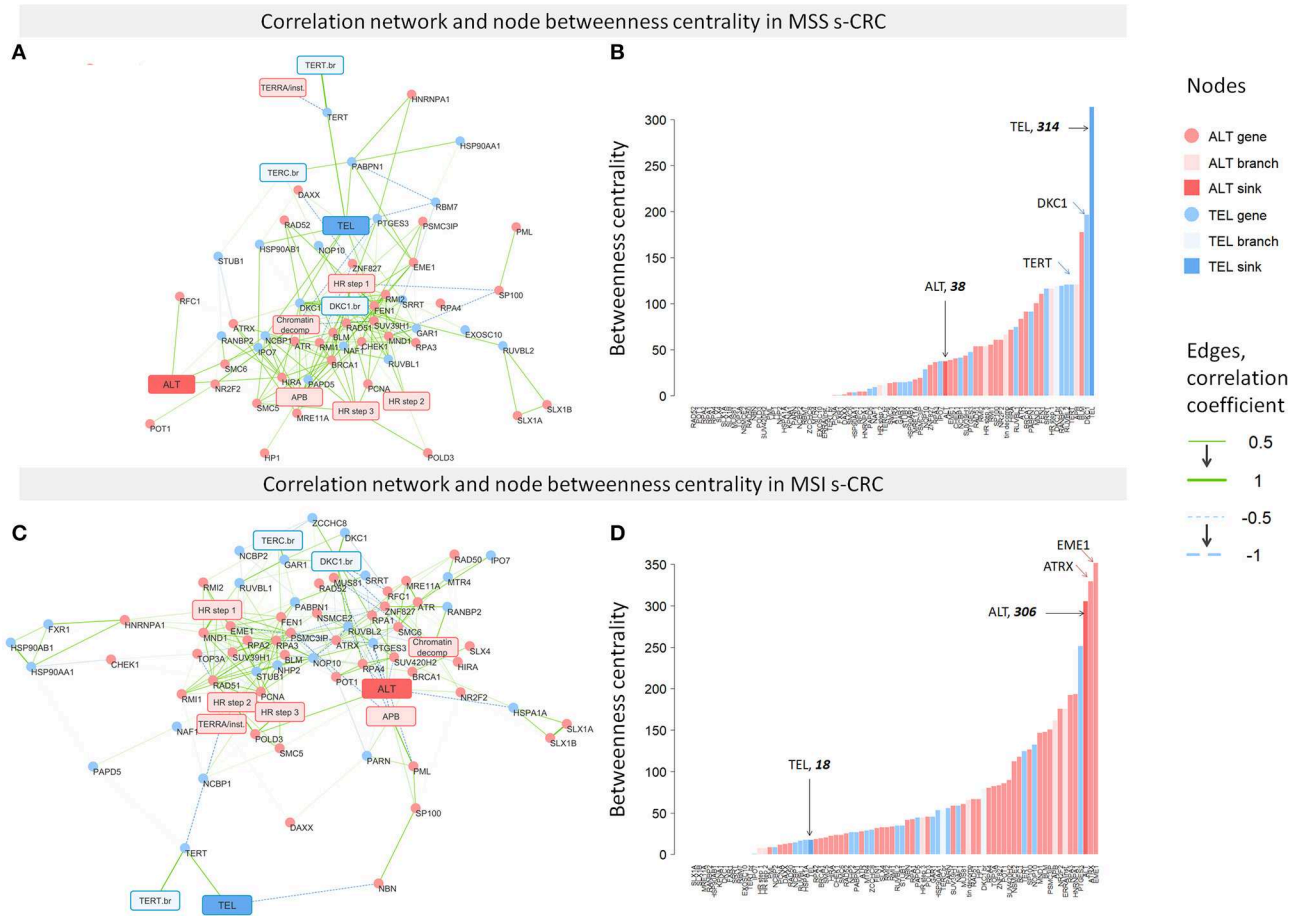


FIGURE 8 | Correlation networks of TMM gene expression and ranked node betweenness centrality (BC) in MSS **(A,B)** and MSI **(C,D)** s-CRC. The edges of the network refer to Pearson correlation coefficients between gene expression and/or PSF-scores of the TMM nodes. It shows that the MSS network is dominated by a large number of connections linked to the TEL-node, while the MSI network is characterized by interconnections between ALT-nodes.

between TEL- and ALT-TMM. In summary, co-regulatory network analysis supports the notion of a more pronounced activation of TEL in MSS, and of ALT in MSI s-CRC (**Figure 4**), at the same time showing no clear-cut decoupling between the two telomere maintenance processes, but rather their coexistence, and co-regulation.

DISCUSSION

We have performed a combined study of telomere length and its transcriptional regulation in selected subtypes of CRC using bioinformatics analysis based on DNA and RNA sequencing data and using a TMM-pathway model. Our analysis provides insights into telomere length regulation in MMR deficient CRCs caused either by constitutional mutations mainly of the MLH1-gene in LS-CRC or by hypermethylation of the MLH1-promotor in MSI s-CRC, both leading to hypermutated cancer phenotypes (**Figure 9**). For comparison, we included MSS s-CRC cases forming a chromosomal instability (CIN) phenotype and specimen of non-tumor mucosa.

Alterations of Telomere Length Indicate Tumor Onset but Are Virtually Insensitive for CRC Subtypes

We have found that all CRC-types studied had on average shorter telomeres than non-tumor colonic mucosa tissues, in agreement with previous reports (30, 32, 34, 61). Gene set analysis of transcriptomic data shows that accelerated cell division rates inversely relate to MTL until telomere length reaches a critical lower limit, which is then maintained after activation of TMM. This scenario is in agreement with the classical model of telomere maintenance. Accordingly, intensive proliferation of cancer cells leads to loss of telomeric caps, which triggers telomere crisis, and chromosomal instability and then drives early carcinogenesis (90–92) enabling cancer cells to bypass telomere-induced apoptosis by activating TMM just on a level which maintains the minimum critical telomere length required for survival (93).

Our observation that the transcriptional level of cell cycle-related genes is proportional to the activity of TMM genes also in non-tumor mucosa suggests that TMM becomes continuously

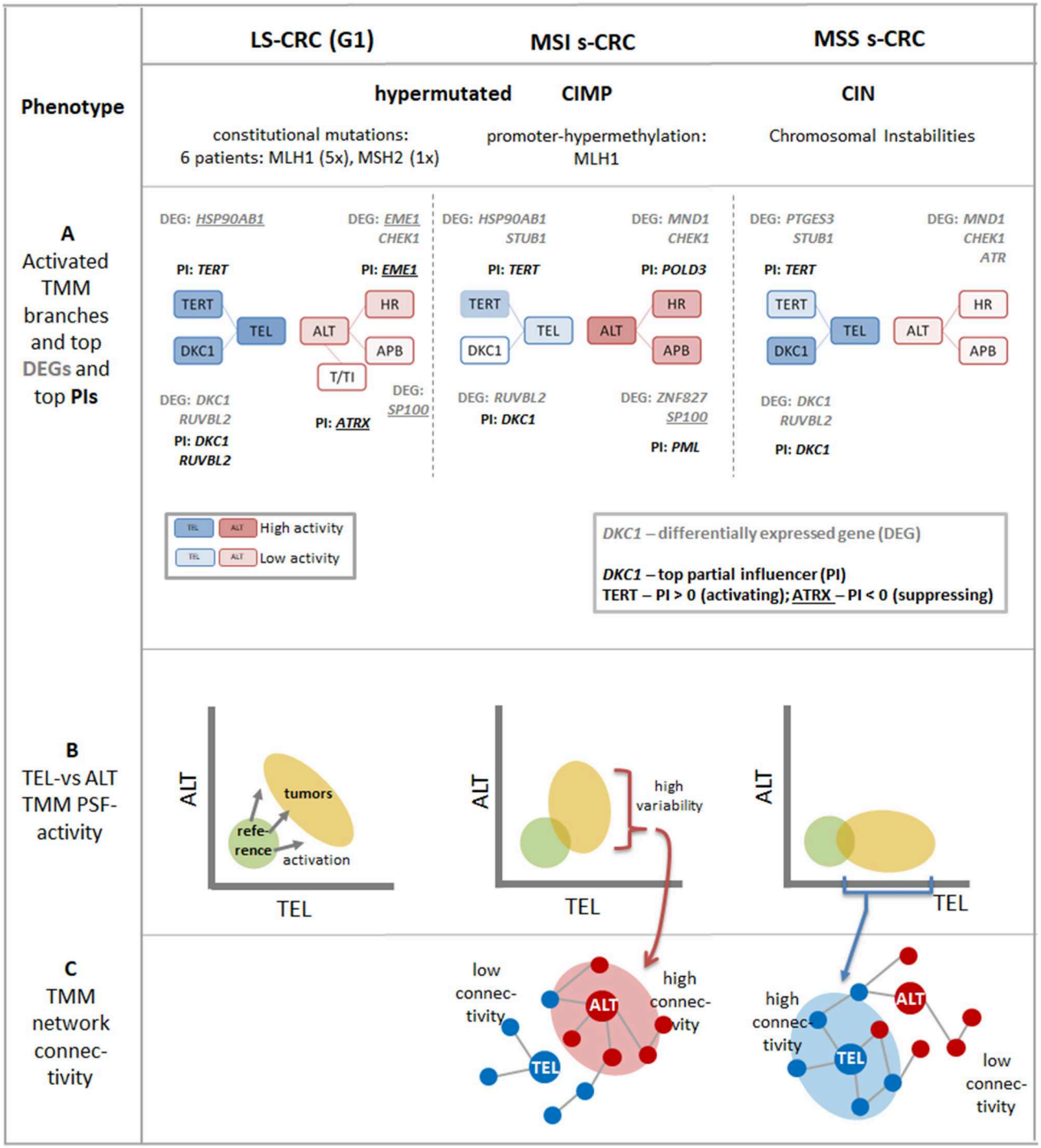


FIGURE 9 | Schematic summary of the major aspects of TMM pathways in LS-CRC and in MSI and MSS s-CRC associating with different tumor phenotypes: **(A)** The transcriptional activation patterns of the TMM pathways lead to a shift from more active ALT-TMM in MSI s-CRC toward more active TEL-TMM in MSS s-CRC and concerted activation of both TMM in LS-CRC (see also **Figure 3**). The top differentially regulated genes (DEGs) and top partial influencers (PI) in cancer vs. reference tissues are depicted. The activity of the final TEL node is strongly "influenced" by *TERT* and *DKC1* genes while ALT is under control of a series of genes exerting activating as well as inhibiting effects of small and moderate amplitudes (see also **Figure 7**). Overall, *TERT* and *DKC1* are key factors leading to activation of TEL-TMM in all cancer subtypes studied while ALT TMM is affected first of all by APB, HR, and other subbranches of the TMM pathways. **(B)** The dark yellow ellipses schematically illustrate the distribution of tumor data. Their more distant position from the coordinate origin compared with the location of mucosa reference samples (green circles) reflects activation of TMM in the tumors (see also **Figure 4**). The decreased variations along the TEL and ALT axes reflect repression of these TMMs in MSI and MSS s-CRC, respectively. **(C)** Stronger activation of TEL- or ALT-TMM accompanies with markedly increased interconnectivities of the correlation networks formed between the genes and sink-nodes of these TMMs, respectively (see also **Figure 8**).

activated in pre-neoplastic mucosa. This view is supported by the continuously decreasing distribution of MTL-values in reference mucosa without clear-cut separation with respect to MTL in the tumors. Moreover, all LS-adenomas show MTL near the minimum values observed in the LS-cancers. Overall these results support the view that telomere attrition is an early event in CRC tumorigenesis (94) and that early carcinomas arise from cells with critically short telomeres (95).

We find that the difference between reference tissue and tumor telomere lengths is larger in MSI s-CRC and LS-CRC, compared to MSS s-CRC which can be rationalized by higher telomere shortening rates in hypermutated tumors (31, 96), or, alternatively, also by earlier diagnosis and the younger mean age of LS- and MSI s-CRC patients, possessing on average longer telomeres in their reference tissues. We find slightly shorter mean MTL in MSI compared with MSS tumors, in agreement with (96), however at low significance level ($p = 0.19$), presumably due to our small sample size. Experiments on mice have indicated that dysfunctional TEL-TMM and MMR-defects can abolish anticancer activity of short telomeres via cell cycle related mechanisms (97).

Telomeric Repeat Variants—Suited Markers for TMM?

Non-canonical telomere repeat variants (TRV) were found to cover up to 2% of the overall telomere length in the tumors and reference tissues studied in agreement with data on other cancer types (66). The most abundant TRVs detected are the substitution variant TGAGGG, previously reported in other studies (66, 98), and a novel insertion variant TTAGGGG. The slight increase of the relative amount of TRVs in tumors (by up to 1.5%) can be rationalized by biased accumulation of TRVs in the proximal telomeric regions virtually not affected by telomere attrition (98).

Only a few studies have explored the difference between TRV generation in tumors with activated telomerase or ALT so far (63, 66), to the best of our knowledge. They have reported differences in TRV abundances between TEL and ALT TMM, mostly based on cell line systems. TEL, on one hand, is found to induce substitutions at repeat positions 1 and 3 due to improper telomerase function (63). ALT, on the other hand, seems to induce random placements of TRV arising from proximal and terminal regions of telomeres via homologous recombination (63). Later, the same group has classified ALT positive(+) from ALT negative(−) cell lines, based on relative TRV content and relative telomere length (66). Most of the ALT-related TRVs had lower relative TRV content, largely attributed to longer telomeres in these cell lines and to "proximity effect." We found a similar trend in MSI-vs.-MSS comparisons (**Supplementary Figure 6**) which corresponds to the slightly enhanced ALT-TMM expression signature in MSI s-CRC reported by us. Interestingly, all the TRVs, except for TTCGGG behaved similarly, showing reduced relative content in MSI vs. MSS s-CRC, in agreement with ALT+ vs. ALT- differences observed in Lee et al. (66) (**Supplementary Table 3B**). Overall, our TRV analysis thus agrees with the previous reports regarding the basic trends to distinguish ALT-vs.-TEL TMM in agreement with our transcriptomic data.

Importantly, TRV studies based on sequencing data are still (very) rare. Absolute quantification of TRV lengths requires systematic methodical studies. Computational telomere and especially TRV length estimates should be interpreted as subjective measures with possible off-sets between the methods. The different approaches in these methods, such as telomeric read capture [alignment (25) vs. repeat count with differing count thresholds (63, 66)] may lead to capturing subtelomeric and interstitial telomeric repeats at varying degrees, which may eventually affect absolute TRV length and relative content. Consequently, they provide consistent quantitative results only within each method used. TRV-estimates are expected to be prone to systematic shifts due to varying GC-content and G-stack formation with strong effects on hybridization chemistry (99) and possible consequences for read-count estimates.

Overall, our results and previously reported findings underline the need for further studies on association of TRV composition with TMM activation across cancers in general and in CRC subtypes in particular. Moreover, the small amplitude of TRV changes and confounding factors affecting, e.g., age and telomere length and their overlay with TRV-proximity effects leaves a series of questions still unanswered.

Different Levels of Expression Analysis Provide Consistent TMM-Related Transcription Patterns Specific to CRC Subtypes

Gene expression data were analyzed making use of pathway models considering a set of 67 genes with relevance for TEL- and ALT-TMM. Analyses have been performed at four levels addressing different aspects of transcriptomic regulation (**Figures 9A–C**): (i) Differential expression analysis, as the most "simple" approach, was applied to estimate expression differences of the genes between cancer and reference mucosa and between the cancer subtypes, as independent entities; (ii) Pathway signal flow (PSF) analysis, has been used to estimate the activity of genes in a certain pathway topology considering their mutual interactions; (iii) The partial influence (PI) was applied to estimate the specific impact of a selected gene on a certain node of the pathway; (iv) Finally, correlation network analysis enabled us to select co-expressed and thus potentially co-regulated genes in an unsupervised fashion, i.e., without assuming a predefined wiring between them.

In all these analyses, we separately considered the TEL- and ALT-TMM in order to compare their particular impact on each of the CRC subtypes. For this purpose, we generated biplots of their pathway activities (**Figure 9B**) and provided TEL-and ALT-specific lists of top genes in units of differential expression, PSF, PI, and BC, respectively (**Figures 9A–C** and **Table 1**). This parallel view on both mechanisms was motivated by recent research indicating that categorization of tumors into either TEL- or ALT-positive ones appears to be imprecise. In other words,

tumors do not necessarily classify into exclusively a single TMM-type. Particularly, TEL- and ALT- TMM can coexist either in different cancer cell sub-populations of the same tumor (12) or within the same cell (15). Moreover, TEL- and ALT-TMM are capable of switching from one mechanism to the other one during different stages of tumor development or upon treatment (15).

Our results support this view. We find concordant activation of both TMM pathways in all the CRC subtypes studied compared with the reference mucosa systems, showing no clear-cut separation between samples in terms of either TEL or ALT pathway activation (**Figure 9**). TEL seems to be the dominating TMM in all analyzed CRC subtypes. However, the branches leading to activation of hTERT (TERT branch) and dyskerin (DKC1 branch) contribute differently with distinctly stronger mean contribution of DKC1 in MSS compared with MSI s-CRC. In turn, ALT-TMM shows stronger effects in MSI compared to MSS s-CRC; mainly via APB formation (APB branch) and homologous recombination events (HR branch) (**Figures 9A,B**). Regulation of ALT pathway is more complex than TEL and involves multiple events. Strikingly, the two TMMs show strong co-regulation of member genes.

Notably, higher mean activity of ALT-TMM in MSI CRCs is accompanied by higher variability of the ALT-PSF values in these samples *and* stronger co-regulations between the ALT-genes in the gene network. Such co-regulations are indicated by higher network connectivity in these CRCs compared with MSS s-CRC, where the relations between these characteristics are reversed. These results stand for a possible trend of increased sensitivity for ALT in MSI and of TEL in MSS s-CRC, which, in turn, can reflect repressive feedback mechanisms between TEL- and ALT-TMM presumably mediated by anti-correlated links detected in network analysis especially in MSI s-CRC. On the other hand, co-activation of TEL and ALT in the tumors, strong co-regulation between the TEL- and ALT-TMM genes and positive correlation of both TMM with cell cycle activity and other cellular processes, indicate that mutual activation of TEL and ALT-TMM is possible in most of the cancer samples. All together, these results support the notion of a TEL-ALT continuum of expression and pathway activation patterns, where both pathways are concertedly regulated in a fine interplay of activating or mutually repressive interactions. This kind of regulation eventually leads to a situation, where TEL and ALT can co-exist in the same tumor, although at different activity levels. These levels can be specific for each tumor subtype.

TMM Genes as Markers of Telomere Attrition and Limitations of the Study

LS-CRC (G1) and MSI s-CRC reveal an increased mutational load compared with MSS s-CRC including the TMM genes (45). However, only few of them were mutated on moderate recurrence levels of <50% mainly in the APB branch of ALT-TMM (**Supplementary Figure 10**). Hence, mutation markers seem not to be suited for judging tumor development, subtypes and/or TMM in CRC. This contrasts to other cancer types, such as gliomas that show strong association between astrocytic and oligodendroglial subtypes and telomere biology, which is driven mainly by mutations of the *ATRX* and *TERT* genes, respectively (100), as well as aggressive metastatic melanomas (101) and other cancers [see (102) and references cited therein] showing a high percentage of *TERT* mutations.

According to our results, RNA-seq data has the promise to offer an alternative and independent option for judging the telomere status of CRC. Fortunately, they are available in many molecular cancer studies. Frequently, *TERT* is used as a gene expression measure of TMM activity, e.g., to estimate tumor progression in CRC [see (28–30, 30–33) and references cited therein]. Here, we found significant differential expression of *TERT* between s-CRC tumors and reference. However, TERT showed by far not the largest effect (position 23, 29, and 31 in the ranked lists of 67 DEGs in MSS, MSI and LS, respectively; see **Supplementary Table 5**). Because of multiple extra-telomeric functions of *TERT*, by *TERT*-bypassing mechanisms of tumor development (103) and because of subtle epigenetic regulatory mechanisms of *TERT* activity (104). Moreover, whether TERT expression translates directly to telomerase activity is unclear because only the full-length transcript (as opposed to known isoforms) has been found to activate telomerase (105, 106). Thus, the transcriptional level of this gene may not serve as a stable indicator of TEL pathway activity. We found that other transcripts, such as *RUVBL2* (telomerase complex assembly), *DKC1* (telomerase subunit) and also *HSP90AB1* (*TERT* nuclear import), show much stronger and more consistent effects in our TEL-TMM data making them suited candidates for estimating TEL-activity (**Table 1**). Interestingly, *DKC1* (and partly also *RUVBL2*) overexpression associates consistently with unfavorable prognosis in renal, liver, head-neck, endometrial and skin (melanoma) cancers (107, 108). We expect these transcripts to function as potential markers with prognostic impact also in CRC.

The partial influence (PI) of *TERT* on TEL pathway activity is highest in all cancer subtypes in contrast to *TERT* differential expression, presumably due to the stabilizing effect of the interaction partners of *TERT* in its local pathway topology. *TERT* also occupies top positions in the betweenness centrality rankings. These two measures together show that consideration of pathway topology and/or degree of co-regulation will increase the impact of *TERT* as TEL-TMM marker. Please, recall also that MTL levels off at shortest boundary values in cancers, which makes it virtually insensitive to cancer progression, while expression of many TEL-genes is still considerably variable, thus making them potentially more sensitive markers for cancer development (**Supplementary Figure 11**).

Limitations of our study are linked to the relative small sample size, which decreases resolution especially of the MTL and TRV data obtained from whole genome DNA sequencing. On the other hand, our dataset of matched tumor-reference and combined whole genome DNA-seq and RNS-seq of LS-CRC is the only presently available data of its kind, to our best knowledge. So it represents a unique data source of this relatively rare disease (about 3% of bowel cancers). It is well-characterized in terms of subtypes, somatic and constitutional

mutations and transcriptional states (45) and, it is reviewed as state of the art study addressing molecular heterogeneity of LS-CRC and providing novel insights into immune escape mechanisms of carcigogenesis of LS-CRC (46). The latter review emphasizes the need for identification of suitable molecular markers for describing tumor development and heter6ogeneity in these cancer types (46). The present study, despite its relative small size, provides a potential starting point for the search of such markers with focus on telomere biology. Please note also, that sub-stratification into molecular subtypes is an intrinsic problem in molecular cancer studies because they naturally reduce sample size in the strata. On the other hand, G1 and G2 behave similarly concerning telomere lengths what, in turn, increases significance in a combined view on the data (**Supplementary Table 4A**).

The supervised pathway approach restricts our results to a limited number of TMM genes selected and curated based on literature knowledge. Our conclusions regarding TEL/ALT-TMM activation thus refer to expression data and the pathway model applied. In a general sense they are not definite, but are indications of trends that have to be further validation by independent experimental approaches. Because of pleiotropic roles of many of these genes, e.g., related to extra-telomeric cellular functions accompanying telomere shortening, their particular function for TMM remains ambiguous in many cases and requires further studies. Selection, specification and extension of genes considered and adjustment of their interactions in terms of pathway topologies, together with systematic study of other cancer entities, are expected to improve the functional understanding of TMM and its impact in the context of tumor biology. Overall these analyses illustrate the general problem, namely that there is no clear-cut separation between "telomere biology" and other cellular functions. Pathways in general (i.e., not only our TMM pathways), represent models which consider direct interactions between genes and proteins on one hand but on the other hand focus on a definite "cutout" of cellular function which neglects relations to functionalities outside this "window." This is their strength on one hand, but also their weakness. Such pathway models have been proven in many applications because of their focused view, which allows description of selected biological processes by means of definite ingredients. Our approach is only a first step in this direction, which needs improvement in future work. On the other hand, application of TMM-pathway models to "real world" data such as CRC omics data as done here are needed for such improvements.

CONCLUSIONS

The present study demonstrated that genome and transcriptome sequencing can provide a detailed picture of alterations of telomere length, sequence composition and of gene expression changes related to transcriptional regulation of telomere maintenance in selected CRC subtypes. Thereby, gene expression data can provide an alternative to genomic data and/or complementary measure of the telomere status in tumors. Consideration of interaction topologies in pathway analysis provided additional information about the mechanisms of telomere length regulation in addition to standard gene expression analysis. Our study thus provides an example how omics data can support understanding of selected aspects of tumor biology.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the dbGaP database (www.ncbi.nlm. nih.gov/gap) under accession number phs001407.

AUTHOR CONTRIBUTIONS

LN and HB conceived this study, performed analyses, and wrote the paper. LH performed mutation and gene set analysis. All authors contributed to data generation, methods development, and read and approved the final version of the manuscript.

FUNDING

This study was supported by the German Federal Ministry of Education and Science (BMBF) grants LHA (idSEM program: FKZ 031L0026 to HB, ML, and LH), HNPCC-SYS (PTJ grant HNPCCSys 031 6065A to HB, LH, JG, and ML), PathwayMaps (DFG: WTZ ARM II-010 and 01ZX1304A; and State Committe of Science of Armenia: 16GE-025; to HB, LN, LH, HL-W, and AA), and oBIG (FFE-0034 to HB, LH, AA, HL-W, and LN). The authors acknowledge support from the German Research Foundation (DFG) and Leipzig University within the program of Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2019.01172/full#supplementary-material

REFERENCES

1. Palm W, de Lange T. How shelterin protects mammalian telomeres. *Annu Rev Genet*. (2008) 42:301–34. doi: 10.1146/annurev.genet.41.110306.130350
2. de Lange T. Shelterin-mediated telomere protection. *Annu Rev Genet*. (2018) 52:223–47. doi: 10.1146/annurev-genet-032918-021921
3. Victorelli S, Passos JF. Telomeres and cell senescence - size matters not. *EBioMedicine*. (2017) 21:14–20. doi: 10.1016/j.ebiom.2017.03.027
4. Kim W, Shay JW. Long-range telomere regulation of gene expression: telomere looping and telomere position effect over long distances (TPE-OLD). *Differentiation*. (2018) 99:1–9. doi: 10.1016/j.diff.2017.11.005
5. Martínez P, Blasco MA. Replicating through telomeres: a means to an end. *Trends Biochem Sci*. (2015) 40:504–15. doi: 10.1016/j.tibs.2015.06.003
6. Shay JW, Wright WE. Role of telomeres and telomerase in cancer. *Semin Cancer Biol*. (2011) 21:349–53. doi: 10.1016/j.semcancer.2011.10.001

7. Ju Z, Rudolph KL. Telomeres and telomerase in stem cells during aging and disease. In: Volff J-N, editor. *Genome and Disease*. Basel: KARGER (2006). p. 84–103.
8. Hiyama E, Hiyama K. Telomere and telomerase in stem cells. *Br J Cancer*. (2007) 96:1020–4. doi: 10.1038/sj.bjc.6603671
9. Reddel RR. Alternative lengthening of telomeres, telomerase, and cancer. *Cancer Lett*. (2003) 194:155–62. doi: 10.1016/S0304-3835(02)00702-4
10. Cesare AJ, Reddel RR. Alternative lengthening of telomeres: models, mechanisms and implications. *Nat Rev Genet*. (2010) 11:319–30. doi: 10.1038/nrg2763
11. Lovejoy CA, Li W, Reisenweber S, Thongthip S, Bruno J, de Lange T, et al. Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of telomeres pathway. *PLoS Genet*. (2012) 8:e1002772. doi: 10.1371/journal.pgen.1002772
12. De Vitis M, Berardinelli F, Sgura A. Telomere length maintenance in cancer: at the crossroad between telomerase and alternative lengthening of telomeres (ALT). *Int J Mol Sci*. (2018) 19:606. doi: 10.3390/ijms19020606
13. Barthel FP, Wei W, Tang M, Martinez-Ledesma E, Hu X, Amin SB, et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat Genet*. (2017) 49:349–57. doi: 10.1038/ng.3781
14. Dagg RA, Pickett HA, Neumann AA, Napier CE, Henson JD, Teber ET, et al. Extensive proliferation of human cancer cells with ever-shorter telomeres. *Cell Rep*. (2017) 19:2544–56. doi: 10.1016/j.celrep.2017.05.087
15. Napier CE, Huschtscha LI, Harvey A, Bower K, Noble JR, Hendrickson EA, et al. ATRX represses alternative lengthening of telomeres. *Oncotarget*. (2015) 6:16543–58. doi: 10.18632/oncotarget.3846
16. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. *Nature*. (2017) 545:175–80. doi: 10.1038/nature22071
17. Wiestler B, Capper D, Holland-Letz T, Korshunov A, von Deimling A, Pfister SM, et al. ATRX loss refines the classification of anaplastic gliomas and identifies a subgroup of IDH mutant astrocytic tumors with better prognosis. *Acta Neuropathol*. (2013) 126:443–51. doi: 10.1007/s00401-013-1156-z
18. Chiba K, Lorbeer FK, Shain AH, McSwiggen DT, Schruf E, Oh A, et al. Mutations in the promoter of the telomerase gene *TERT* contribute to tumorigenesis by a two-step mechanism. *Science*. (2017) 357:1416–20. doi: 10.1126/science.aao0535
19. Pickett HA, Reddel RR. Molecular mechanisms of activity and derepression of alternative lengthening of telomeres. *Nat Struct Mol Biol*. (2015) 22:875–80. doi: 10.1038/nsmb.3106
20. Garcia-Aranda C, de Juan C, Diaz-Lopez A, Sanchez-Pernaute A, Torres A-J, Diaz-Rubio E, et al. Correlations of telomere length, telomerase activity, and telomeric-repeat binding factor 1 expression in colorectal carcinoma. *Cancer*. (2006) 106:541–51. doi: 10.1002/cncr.21625
21. Henson JD, Reddel RR. Assaying and investigating alternative lengthening of telomeres activity in human cells and cancers. *FEBS Lett*. (2010) 584:3800–11. doi: 10.1016/j.febslet.2010.06.009
22. Zhang J-M, Yadav T, Ouyang J, Lan L, Zou L. Alternative lengthening of telomeres through two distinct break-induced replication pathways. *Cell Rep*. (2019) 26:955–68.e3. doi: 10.1016/j.celrep.2018.12.102
23. Lee M, Napier CE, Yang SF, Arthur JW, Reddel RR, Pickett HA. Comparative analysis of whole genome sequencing-based telomere length measurement techniques. *Methods*. (2017) 114:4–15. doi: 10.1016/j.ymeth.2016.08.008
24. Nersisyan L. Integration of telomere length dynamics into systems biology framework: a review. *Gene Regul Syst Bio*. (2016) 10:GRSB.S39836. doi: 10.4137/GRSB.S39836
25. Nersisyan L, Arakelyan A. Computel: computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS ONE*. (2015) 10:e0125201. doi: 10.1371/journal.pone.0125201
26. Reyes-Uribe P, Paz Adrianzen-Ruesta M, Deng Zhong, Echevarria-Vargas I, Mender I, Saheb S, et al. Exploiting TERT dependency as a therapeutic strategy for NRAS-mutant melanoma. *Oncogene*. (2018) 37:4058–72. doi: 10.1038/s41388-018-0247-7
27. Xu Y, Goldkorn A, Xu Y, Goldkorn A. Telomere and telomerase therapeutics in cancer. *Genes*. (2016) 7:22. doi: 10.3390/genes7060022
28. Bertorelle R, Rampazzo E, Pucciarelli S, Nitti D, De Rossi A. Telomeres, telomerase and colorectal cancer. *World J Gastroenterol*. (2014) 20:1940–50. doi: 10.3748/wjg.v20.i8.1940
29. Bisoffi M, Heaphy CM, Griffith JK. Telomeres: prognostic markers for solid tumors. *Int J Cancer*. (2006) 119:2255–60. doi: 10.1002/ijc.22120
30. Gertler R, Rosenberg R, Stricker D, Friederichs J, Hoos A, Werner M, et al. Telomere length and human telomerase reverse transcriptase expression as markers for progression and prognosis of colorectal carcinoma. *J Clin Oncol*. (2004) 22:1807–14. doi: 10.1200/JCO.2004.09.160
31. Rampazzo E, Bertorelle R, Serra L, Terrin L, Candiotto C, Pucciarelli S, et al. Relationship between telomere shortening, genetic instability, and site of tumour origin in colorectal cancers. *Br J Cancer*. (2010) 102:1300–5. doi: 10.1038/sj.bjc.6605644
32. Balc'h E Le, Grandin N, Demattei M-V, Guyétant S, Tallet A, Pagès J-C, et al. Measurement of telomere length in colorectal cancers for improved molecular diagnosis. *Int J Mol Sci*. (2017) 18:1871. doi: 10.3390/ijms18091871
33. Baichoo E, Boardman LA. Toward a molecular classification of colorectal cancer: the role of telomere length. *Front Oncol*. (2014) 4:158. doi: 10.3389/fonc.2014.00158
34. Fernández-Marcelo T, Sánchez-Pernaute A, Pascua I, De Juan C, Head J, Torres-García A-J, et al. Clinical relevance of telomere status and telomerase activity in colorectal cancer. *PLoS ONE*. (2016) 11:e0149626. doi: 10.1371/journal.pone.0149626
35. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of lynch syndrome: 1895–2015. *Nat Rev Cancer*. (2015) 15:181–94. doi: 10.1038/nrc3878
36. de la Chapelle A, Hampel H. Clinical relevance of microsatellite instability in colorectal cancer. *J Clin Oncol*. (2010) 28:3380–7. doi: 10.1200/JCO.2009.27.0652
37. Lagerstedt Robinson K, Liu T, Vandrovcova J, Halvarsson B, Clendenning M, Frebourg T, et al. Lynch syndrome (hereditary nonpolyposis colorectal cancer) diagnostics. *J Natl Cancer Inst*. (2007) 99:291–9. doi: 10.1093/jnci/djk051
38. Kastrinos F, Stoffel EM, Balmaña J, Steyerberg EW, Mercado R, Syngal S. Phenotype comparison of MLH1 and MSH2 mutation carriers in a cohort of 1,914 individuals undergoing clinical genetic testing in the United States. *Cancer Epidemiol Biomarkers Prev*. (2008) 17:2044–51. doi: 10.1158/1055-9965.EPI-08-0301
39. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci USA*. (1999) 96:8681–6. doi: 10.1073/pnas.96.15.8681
40. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. (2012) 487:330–7. doi: 10.1038/nature11252
41. Dilley RL, Verma P, Cho NW, Winters HD, Wondisford AR, Greenberg RA. Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature*. (2016) 539:54–8. doi: 10.1038/nature20099
42. Mendez-Bermudez A, Royle NJ. Deficiency in DNA mismatch repair increases the rate of telomere shortening in normal human cells. *Hum Mutat*. (2011) 32:939–46. doi: 10.1002/humu.21522
43. Bechter OE, Zou Y, Walker W, Wright WE, Shay JW. Telomeric recombination in mismatch repair deficient human colon cancer cells after telomerase inhibition. *Cancer Res*. (2004) 64:3444–51. doi: 10.1158/0008-5472.CAN-04-0323
44. Omori Y, Nakayama F, Li D, Kanemitsu K, Semba S, Ito A, et al. Alternative lengthening of telomeres frequently occurs in mismatch repair system-deficient gastric carcinoma. *Cancer Sci*. (2009) 100:413–8. doi: 10.1111/j.1349-7006.2008.01063.x
45. Binder H, Hopp L, Schweiger MR, Hoffmann S, Jühling F, Kerick M, et al. Genomic and transcriptomic heterogeneity of colorectal tumours arising in Lynch syndrome. *J Pathol*. (2017) 243:242–54. doi: 10.1002/path.4948
46. Seth S, Ager A, Arends MJ, Frayling IM. Lynch syndrome—cancer pathways, heterogeneity and immune escape. *J Pathol*. (2018) 246:129–33. doi: 10.1002/path.5139
47. Pawlik TM, Raut CP, Rodriguez-Bigas MA. Colorectal carcinogenesis: MSI-H vs. MSI-L. *Dis Markers*. (2004) 20:199–206. doi: 10.1155/2004/368680
48. Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. *Nature*. (1997) 386:623–7. doi: 10.1038/386623a0
49. Clark CR, Maile M, Blaney P, Hellweg SR, Strauss A, Durose W, et al. Transposon mutagenesis screen in mice identifies TM9SF2

as a novel colorectal cancer oncogene. *Sci Rep*. (2018) 8:15327. doi: 10.1038/s41598-018-33527-3
50. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. (2014) 15:550. doi: 10.1186/s13059-014-0550-8
51. Törönen P, Ojala PJ, Marttinen P, Holm L. Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. *BMC Bioinformatics*. (2009) 10:307. doi: 10.1186/1471-2105-10-307
52. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
53. Nersisyan L, Arakelyan A. A transcriptome and literature guided algorithm for reconstruction of pathways to assess activity of telomere maintenance mechanisms. *bioRxiv*. (2017) 200535. doi: 10.1101/200535
54. Nersisyan L. *Telomere Analysis Based on High-Throughput Multi -Omics Data*. (2017) Available online at: urn:nbn:de:bsz:15-qucosa2-162974
55. Nersisyan L, Johnson G, Riel-Mehan M, Pico A, Arakelyan A. PSFC: a pathway signal flow calculator app for cytoscape. *F1000Research*. (2015) 4:480. doi: 10.12688/f1000research.6706.1
56. Nersisyan L, Löffler-Wirth H, Arakelyan A, Binder H. Gene Set- and pathway- centered knowledge discovery assigns transcriptional activation patterns in brain, blood, and colon cancer. *Int J Knowl Discov Bioinforma*. (2014) 4:46–69. doi: 10.4018/IJKDB.2014070104
57. Loeffler-Wirth H, Kreuz M, Hopp L, Arakelyan A, Haake A, Cogliatti SB, et al. A modular transcriptome map of mature B cell lymphomas. *Genome Med*. (2019) 11:27. doi: 10.1186/s13073-019-0637-7
58. Hopp L, Loeffler-Wirth H, Nersisyan L, Arakelyan A, Binder H. Footprints of sepsis framed within community acquired pneumonia in the blood transcriptome. *Front Immunol*. (2018) 9:1620. doi: 10.3389/fimmu.2018.01620
59. Volkan Çakir M, Wirth H, Arakelyan A, Binder H. Dysregulated signal propagation in a MYC-associated boolean gene network in B-cell lymphoma. *Biol Eng Med*. (2017) 2:2–11. doi: 10.15761/BEM.1000115
60. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with cytoscape 3. *Curr Protoc Bioinforma*. (2014) 13:1–24. doi: 10.1002/0471250953.bi0813s47
61. O'Sullivan J, Risques RA, Mandelson MT, Chen L, Brentnall TA, Bronner MP, et al. Telomere length in the colon declines with age: a relation to colorectal cancer? *Cancer Epidemiol Biomarkers Prev*. (2006) 15:573–7. doi: 10.1158/1055-9965.EPI-05-0542
62. Nakamura K, Furugori E, Esaki Y, Arai T, Sawabe M, Okayasu I, et al. Correlation of telomere lengths in normal and cancers tissue in the large bowel. *Cancer Lett*. (2000) 158:179–84. doi: 10.1016/S0304-3835(00)00521-8
63. Lee M, Hills M, Conomos D, Stutz MD, Dagg RA, Lau LMS, et al. Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res*. (2014) 42:1733–46. doi: 10.1093/nar/gkt1117
64. Baird DM, Jeffreys AJ, Royle NJ. Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *EMBO J*. (1995) 14:5433–43. doi: 10.1002/j.1460-2075.1995.tb00227.x
65. Allshire RC, Dempster M, Hastie ND. Human telomeres contain at least three types of G-rich repeat distributed non-randomly. *Nucleic Acids Res*. (1989) 17:4611–27. doi: 10.1093/nar/17.12.4611
66. Lee M, Teber ET, Holmes O, Nones K, Patch A-M, Dagg RA, et al. Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic Acids Res*. (2018) 46:4903–18. doi: 10.1093/nar/gky297
67. Nabetani A, Ishikawa F. Alternative lengthening of telomeres pathway: recombination-mediated telomere maintenance mechanism in human cells. *J Biochem*. (2011) 149:5–14. doi: 10.1093/jb/mvq119
68. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. *Nat Rev Cancer*. (2006) 6:99–106. doi: 10.1038/nrc1802
69. Artandi SE, Attardi LD. Pathways connecting telomeres and p53 in senescence, apoptosis, and cancer. *Biochem Biophys Res Commun*. (2005) 331:881–90. doi: 10.1016/j.bbrc.2005.03.211
70. Lucien G, Wang X. "Extra-Telomeric Effects of Telomerase (hTERT) in Cell Death," in *Apoptosis*
71. Chin L, Artandi SE, Shen Q, Tam A, Lee SL, Gottlieb GJ, et al. p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. *Cell*. (1999) 97:527–38. doi: 10.1016/S0092-8674(00)80762-X
72. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat Commun*. (2017) 8:15107. doi: 10.1038/ncomms15107
73. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, et al. Stromal contribution to the colorectal cancer transcriptome. *Nat Genet*. (2015) 47:312–9. doi: 10.1038/ng.3224
74. Forsythe HL, Jarvis JL, Turner JW, Elmore LW, Holt SE. Stable association of hsp90 and p23, but not hsp70, with active human telomerase. *J Biol Chem*. (2001) 276:15571–4. doi: 10.1074/jbc.C100055200
75. Frohnert C, Hutten S, Wälde S, Nath A, Kehlenbach RH. Importin 7 and Nup358 promote nuclear import of the protein component of human telomerase. *PLoS ONE*. (2014) 9:e88887. doi: 10.1371/journal.pone.0088887
76. Jeong SA, Kim K, Lee JH, Cha JS, Khadka P, Cho H-S, et al. Akt-mediated phosphorylation increases the binding affinity of hTERT for importin to promote nuclear translocation. *J Cell Sci*. (2015) 128:2951. doi: 10.1242/jcs.176453
77. Fernandez-Garcia I, Marcos T, Muñoz-Barrutia A, Serrano D, Pio R, Montuenga LM, et al. Multiscale *in situ* analysis of the role of dyskerin in lung cancer cells. *Integr Biol*. (2013) 5:402–13. doi: 10.1039/c2ib20219k
78. Alawi F, Lin P, Ziober B, Patel R. Correlation of dyskerin expression with active proliferation independent of telomerase. *Head Neck*. (2011) 33:1041–51. doi: 10.1002/hed.21579
79. Zhong Z-H, Jiang W-Q, Cesare AJ, Neumann AA, Wadhwa R, Reddel RR. Disruption of telomere maintenance by depletion of the MRE11/RAD50/NBS1 complex in cells that use alternative lengthening of telomeres. *J Biol Chem*. (2007) 282:29314–22. doi: 10.1074/jbc.M701413200
80. Jiang W-Q, Zhong Z-H, Henson JD, Neumann AA, Chang AC-M, Reddel RR. Suppression of alternative lengthening of telomeres by Sp100-mediated sequestration of the MRE11/RAD50/NBS1 complex. *Mol Cell Biol*. (2005) 25:2708–21. doi: 10.1128/MCB.25.7.2708-2721.2005
81. Scherer M, Stamminger T. Emerging role of PML nuclear bodies in innate immune signaling. *J Virol*. (2016) 90:5850–4. doi: 10.1128/JVI.01979-15
82. Sahin U, Ferhi O, Jeanne M, Benhenda S, Berthier C, Jollivet F, et al. Oxidative stress-induced assembly of PML nuclear bodies controls sumoylation of partner proteins. *J Cell Biol*. (2014) 204:931–45. doi: 10.1083/jcb.201305148
83. Wang Z, Deng Z, Tutton S, Lieberman PM, Wang Z, Deng Z, et al. The telomeric response to viral infection. *Viruses*. (2017) 9:218. doi: 10.3390/v9080218
84. Ueno H, Hashiguchi Y, Shimazaki H, Shinto E, Kajiwara Y, Nakanishi K, et al. Objective criteria for crohn-like lymphoid reaction in colorectal cancer. *Am J Clin Pathol*. (2013) 139:434–41. doi: 10.1309/AJCPWHUEFTGBWKE4
85. Blasco MA. The epigenetic regulation of mammalian telomeres. *Nat Rev Genet*. (2007) 8:299–309. doi: 10.1038/nrg2047
86. Conomos D, Reddel RR, Pickett HA. NuRD–ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination. *Nat Struct Mol Biol*. (2014) 21:760–70. doi: 10.1038/nsmb.2877
87. Wu G, Jiang X, Lee W-H, Chen P-L. Assembly of functional ALT-associated promyelocytic leukemia bodies requires nijmegen breakage syndrome 1. *Cancer Res*. (2003) 63:2589–95.
88. Naderlinger E, Holzmann K. Epigenetic regulation of telomere maintenance for therapeutic interventions in gliomas. *Genes*. (2017) 8:145. doi: 10.3390/genes8050145
89. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform*. (2018) 19:575–92. doi: 10.1093/bib/bbw139
90. Cacchione S, Biroccio A, Rizzo A. Emerging roles of telomeric chromatin alterations in cancer. *J Exp Clin Cancer Res*. (2019) 38:21. doi: 10.1186/s13046-019-1030-5
91. Xu X, Qu K, Pang Q, Wang Z, Zhou Y, Liu C. Association between telomere length and survival in cancer patients: a meta-analysis and review of literature. *Front Med*. (2016) 10:191–203. doi: 10.1007/s11684-016-0450-2

92. Shay JW. Role of telomeres and telomerase in aging and cancer. *Cancer Discov*. (2016) 6:584–93. doi: 10.1158/2159-8290.CD-16-0062
93. Okamoto K, Seimiya H, Okamoto K, Seimiya H. Revisiting telomere shortening in cancer. *Cells*. (2019) 8:107. doi: 10.3390/cells8020107
94. Basu N, Skinner HG, Litzelman K, Vanderboom R, Baichoo E, Boardman LA. Telomeres and telomere dynamics: relevance to cancers of the GI tract. *Expert Rev Gastroenterol Hepatol*. (2013) 7:733–48. doi: 10.1586/17474124.2013.848790
95. Plentz RR, Wiemann SU, Flemming P, Meier PN, Kubicka S, Kreipe H, et al. Telomere shortening of epithelial cells characterises the adenoma-carcinoma transition of human colorectal cancer. *Gut*. (2003) 52:1304–7. doi: 10.1136/gut.52.9.1304
96. Takagi S, Kinouchi Y, Hiwatashi N, Nagashima F, Chida M, Takahashi S, et al. Relationship between microsatellite instability and telomere shortening in colorectal cancer. *Dis Colon Rectum*. (2000) 43:S12–7. doi: 10.1007/BF02237220
97. Martinez P, Siegl-Cachedenier I, Flores JM, Blasco MA. MSH2 deficiency abolishes the anticancer and pro-aging activity of short telomeres. *Aging Cell*. (2009) 8:2–17. doi: 10.1111/j.1474-9726.2008.00441.x
98. Conomos D, Stutz MD, Hills M, Neumann AA, Bryan TM, Reddel RR, et al. Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J Cell Biol*. (2012) 199:893–906. doi: 10.1083/jcb.201207189
99. Fasold M, Stadler PF, Binder H. G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration. *BMC Bioinformatics*. (2010) 11:207. doi: 10.1186/1471-2105-11-207
100. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*. (2016) 164:550–63. doi: 10.1016/j.cell.2015.12.028
101. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. *Science*. (2013) 339:957–9. doi: 10.1126/science.1229259
102. Heidenreich B, Kumar R. TERT promoter mutations in telomere biology. *Mutat Res Mutat Res*. (2017) 771:15–31. doi: 10.1016/j.mrrev.2016.11.002
103. Viceconte N, Dheur M-S, Majerova E, Pierreux CE, Baurain J-F, van Baren N, et al. Highly aggressive metastatic melanoma cells unable to maintain telomere length. *Cell Rep*. (2017) 19:2529–43. doi: 10.1016/j.celrep.2017.05.046
104. Stern JL, Paucek RD, Huang FW, Ghandi M, Nwumeh R, Costello JC, et al. Allele-specific DNA methylation and its interplay with repressive histone marks at promoter-mutant TERT genes. *Cell Rep*. (2017) 21:3700–7. doi: 10.1016/j.celrep.2017.12.001
105. Hrdlickova R, Nehyba J, Bose HR. Alternatively spliced telomerase reverse transcriptase variants lacking telomerase activity stimulate cell proliferation. *Mol Cell Biol*. (2012). doi: 10.1128/MCB.00550-12
106. Wong MS, Wright WE, Shay JW. Alternative splicing regulation of telomerase: a new paradigm? *Trends Genet*. (2014) 30:430–8. doi: 10.1016/j.tig.2014.07.006
107. DKC1. *Hum Protein Atlas*. Available online at: https://www.proteinatlas.org/ENSG00000130826-DKC1/pathology (accessed April 27, 2019).
108. RUVBL2. *Hum Protein Atlas*. Available onlin at: https://www.proteinatlas.org/ENSG00000183207-RUVBL2/pathology (accessed April 27, 2019).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fgene.2019.01408

Identification of Prognostic Genes in Leiomyosarcoma by Gene Co-Expression Network Analysis

Jun Yang[1], *Cuili Li*[1], *Jiaying Zhou*[1], *Xiaoquan Liu*[1] *and Shaohua Wang*[2*]

[1] Department of Pediatrics, The University of Hong Kong-Shenzhen Hospital, ShenZhen, China, [2] Department of Pediatrics, Women and Children Health Institute of FuTian, University of South China, ShenZhen, China

Background/Aims: Leiomyosarcoma (LMS) is a tumor derived from malignant mesenchymal tissue associated with poor prognosis. Determining potential prognostic markers for LMS can provide clues for early diagnosis, recurrence, and treatment.

Methods: RNA sequence data and clinical features of 103 LMS were obtained from the Cancer Genome Atlas (TCGA) database. Application Weighted Gene Co-Expression Network Analysis (WGCNA) was used to construct a free-scale gene co-expression network, to study the interrelationship between its potential modules and clinical features, and to identify hub genes in the module. The hub gene function was verified by an external database.

Results: Twenty-four co-expression modules were constructed using WGCNA. A dark red co-expression module was found to be significantly associated with disease recurrence. Functional enrichment analysis and GEPIA and ONCOMINE database analyses demonstrated that hub genes CDK4, CCT2, and MGAT1 may play an important role in LMS recurrence.

Conclusion: Our study constructed an LMS co-expressing gene module and identified prognostic markers for LMS recurrence detection and treatment.

Keywords: leiomyosarcoma, prognosis, weighted gene co-expression network analysis (WGCNA), TCGA, recurrence

OPEN ACCESS

Edited by:
Barbara Karen Dunn, National Institutes of Health, United States

Reviewed by:
*Akshay Bhinge, Genome Institute of Singapore (A*STAR), Singapore*
Xiangqian Guo, Henan University, China

***Correspondence:**
Shaohua Wang
lwwsbs22@163.com

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: *21 February 2019*
Accepted: *24 December 2019*
Published: *04 February 2020*

Citation:
Yang J, Li C, Zhou J, Liu X and Wang S (2020) Identification of Prognostic Genes in Leiomyosarcoma by Gene Co-Expression Network Analysis. Front. Genet. 10:1408. doi: 10.3389/fgene.2019.01408

INTRODUCTION

Leiomyosarcoma (LMS) is a highly malignant mesenchymal-derived tumor with varying degrees of smooth muscle differentiation, accounting for approximately 10% of soft tissue sarcomas (Noujaim et al., 2015; Pautier et al., 2015). These tumors occur mainly in adults in any body location and are associated with very high mortality. Leiomyosarcoma is divided into a variety of pathological subtypes according to cell morphology and molecular atypia, including typical leiomyosarcoma, epithelioid leiomyosarcoma, and pleomorphic leiomyosarcoma. Because this type of tumor is prone to recurrence and metastasis, it often has invasive clinical characteristics and poor prognosis. The 5-year recurrence rate is less than 40% (Serrano and George, 2013). Although many genes and signaling pathways have been identified to improve detection and treatment of LMS, surgical removal of tumors is currently the most effective way to treat leiomyosarcoma. Poor prognosis of

LMS is related to a higher degree of malignancy, larger tumor volume, and deeper tumor site (Hayashi et al., 2010; Ognjanovic et al., 2012; Croce and Chibon, 2015). Therefore, identification of new biomarkers to assess malignancy and prognosis of LMS is essential.

Weighted correlation network analysis (WGCNA) is a systematic biological approach used to describe the pattern of gene association between different samples. WGCNA analysis uses correlation coefficient weights to make the connections between genes in the network obey scale-free networks, which is more biologically significant (Langfelder and Horvath, 2008). WGCNA can be used to identify highly synergistically altered gene sets and identify candidate biomarker genes or therapeutic targets based on the association of gene set connectivity and phenotype (Radulescu et al., 2018). Compared to genes that only focus on differential expression, WGCNA uses thousands of the most variable genes or all of the genes to identify the set of genes of interest and conducts a significant association analysis with the phenotype. WGCNA may make full use of information, and to convert thousands of genes and phenotypes into several gene sets and phenotypes, eliminating the need for multiple hypothesis testing (Zuo et al., 2018).

In this study, we constructed a co-expression network of LMS through WGCNA to systematically analyze the pathogenesis of LMS and tumorigenesis. Our goal is to study new and key biomarkers and to develop a better understanding of the molecular mechanisms of LMS to provide new strategies for diagnosis and treatment of diseases.

MATERIALS AND METHODS

Data Collection

The mRNA sequence data and corresponding clinical traits of LMS were downloaded from the TCGA database (https://tcga-data.nci.nih.gov/tcga/), which contained 103 tumor tissues. Gene symbol annotation information was used to match probes with corresponding genes. TCGA was publicly available and in an open access platforms. As a result, ethics committee approval was not required.

Co-Expression Network Construction With WGCNA and Target Prediction

The WGCNA algorithm runs in the R software package (http://www.r-project.org/) to assess the importance of genes and their associated modules by calculation the correlation coefficient between any two genes (Person Coefficient). To measure whether two genes have similar expression patterns, screening is performed and values above a pre-determined threshold are considered similar. WGCNA analysis uses the correlation coefficient weighting value, which is the N^{th} power of the gene correlation coefficient, so that the connections between the genes in the network obey the scale-free networks, which is more biologically significant. A hierarchical clustering tree was constructed based on the weighted correlation coefficients of genes. Genes were classified according to expression patterns, and genes with similar patterns were classified into one module. Different branches of the cluster tree represent different gene modules, and different colors represent different modules. This strategy allows for tens of thousands of genes can be divided into dozens of modules based on gene expression patterns, which is a process of extracting information. After weighted correlation analysis, we predicted target genes using a co-expression network produced using Cytoscape 3.7.0 software.

Construct Module-Trait Relationships of LMS

Gene modules are linked to the traits of the study to screen for key gene modules. We used "module eigenvalues" to represent the combined value of the gene set expression of the module. Therefore, each module can be associated with a trait by the eigenvector of the module and the correlation coefficient of the phenotype or the saliency P value of the module. In addition, the modules do not exist in isolation, but are related to each other. Using a network heat map, the connections between the trait association module and other modules can be visualized.

Functional Enrichment Analysis of Co-Expression Module

To explore the function of genes in key co-expression modules, we uploaded the data to DAVID for analysis. DAVID is an online database (https://david.ncifcrf.gov/) (Tang et al., 2017; Nagy et al., 2018). It is a classic gene enrichment analysis website, mainly used for differential gene function and pathway enrichment analysis.

Identification of Hub Genes In Key Module

After screening the key gene modules associated with the traits, the gene co-expression network map was drawn based on the relationships of the genes within the module. This network diagram belongs to the scale-free network. Mathematically, for a network graph, each node is given the concept of a degree, and the degree of a point refers to the number of edges associated with that point. In a scale-free network visualized by Cytoscape (3.7.0), the degree of a few nodes is significantly higher than the average point, and these points become hubs. A small number of hubs are associated with other nodes to form the entire network. The gene in the gene module that regulates the network center is the hub gene. At last, we decided the key genes through K-M survival analysis in the GEPIA database (http://gepia.cancer-pku.cn/).

Validation of the Key Genes

We validated the function of candidate genes through a public databases, the ONCOMINE database (https://www.oncomine.org/) (Tang et al., 2017). Then, the overall survival and event-free survival analysis of hub genes were performed using Kaplan-Meier curve in Kaplan Meier plotter (https://kmplot.com/analysis/index.php?p=background) (Nagy et al., 2018) and LOGpc (Long-term Outcome and Gene Expression Profiling Database of pan-cancers) (http://bioinfo.henu.edu.cn/DatabaseList.jsp) (Wang et al., 2019). Finally, we performed multi-factor COX analysis on three key genes, established a

risk model, and performed survival analysis and model identification.

RESULTS

Data Preprocessing

Gene annotation of gene expression data obtained from TCGA, matching probes and genes, removal of probes matching multiple genes, and gene annotation of the genes were matched by multiple probes using the median value as the final expression value. A total of 20,098 genes were identified. We calculated the variance of each gene and then selected the top 25% (5,025) of genes with the largest variance for WGCNA and sample cluster analysis (**Figure 1**).

Construction of Co-Expression Modules

A gene co-expression network was constructed using weighted expression correlation. The soft threshold power value was used for initial screening. When the soft threshold power was equal to 4, the degree of independence reached 0.9 and the average connectivity was higher (**Figure 1**). Therefore, based on the weighted correlation, the WGCNA package automatically constructed a co-expression network, performed hierarchical clustering analysis, and segmented the clustered results according to the predetermined thresholds to obtain different gene modules. Of all the genes in the LMS network, 4,255 were assigned to 24 modules (**Table 1**), and the remaining 770 genes were assigned to the same "gray" module (**Figure 2**) and were included in the heat map. Branches of cluster trees and different color represent different clustering modules.

Correlation Between Modules and Identification of Key Modules

We calculated the eigengenes in-modules connectivity and clustered them to study the co-expression relationships of all modules. The results showed that each module was independent of the others, demonstrating the high degree of independence between modules and the relative independence of gene expression in each module. A heat map drawn from adjacent relationships showed similar results. The dark module ME was highly correlated with recurrence compared to other modules, suggesting that the dark module may play a key role in disease recurrence (**Figures 3** and **4**).

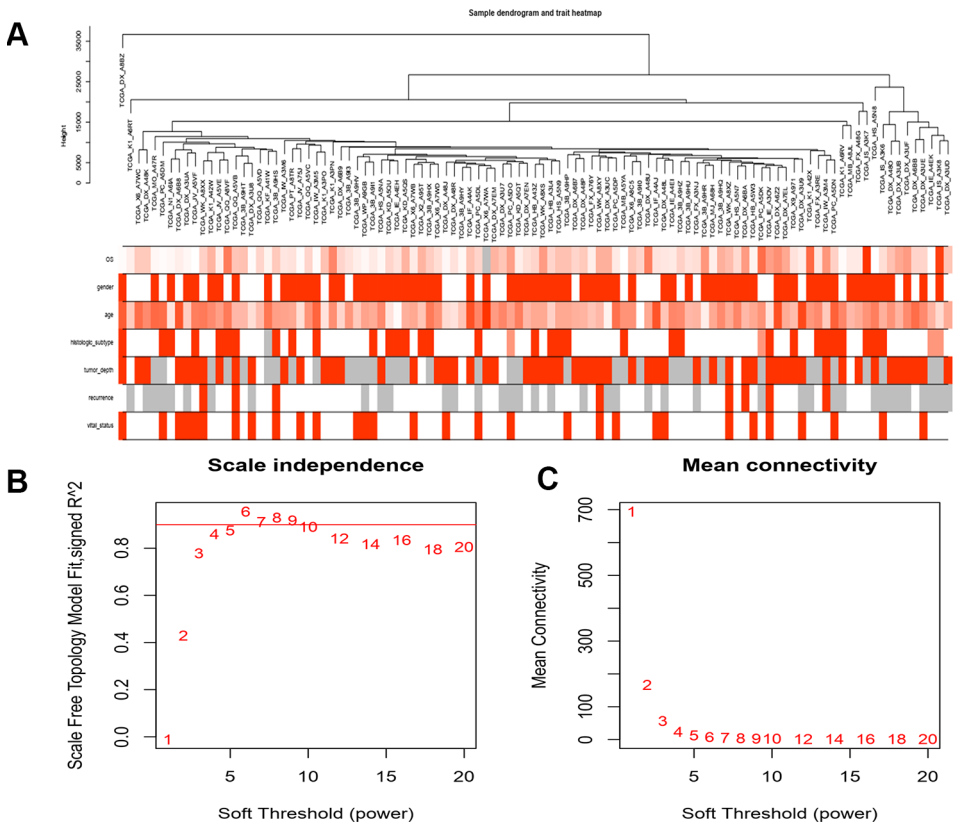


FIGURE 1 | Clustering of samples and determination of soft-thresholding power. **(A)** The clustering was based on the expression data of LMS, which contained 103 LMS tumor tissue. The top 25% genes with the highest SD values were used for the analysis by WGCNA. **(B)** Analysis if the scale free fit index for various softthresholding powers (β). **(C)** Analysis of the mean connectivity of various soft-thresholding powers. In all, 4 was the most fit power value.

TABLE 1 | Co-expressions modules.

Module color	Genes
Black	144
Blue	559
Brown	412
Cyan	112
Dark green	63
Dark gray	45
Dark red	68
Dark turquoise	58
Green	221
Green yellow	124
Gray	770
Gray 60	89
Light cyan	93
Light green	82
Light yellow	74
Magenta	132
Midnight blue	105
Pink	134
Purple	131
Red	151
Royal blue	68
Salmon	113
Tan	115
Turquoise	816
Yellow	346

Function Enrichment Analysis

To clarify the gene functions in the modules, we performed gene ontology enrichment analysis of the identified genes using DAVID, and explored combination of genes related to biological processes (BP), molecular functions (MF), and cellular components (CC) in key modules. (Details of GO enrichment are given in **Table 2**). GO analysis showed that these genes are involved in the components of the cell, embryo development, and transcription, and play an important role in the biological processes of cell division, signal transduction, and transcriptional regulation. The result of functional enrichment analysis showed that genes associated with biology processes were mainly enriched in GO:0060070 (canonical Wnt signaling pathway), GO:0009636 (response to toxic substance), GO:0060349 (bone morphogenesis), GO:0001657 (ureteric bud development), GO:0045892 (negative regulation of transcription, DNA-templated). Genes associated with Molecular Function were enriched in in GO:0043237 (laminin-1 binding), GO:0008201 (heparin binding), GO:0001948 (glycoprotein binding), GO:0005578 (proteinaceous extracellular matrix), and GO:0005737 (cytoplasm). According to the Kyoto Gene and Genomic Encyclopedia (KEGG) pathway analysis, the dark red module genes were mainly enriched in the p53 signaling pathway and the bladder cancer signaling pathway (**Table 3** and **Figure 4**).

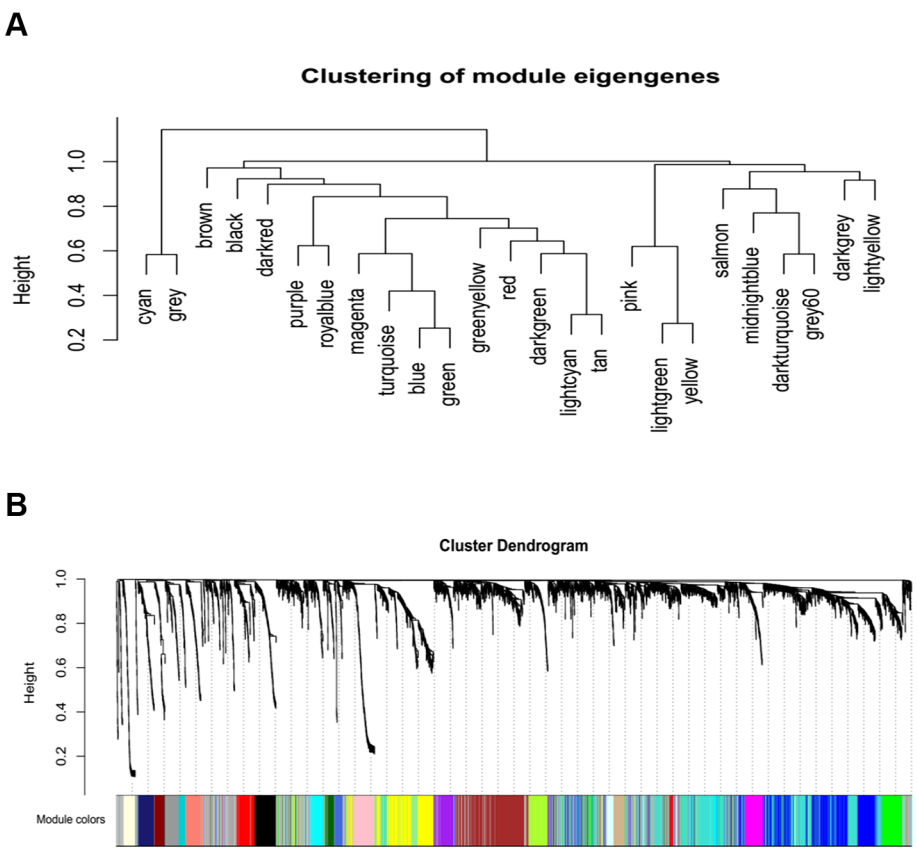


FIGURE 2 | Construction of co-expression modules by WGCNA package in R. **(A)** The cluster dendrogram of module eigengenes. **(B)** The cluster dendrogram of genes. Each branch in the figure represents one gene, and every color below represents one co-expression module.

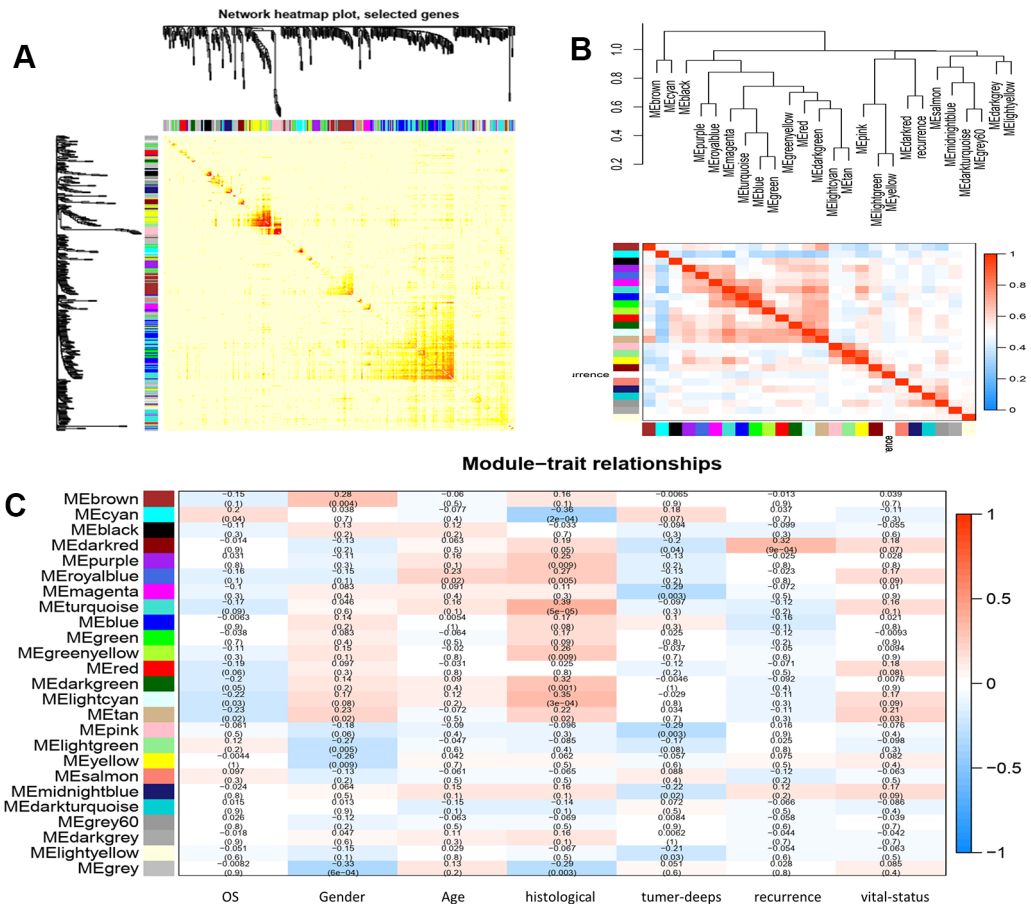


FIGURE 3 | (A) interaction relationship analysis of co-expression genes. Different colors of horizontal axis and vertical axis represent different modules. The brightness of yellow in the middle represents the degree of connectivity of different modules. There was no significant difference in interactions among different modules, indicating a high-scale independence degree among these modules. **(B)** hierarchical clustering of module hub genes that summarize the modules yielded in the clustering analysis and heat map plot of the adjacencies in the hub gene network. **(C)** heat map of the correlation between module eigengenes and the clinical traits of LMS. The dark red module was the most positively correlated with recurrence of disease.

Practical ClueGo was used for visual analysis of KEGG pathway (Bindea et al., 2013).

Identification of Hub Genes

We used Cytoscape software to visualize the dark red module network to build the module and calculate connectivity within the module (Shannon et al., 2003). Genes with high connectivity were identified as hub genes based on connectivity within the module. Genes with significant survival results were selected and sorted by node degree (**Figure 5**). Twelve genes in the selected modules were considered hub genes: TSFM, AATF, BBS10, CDK4, CTDSP2, PLAGL1, DYRK2, FGFR3, CNOT2, METTL1, CCT2, and MGAT1. These hub genes were selected using cytoHubba (Chin et al., 2014). We used GEPIA (http://gepia.cancer-pku.cn/) to perform survival analysis on these hub genes to determine their biological significance (Tang et al., 2017). GEPIA was used to verify the expression characteristics of the twelve genes selected. Among these genes, CDK4, CCT2, and MGAT1 were associated with overall survival and recurrence-free survival, and the expression levels of these three genes were significantly higher in tumor tissues (**Figure 6**). Therefore, these genes were identified as key genes.

Validation of Key Genes

Using the data from ONCOMINE database (https://www.oncomine.org/), we noted that leiomyosarcoma patients who had an association of genomic alterations in CDK4,CCT2, and MGAT1 (**Figure 7**). The expression of the three key genes in the dark red module positively correlated with the disease state. Oncomine analysis of cancer vs. normal tissue showed that cdk4, cct2, and MGAT1 were significantly overexpressed in leiomyosarcoma in the different datasets (**Figure 8**) (Quade et al., 2004; Detwiller et al., 2005; Nakayama et al., 2007; Barretina et al., 2010; Chibon et al., 2010). Using the data from Kaplan

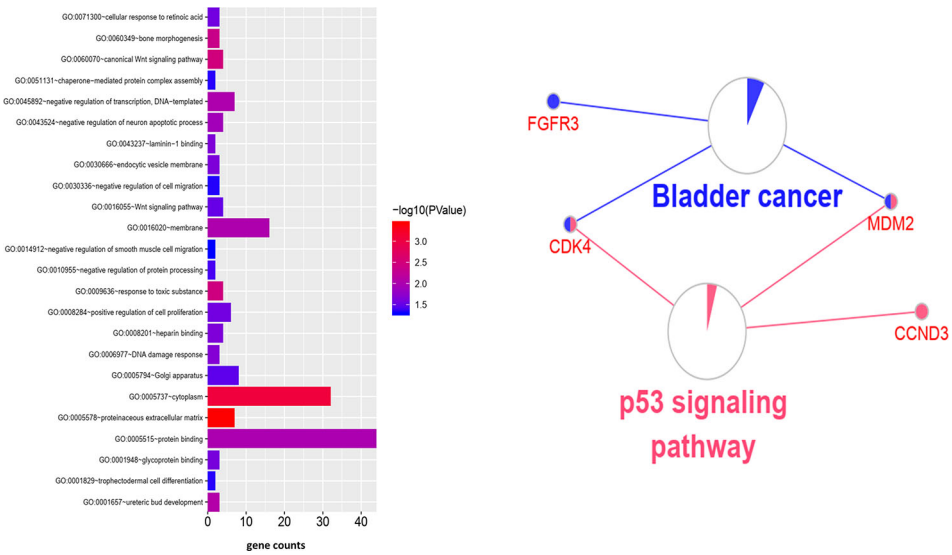


FIGURE 4 | Go and KEGG enrichment analysis of genes in dark red module.

TABLE 2 | GO enrichment analysis of genes in co-expression modules.

Category	ID	Term	Count	P value
BP	GO:0060070	Canonical Wnt signaling pathway	4	3.4092E-03
BP	GO:0009636	Response to toxic substance	4	3.6460E-03
BP	GO:0060349	Bone morphogenesis	3	4.2975E-03
BP	GO:0001657	Ureteric bud development	3	8.3898E-03
BP	GO:0045892	Negative regulation of transcription, DNA-templated	7	9.3260E-03
BP	GO:0043524	Negative regulation of neuron apoptotic process	4	1.2264E-02
BP	GO:0006977	DNA damage response	3	2.1347E-02
BP	GO:0071300	Cellular response to retinoic acid	3	2.6763E-02
BP	GO:0008284	Positive regulation of cell proliferation	6	2.6878E-02
BP	GO:0016055	Wnt signaling pathway	4	3.0529E-02
BP	GO:0010955	Negative regulation of protein processing	2	3.9253E-02
BP	GO:0001829	Trophectodermal cell differentiation	2	4.6225E-02
BP	GO:0051131	Chaperone-mediated protein complex assembly	2	4.6225E-02
BP	GO:0030336	Negative regulation of cell migration	3	4.6720E-02
BP	GO:0014912	Negative regulation of smooth muscle cell migration	2	4.9693E-02
CC	GO:0005578	proteinaceous extracellular matrix	7	3.8200E-04
CC	GO:0005737	Cytoplasm	32	8.7000E-04
CC	GO:0016020	Membrane	16	9.3874E-03
CC	GO:0030666	Endocytic vesicle membrane	3	2.3207E-02
CC	GO:0005794	Golgi apparatus	8	3.3526E-02
MF	GO:0005515	Protein binding	44	1.0017E-02
MF	GO:0043237	Laminin-1 binding	2	2.2536E-02
MF	GO:0008201	heparin binding	4	2.2845E-02
MF	GO:0001948	Glycoprotein binding	3	2.5251E-02

TABLE 3 | KEGG analysis of genes in co-expression modules.

Term	Count	P value	Genes
hsa05219: Bladder cancer	3	0.009534445	FGFR3, MDM2, CDK4
hsa05200: Pathways in cancer	6	0.01219952	WNT10B, FGFR3, BIRC7, MDM2, WNT11, CDK4
hsa04550: Signaling pathways regulating pluripotency of stem cells	4	0.013674563	WNT10B, FGFR3, WNT11, KLF4
hsa04115: p53 signaling pathway	3	0.02427143	CCND3, MDM2, CDK4
hsa05205: Proteoglycans in cancer	4	0.034788315	WNT10B, SDC1, MDM2, WNT11

Meier plotter and LOGpc, we noted that leiomyosarcoma patients who had an association of genomic alterations in CDK4, CCT2, and MGAT1 showed reductions in overall and disease-free survival. However, those observations were statistically significant for overall survival time and no statistically significant for event-free survival (**Figure 9**). Finally, through multi-factor COX analysis of the three genes, we obtained the risk prediction formula, risk score = CDK4* 0.00848 + MAGT1*(−0.01012) (**Table 4**). The model is analyzed for survival and the ROC value is calculated for verification (**Figure 10**).

DISCUSSION

Leiomyosarcoma, which occurs in smooth muscle connective tissue, accounts for ten percent of all soft tissue sarcomas. LMS is malignant and exhibits a high degree of invasiveness, high

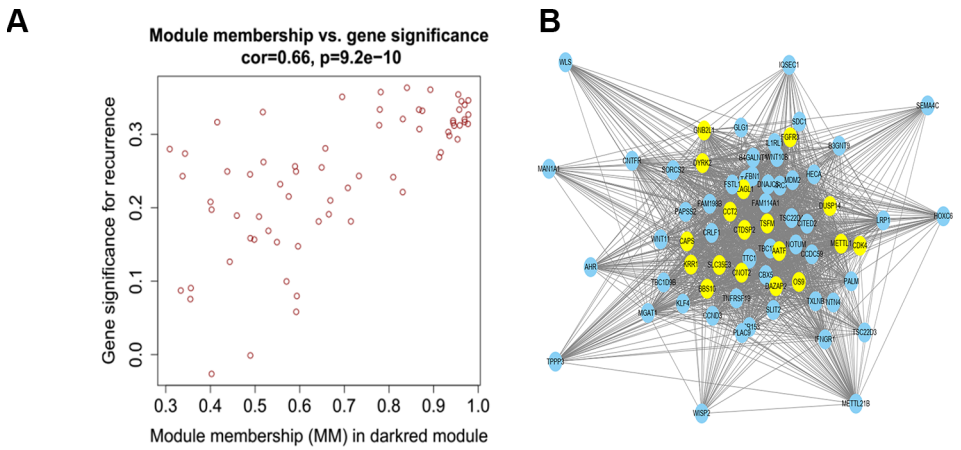


FIGURE 5 | (A) Scatter plot of module eigengenes in the dark red module. **(B)** The hub genes in the dark red module and node size is correlated with connectivity of the gene by degree. Hubgene is represented as a bright yellow node in **(B).**

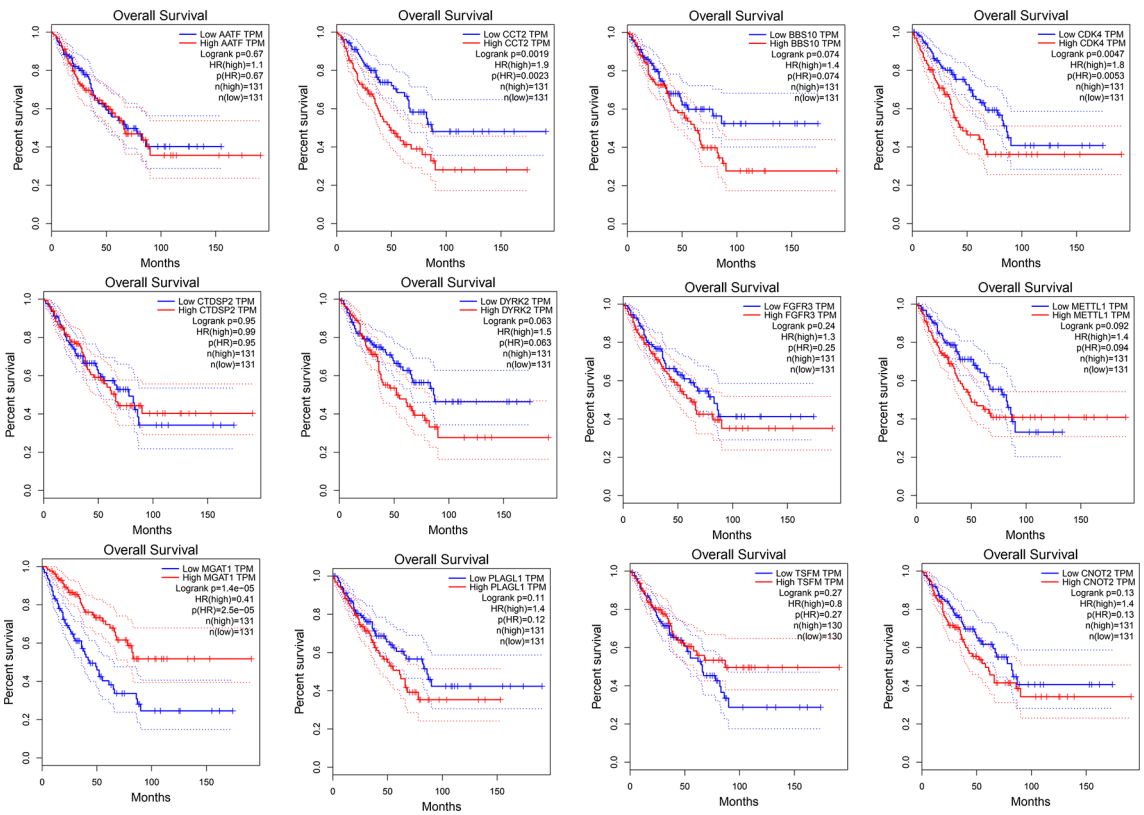


FIGURE 6 | Survival analysis of 12 hub genes identified by WGCNA.

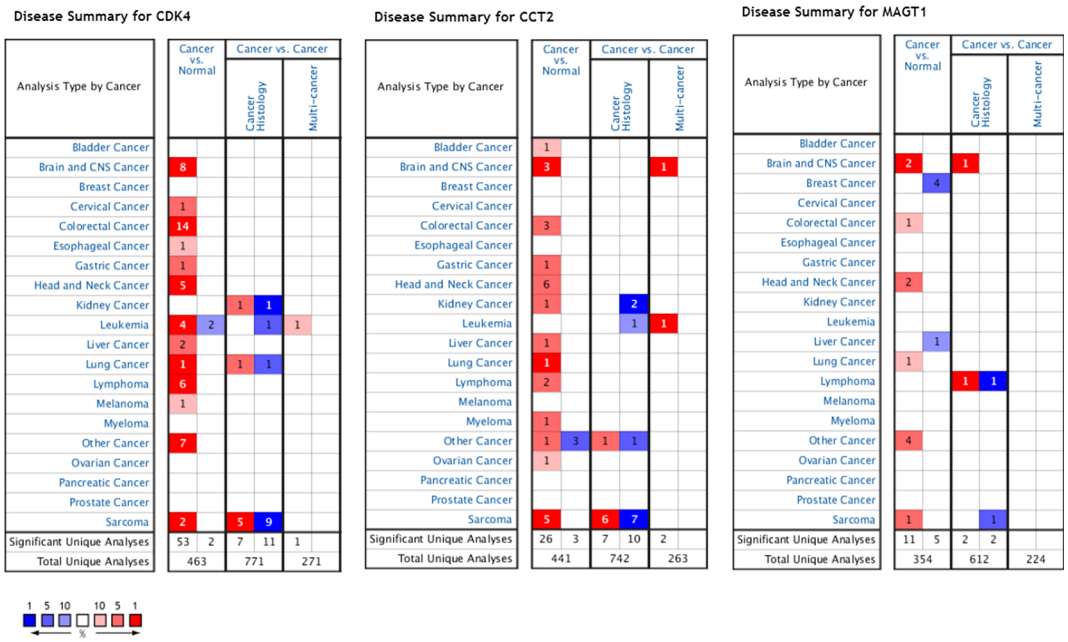


FIGURE 7 | Expression profiles and analysis of cancer vs. normal tissue for CDK4, CCT2, and MAGT1 in human cancers analyzed using Oncomine.

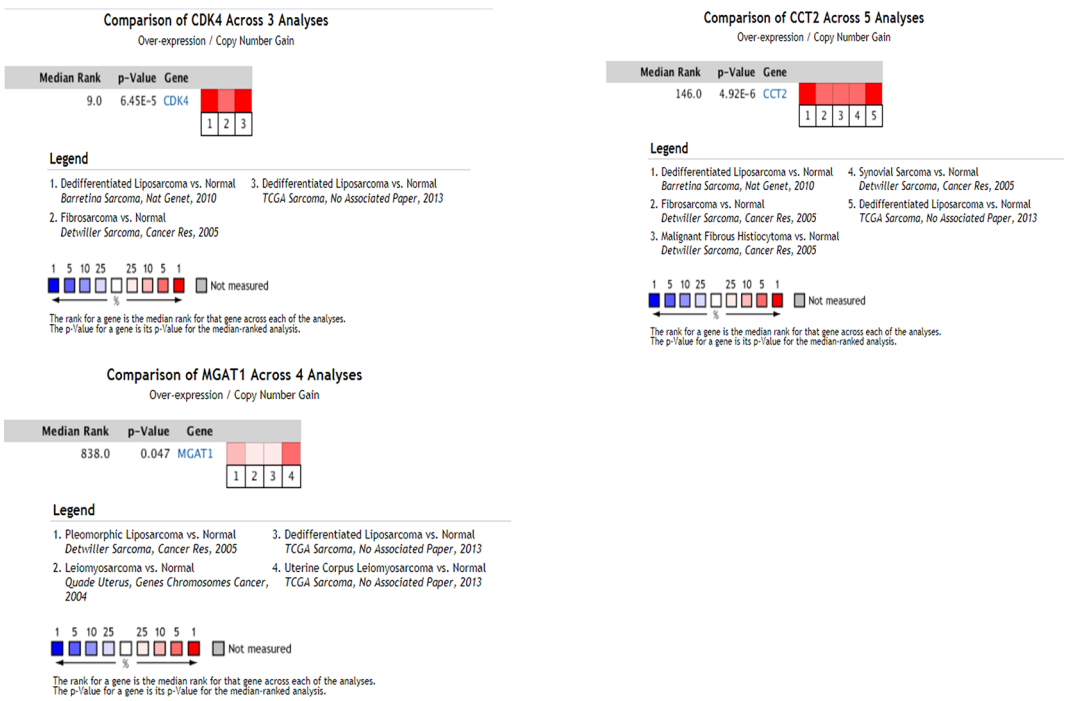


FIGURE 8 | Oncomine analysis of cancer vs. normal tissue of CDK4,CCT2, and MAGT1. Heat maps of CDK4, CCT2, and MGAT1 gene expression in clinical sarcoma samples vs. normal tissues.

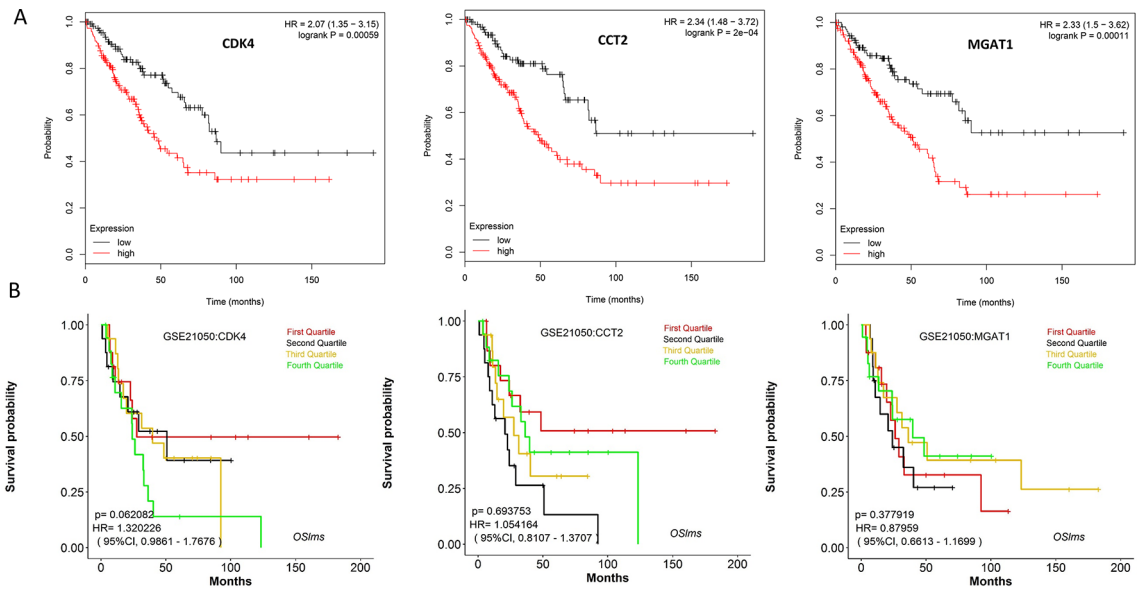


FIGURE 9 | (A) Overall survival analyses of key genes were performed using Kaplan-Meier plotter **(B)** Event-free survival of key genes were performed using LOGpc. ($P < 0.05$ was considered statistically significant.)

TABLE 4 | COX analysis of key genes.

id	Coef	HR	HR.95L	HR.95H	P value
CDK4	0.008487993	1.008524118	1.000962917	1.016142437	0.027061684
MGAT1	−0.010120172	0.989930865	0.977568773	1.002449286	0.114468861

recurrence rate, and high mortality (George et al., 2018). LMS can occur in any location of the body such as in the extremities, small intestine, or retroperitoneal space. As LMS is most common in the uterus, it can be classified as uterine LMS (ULMS) or non-uterine LMS (NULMS) (Guo et al., 2015). ULMS is highly aggressive and is not sensitive to chemotherapy and radiation therapy. Surgical resection is currently the best treatment. The median survival of individuals with NULMS and ULMS is less than 5 years (Eriksson, 2010). In-depth study of the biological behavior and potential molecular mechanisms of LMS is of great significance for improving the efficacy and prognosis of LMS (Pautier et al., 2015; Schoffski et al., 2016; Mir et al., 2016; Tawbi et al., 2017; Gronchi et al., 2017).

In this study, we assessed gene expression to identify potential biomarkers for LMS using WGCNA. Twenty-four co-expression modules were constructed for 5,025 genes from 103 human LMS samples. Because WGCNA focuses on the association between co-expression modules and clinical features, the results are more reliable and biologically meaningful. Genes that are functionally related to each other are clustered together in the same module. Thus, WGCNA can identify biologically relevant modules and central genes that can ultimately become biomarkers for detection or treatment. We found that the dark red module was most significantly associated with disease recurrence. Using a total of 68 genes in the dark red module were screened. GO and KEGG analyses showed that these genes are involved in the components of the cell, embryo development, and transcription, and play an important role in the biological processes of cell division, signal transduction, and transcriptional regulation. We believe that the dark red module was the most important module for characterization of the LMS recurrence mechanism.

Further analysis of the dark red module showed that three genes (CDK4, CCT2, and MGAT1) significantly correlated with survival analysis and were identified as hub genes. The hub genes were further validated in GEPIA and ONCOMINE. Some studies have reported that these three key genes are cancer-associated genes involved in mitotic regulation in cancer cells and inhibition of cell proliferation, which may contribute to tumorigenesis and malignant phenotype. CDK4 is an important effector of the P53 signaling pathway. CDK4 encodes a member of the Ser/Thr protein kinase family, which is important for cell cycle G1 progression. Mutations of this gene and its related proteins, including D-type cyclins, p16 (INK4a), and retinoblastoma gene product (Rb), have been found to be involved in tumorigenesis in a variety of cancers. Multiple polyadenylation sites of this gene have been reported (O'Leary et al., 2016). Increased expression of CDK4 is associated with

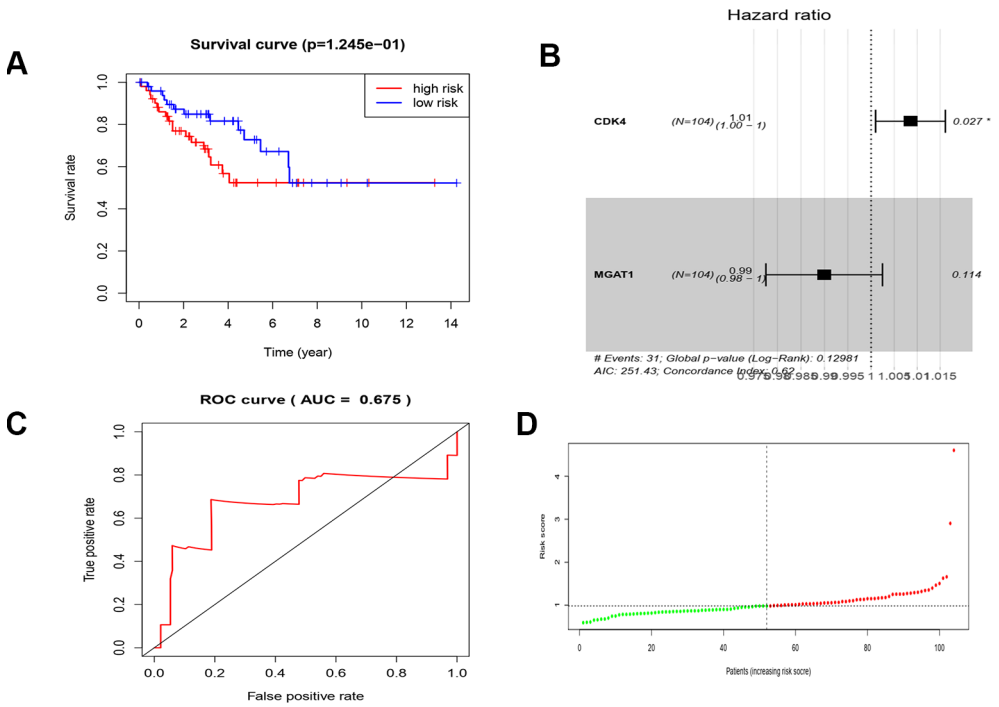


FIGURE 10 | COX analysis of the key genes. **(A)** Survival analysis of high risk and low risk. **(B)** The hazard ratio of key genes of CDK4 and MGAT1. **(C)** ROC curve and AUC value. **(D)** Risk score of the patients.

advanced soft tissue sarcomas and is often observed in many types of cancer. This may be due to an imbalance in the cyclin D-CDK4/6-INK4-Rb pathway, leading increased abnormal cell proliferation (Lucchesi et al., 2018). The expression of CDK4 in tumor tissues is specific and can provide a sensitive marker for diagnosis of low-grade osteosarcoma (Dujardin et al., 2011). Targeted therapy has recently received increased attention. Inhibitors of CDK4/6 have been shown to have significant activity against several solid tumors, increase intracellular double-stranded RNA levels, and activate endogenous retroviral elements to inhibit tumor cell expression (Goel et al., 2017). Due to the importance of CDK4/6 activity in tumorigenesis, targeted inhibitors of the CDK4/6 gene have become new candidates for tumor therapy. The CDK4 inhibitor letrozole has been used to successfully treat breast cancer and has recently entered clinical trials for treatment of various diseases (Klein et al., 2018). CCT2 is a member of a chaperone protein containing the TCP1 complex (CCT), also known as the TCP1 loop complex (TRiC). It is a macromolecular complex of 16 subunits forming a back-to-back bicyclic structure, each ring containing eight different subunits α, β, γ, δ, ϵ, ζ, η, and θ (CCT1–8). This gene has been found to encode two different transcript variants. CCT has the function of assisting the correct folding of proteins, and cytoskeletal proteins and cell cycle regulators are the most important substrates. Blocking CCT activity can cause significant morphological changes and cell cycle arrest. Previous studies have found that CT2 is overexpressed in some tumors, and CCT2 expression in intestinal and hepatocarcinoma tissues is significantly higher than that in adjacent tissues, and its expression is highly correlated with PCNA, suggesting that CCT2 may be involved in cell proliferation. The positive expression of CCT2 in gallbladder carcinoma is associated with TNM stage and lymph node metastasis. In addition, the expression of CCT2 is associated with histological grade, suggesting that it is associated with tumor differentiation and progression. Recent studies have found that CCT2 is critical for the survival of breast cancer patients, and is significantly higher in hepatocellular carcinoma, colon cancer, extrahepatic cholangiocarcinoma, gallbladder cancer, and gastric cancer than benign lesions and normal tissues. However, how CCT2 affects HCC proliferation and progression remains to be explored. In conclusion, CCT2 is closely related to the development of HCC, which provides a theoretical basis for CCT2 to become a target for HCC molecular targeted therapy (Amit et al., 2010; Zou et al., 2013; Guest et al., 2015; Pavel et al., 2016; Minegishi et al., 2018). In our study, positive expression of CCT2 was negatively correlated with survival time, and was an independent risk factor for prognosis of LMS. The main function of monoacylglycerol acyltransferase (MGAT) is to catalyze the synthesis of diacylglycerol by monoacylglycerol. Currently, three genes encoding MGAT have been found, namely MGAT1, MGAT2, and MGAT3. MGAT is an important gene for the synthesis of diacylglycerol during fat deposition, and is closely

related to the absorption of fat in the intestine, the synthesis and storage of lipids, and intracellular signal transduction. An important function of MGAT1 is an important target of the Wnt/β-catenin signal. The protein has a characteristic type II transmembrane protein characteristic and plays an important role in the early development of animal embryos, organ formation, tissue regeneration, and other physiological processes, and is considered to be essential for normal embryogenesis. Stable overexpression of the MGAT1 gene in the Huh7 cell line resulted in a significant increase in tumor growth rate in severe combined immunodeficiency (SCID) mice. Down-regulation of MGAT1 expression in the liver can significantly reduce hepatic steatosis in mice, while reducing body weight and increasing glucose tolerance (Lee et al., 2012; Akiva and Birgul Iyison, 2018).

The results showed that the expression of CDK4, CCT2, and MGAT1 in LMS tissues was significantly higher than that in adjacent tissues and an important member of the cancer signaling pathway. Clinical data from the GEPIA dataset confirms that CDK4, CCT2, and MGAT1 expression levels are highly correlated with prognosis, and that up-regulation may lead to a significant reduction in survival time in patients with soft tissue sarcoma. At last, the result of cox analysis suggests that CDK4 and MGAT1 may play an important role in the development of LMS and can be used as predictors of LMS patients as a post-evaluation indicator. A recent large cohort study of 99 patients with LMS found that CDK4 may be a key gene for leiomyosarcoma recurrence, and palbociclib, an inhibitor of CDK4, may provide a new option for targeted therapy in patients with LMS (Bohm et al., 2019). However, LMS tumorigenesis is not well understood, and further evaluation of large sample clinical data is critical.

We studied co-expressed gene modules that were highly correlated with tumor recurrence, and determination of hub genes in these module helped to determine the major functions of the genes in these modules. The study of these three central genes may help us to understand the molecular mechanisms of tumorigenesis and these genes may represents new diagnostic marker and therapeutic target for LMS.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://tcga-data.nci.nih.gov/tcga/.

AUTHOR CONTRIBUTIONS

JY and CL reviewed relevant literature and drafted the manuscript. JZ, XL and SW conducted all statistical analyses. JY and SW were responsible for the supervision of the project and final approval of the version. All authors read and approved the final manuscript.

REFERENCES

Akiva, I., and Birgul Iyison, N. (2018). MGAT1 is a novel transcriptional target of Wnt/beta-catenin signaling pathway. *BMC cancer* 18 (1), 60. doi: 10.1186/s12885-017-3960-7

Amit, M., Weisberg, S. J., Nadler-Holly, M., McCormack, E. A., Feldmesser, E., Kaganovich, D., et al. (2010). Equivalent mutations in the eight subunits of the chaperonin CCT produce dramatically different cellular and gene expression phenotypes. *J. Mol. Biol.* 401 (3), 532–543. doi: 10.1016/j.jmb.2010.06.037

Barretina, J., Taylor, B. S., Banerji, S., Ramos, A. H., Lagos-Quintana, M., Decarolis, P. L., et al. (2010). Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat. Genet.* 42 (8), 715–721. doi: 10.1038/ng.619

Bindea, G., Galon, J., and Mlecnik, B. (2013). CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics* 29 (5), 661–663. doi: 10.1093/bioinformatics/btt019

Bohm, M. J., Marienfeld, R., Jager, D., Mellert, K., von Witzleben, A., Bruderlein, S., et al. (2019). Analysis of the CDK4/6 cell cycle pathway in Leiomyosarcomas as a potential target for inhibition by palbociclib. *Sarcoma* 2019, 3914232. doi: 10.1155/2019/3914232

Chibon, F., Lagarde, P., Salas, S., Perot, G., Brouste, V., Tirode, F., et al. (2010). Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* 16 (7), 781–787. doi: 10.1038/nm.2174

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8 Suppl 4, S11. doi: 10.1186/1752-0509-8-S4-S11

Croce, S., and Chibon, F. (2015). MED12 and uterine smooth muscle oncogenesis: State of the art and perspectives. *Eur. J. Cancer* 51 (12), 1603–1610. doi: 10.1016/j.ejca.2015.04.023

Detwiller, K. Y., Fernando, N. T., Segal, N. H., Ryeom, S. W., D'Amore, P. A., and Yoon, S. S. (2005). Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A. *Cancer Res.* 65 (13), 5881–5889. doi: 10.1158/0008-5472.CAN-04-4078

Dujardin, F., Binh, M. B., Bouvier, C., Gomez-Brouchet, A., Larousserie, F., Muret, A., et al. (2011). MDM2 and CDK4 immunohistochemistry is a valuable tool in the differential diagnosis of low-grade osteosarcomas and other primary fibro-osseous lesions of the bone. *Mod. Pathol. : an Off. J. United States Can. Acad. Pathol. Inc* 24 (5), 624–637. doi: 10.1038/modpathol.2010.229

Eriksson, M. (2010). Histology-driven chemotherapy of soft-tissue sarcoma. *Ann. Oncol.* 21 Suppl 7, vii270–vii276. doi: 10.1093/annonc/mdq285

George, S., Serrano, C., Hensley, M. L., and Ray-Coquard, I. (2018). Soft Tissue and Uterine Leiomyosarcoma. *J. Clin. Oncol. : Off. J. Am. Soc. Clin. Oncol.* 36 (2), 144–150. doi: 10.1200/JCO.2017.75.9845

Goel, S., DeCristo, M. J., Watt, A. C., BrinJones, H., Sceneay, J., Li, B. B., et al. (2017). CDK4/6 inhibition triggers anti-tumour immunity. *Nature* 548 (7668), 471–475. doi: 10.1038/nature23465

Gronchi, A., Ferrari, S., Quagliuolo, V., Broto, J. M., Pousa, A. L., Grignani, G., et al. (2017). Histotype-tailored neoadjuvant chemotherapy versus standard chemotherapy in patients with high-risk soft-tissue sarcomas (ISG-STS 1001): an international, open-label, randomised, controlled, phase 3, multicentre trial. *Lancet Oncol.* 18 (6), 812–822. doi: 10.1016/S1470-2045(17)30334-0

Guest, S. T., Kratche, Z. R., Bollig-Fischer, A., Haddad, R., and Ethier, S. P. (2015). Two members of the TRiC chaperonin complex, CCT2 and TCP1 are essential for survival of breast cancer cells and are linked to driving oncogenes. *Exp. Cell Res.* 332 (2), 223–235. doi: 10.1016/j.yexcr.2015.02.005

Guo, X., Jo, V. Y., Mills, A. M., Zhu, S. X., Lee, C. H., Espinosa, I., et al. (2015). Clinically Relevant Molecular Subtypes in Leiomyosarcoma. *Clin. Cancer Res.* 21 (15), 3501–3511. doi: 10.1158/1078-0432.CCR-14-3141

Hayashi, T., Horiuchi, A., Sano, K., Hiraoka, N., Kanai, Y., Shiozawa, T., et al. (2010). Mice-lacking LMP2, immuno-proteasome subunit, as an animal model of spontaneous uterine leiomyosarcoma. *Protein Cell* 1 (8), 711–717. doi: 10.1007/s13238-010-0095-x

Klein, M. E., Kovatcheva, M., Davis, L. E., Tap, W. D., and Koff, A. (2018). CDK4/6 inhibitors: the mechanism of action may not be as simple as once thought. *Cancer Cell* 34 (1), 9–20. doi: 10.1016/j.ccell.2018.03.023

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559

Lee, Y. J., Ko, E. H., Kim, J. E., Kim, E., Lee, H., Choi, H., et al. (2012). Nuclear receptor PPARgamma-regulated monoacylglycerol O-acyltransferase 1 (MGAT1) expression is responsible for the lipid accumulation in diet-induced hepatic steatosis. *Proc. Natl. Acad. Sci. United States America* 109 (34), 13656–13661. doi: 10.1073/pnas.1203218109

Lucchesi, C., Khalifa, E., Laizet, Y., Soubeyran, I., Mathoulin-Pelissier, S., Chomienne, C., et al. (2018). Targetable alterations in adult patients with soft-tissue sarcomas: insights for personalized therapy. *JAMA Oncol.* 4 (10), 1398–1404. doi: 10.1001/jamaoncol.2018.0723

Minegishi, Y., Nakaya, N., and Tomarev, S. I. (2018). Mutation in the Zebrafish cct2 gene leads to abnormalities of cell cycle and cell death in the retina: a model of CCT2-related Leber Congenital Amaurosis. *Investig. Ophthalmol. Vis. Sci.* 59, 995–1004. doi: 10.1167/iovs.17-22919

Mir, O., Brodowicz, T., Italiano, A., Wallet, J., Blay, J. Y., Bertucci, F., et al. (2016). Safety and efficacy of regorafenib in patients with advanced soft tissue sarcoma (REGOSARC): a randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Oncol.* 17 (12), 1732–1742. doi: 10.1016/S1470-2045(16)30507-1

Nagy, A., Lanczky, A., Menyhart, O., and Gyorffy, B. (2018). Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci. Rep.* 8 (1), 9227. doi: 10.1038/s41598-018-29514-3

Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Seki, K., et al. (2007). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Mod. Pathol.* 20 (7), 749–759. doi: 10.1038/modpathol.3800794

Noujaim, J., van der Graaf, W. T., and Jones, R. L. (2015). Redefining the standard of care in metastatic leiomyosarcoma. *Lancet Oncol.* 16 (4), 360–362. doi: 10.1016/S1470-2045(15)70107-5

O'Leary, B., Finn, R. S., and Turner, N. C. (2016). Treating cancer with selective CDK4/6 inhibitors. *Nat. Rev. Clin. Oncol.* 13 (7), 417–430. doi: 10.1038/nrclinonc.2016.26

Ognjanovic, S., Olivier, M., Bergemann, T. L., and Hainaut, P. (2012). Sarcomas in TP53 germline mutation carriers: a review of the IARC TP53 database. *Cancer* 118 (5), 1387–1396. doi: 10.1002/cncr.26390

Pautier, P., Floquet, A., Chevreau, C., Penel, N., Guillemet, C., Delcambre, C., et al. (2015). Trabectedin in combination with doxorubicin for first-line treatment of advanced uterine or soft-tissue leiomyosarcoma (LMS-02): a non-randomised, multicentre, phase 2 trial. *Lancet Oncol.* 16 (4), 457–464. doi: 10.1016/S1470-2045(15)70070-7

Pavel, M., Imarisio, S., Menzies, F. M., Jimenez-Sanchez, M., Siddiqi, F. H., Wu, X., et al. (2016). CCT complex restricts neuropathogenic protein aggregation via autophagy. *Nat. Commun.* 7, 13821. doi: 10.1038/ncomms13821

Quade, B. J., Wang, T. Y., Sornberger, K., Dal Cin, P., Mutter, G. L., and Morton, C. C. (2004). Molecular pathogenesis of uterine smooth muscle tumors from transcriptional profiling. *Genes Chromosomes Cancer* 40 (2), 97–108. doi: 10.1002/gcc.20018

Radulescu, E., Jaffe, A. E., Straub, R. E., Chen, Q., Shin, J. H., Hyde, T. M., et al. (2018). Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol. Psychiatry*. 26 (11), 1320–1331. doi: 10.1038/s41380-018-0304-1

Schoffski, P., Chawla, S., Maki, R. G., Italiano, A., Gelderblom, H., Choy, E., et al. (2016). Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: a randomised, open-label, multicentre, phase 3 trial. *Lancet* 387 (10028), 1629–1637. doi: 10.1016/S0140-6736(15)01283-0

Serrano, C., and George, S. (2013). Leiomyosarcoma. *Hematol. Oncol. Clin. North Am.* 27 (5), 957–974. doi: 10.1016/j.hoc.2013.07.002

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi: 10.1101/gr.1239303

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45 (W1), W98–w102. doi: 10.1093/nar/gkx247

Tawbi, H. A., Burgess, M., Bolejack, V., Van Tine, B. A., Schuetze, S. M., Hu, J., et al. (2017). Pembrolizumab in advanced soft-tissue sarcoma and bone sarcoma (SARC028): a multicentre, two-cohort, single-arm, open-label, phase 2 trial. *Lancet Oncol.* 18 (11), 1493–1501. doi: 10.1016/S1470-2045(17)30624-1

Wang, Q., Xie, L., Dang, Y., Sun, X., Xie, T., Guo, J., et al. (2019). OSlms: a web server to evaluate the prognostic value of genes in Leiomyosarcoma. *Front. Oncol.* 9, 190. doi: 10.3389/fonc.2019.00190

Zou, Q., Yang, Z. L., Yuan, Y., Li, J. H., Liang, L. F., Zeng, G. X., et al. (2013). Clinicopathological features and CCT2 and PDIA2 expression in gallbladder squamous/adenosquamous carcinoma and gallbladder adenocarcinoma. *World J. Surg. Oncol.* 11, 143. doi: 10.1186/1477-7819-11-143

Zuo, Z., Shen, J. X., Pan, Y., Pu, J., Li, Y. G., Shao, X. H., et al. (2018). Weighted Gene Correlation Network Analysis (WGCNA) detected loss of MAGI2 promotes Chronic Kidney Disease (CKD) by podocyte damage. *Cell Physiol. Biochem.* 51 (1), 244–261. doi: 10.1159/000495205

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

MINI REVIEW
published: 13 March 2020
doi: 10.3389/fonc.2020.00186

Humanizing Big Data: Recognizing the Human Aspect of Big Data

Kathy Helzlsouer[1], *Daoud Meerzaman*[2], *Stephen Taplin*[3] *and Barbara K. Dunn*[4]*

[1] Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, United States, [2] Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD, United States, [3] Center for Global Health, National Cancer Institute, Bethesda, MD, United States, [4] Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, United States

The term "big data" refers broadly to large volumes of data, often gathered from several sources, that are then analyzed, for example, for predictive analytics. Combining and mining genetic data from varied sources including clinical genetic testing, for example, electronic health records, what might be termed as "recreational" genetic testing such as ancestry testing, as well as research studies, provide one type of "big data." Challenges and cautions in analyzing big data include recognizing the lack of systematic collection of the source data, the variety of assay technologies used, the potential variation in classification and interpretation of genetic variants. While advanced technologies such as microarrays and, more recently, next-generation sequencing, that enable testing an individual's DNA for thousands of genes and variants simultaneously are briefly discussed, attention is focused more closely on challenges to analysis of the massive data generated by these genomic technologies. The main theme of this review is to evaluate challenges associated with big data in general and specifically to bring the sophisticated technology of genetic/genomic testing down to the individual level, keeping in mind the human aspect of the data source and considering where the impact of the data will be translated and applied. Considerations in this "humanizing" process include providing adequate counseling and consent for genetic testing in all settings, as well as understanding the strengths and limitations of assays and their interpretation.

Keywords: big data, predictive analytics, precision medicine, cancer risk prediction, clinical genetics/genomics, direct-to-consumer testing, data sharing

OPEN ACCESS

Edited by:
Michael Eccles,
University of Otago, New Zealand

Reviewed by:
Banu Arun,
University of Texas MD Anderson Cancer Center, United States
Pietro Pinoli,
Politecnico di Milano, Italy

***Correspondence:**
Barbara K. Dunn
dunnb@mail.nih.gov

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *21 June 2019*
Accepted: *04 February 2020*
Published: *13 March 2020*

Citation:
Helzlsouer K, Meerzaman D, Taplin S and Dunn BK (2020) Humanizing Big Data: Recognizing the Human Aspect of Big Data. Front. Oncol. 10:186. doi: 10.3389/fonc.2020.00186

INTRODUCTION

Precision medicine in cancer treatment is defined by the National Cancer Institute as a "genetic understanding" of cancer, offering a specific treatment tailored to an individual (1). Cancer results from a variety of factors, both genetic and environmental. The developmental path to the actual tumor results from an accumulation of genetic changes which vary across and within tumors. Some of these genetic changes are inherited germline mutations, but the majority are somatic changes, uncorrected by DNA repair processes, that result from exposures or random events. These genetic changes may present treatment targets; however, the genetic changes are heterogeneous and specific actionable treatment targets may be rare. To detect these changes, data on many tumors in many patients are required. Similarly, germline genetic changes that are inherited may increase susceptibility to cancer either directly by affecting key proteins such as those critical to repairing DNA damage or by increasing susceptibility to effects of cancer-causing environmental

factors. These germline changes may also be very rare; thus, analysis of large datasets is required to determine if there is an association with cancer development and to determine if the changes are useful in predicting risk.

The search for treatment targets and for predictive analytics has fueled the demand for large data sets, i.e., "big data." Despite the current widespread use of the term, no consistent or single definition of "big data" has been agreed on (2–5). The online Oxford Dictionaries definition is: "extremely large data sets that may be *analyzed computationally* to reveal patterns, trends, and associations, especially relating to human behavior and interactions" (6). In essence, "big data" denotes any data set large enough to permit valid use of *statistically based analytical methods* to extract a level of knowledge in an area of interest.

This massive data collection requires combining data from varied sources, collected in disparate manners and assayed using multiple techniques. The specific application of big data to be discussed in this paper is genomics and related omics as they feed into clinical management of patients.

These large data sets can be extremely complex, typically characterized by references to the "Vs" [high volume, velocity, variety, veracity, value, variability (4, 5, 7, 8)]. The growth in acquiring and using "big data" is due to a variety of factors including an increase in research and clinical applications of genetic findings, pharmaceutical company interest in large datasets to develop and apply targeted treatments, consumer interest in genetic tests for ancestry and medical applications, and a growth in the direct-to-consumer genetic test market. "Big Data" is now big business and growing. The market for genetic testing is projected to exceed $22 billion by 2024 (9). Companies now produce, buy, and sell genetic data. Buyers of data include researchers and pharmaceutical companies. Sellers include companies that provide genetic testing and/or companies that build and sell access to large data sets (*data aggregators*), as well as a new developing market for individuals, not just the companies, to benefit monetarily from the selling of their data to companies (10). Companies that market DNA data also may offer to perform testing. With this developing business around producing and sharing data, outside of the clinical setting, the danger exists of losing site not only of both how the data were collected and assayed, but also of the individual who is sharing the most intimate of data, their genetic profile.

The sharing and aggregation of genetic information into large data sets may obscure the fact that the basic underlying source of each data point is an individual. Individuals provide the data, the data from many are aggregated, and ultimately the information is translated back to an individual. Thus, analyzing and interpreting big data require recognizing the individual source of the data, how the data are obtained, stored, and assayed and analyzed, and how, ultimately, to apply them. In essence, the data must always be viewed and used with the humanity of the individuals providing their genetic material kept in mind.

The main theme of this review is to discuss challenges associated with big data in general and specifically to bring the sophisticated technology of genetic/genomic testing down to the individual level, where the impact of the data will be translated and applied. The latter activities reflect the "humanizing" of big data as applied to genomic medicine. This article will address analytical aspects of both genetics and genomics data and their evolution over time. Whereas, "genetics" involves the functioning and make-up of individual genes, the field addressed by big data sets containing genetic information is "genomics": genomics deals with *all* genes in an organism and their inter-relationships (11). The additional complexity in such big data has downstream implications for clinical interpretation and management for the individual. For this article, we will use the term "genetics" to include both genetics and genomics data, and we will address primarily germline genetics (i.e., also genomics), as elaborated below. Finally, as we review the sequential stages of genetic testing, we wish to re-emphasize the need to consider the relationship of each technical phase of the pipeline to the human being who is the source of the genetic material being analyzed.

CHALLENGES TO ANALYSIS OF GENOMIC AND MEDICAL DATA FOR DISCOVERY OF CLINICALLY RELEVANT GENETIC VARIANTS

Laboratory Testing of Germline DNA Variants

Testing of the germline for DNA variants, passed from one generation to the next, that confer deleterious phenotypic attributes has evolved radically over the years. This is largely in response to the evolution of technologies that enable massive testing of the genome (12), including microarrays but especially next-generation sequencing (13, 14). The huge data sets generated by these methods pose major challenges to the next stage in the pipeline: bioinformatic analysis and statistical validation. Such laboratory technologies allied with their follow-up bioinformatic analyses provide the venue through which "big data" are generated, and then funneled down into clinically interpretable genetic information, i.e., that which is directly relevant to the patient.

Challenges to analyzing genomic data for knowledge discovery begin in the laboratory at the technical level in the choice and conduct of specific approaches to sample preparation and laboratory analysis (15). The challenges continue downstream with the initial phases of the bioinformatic pipeline for identification of clinically relevant variants. These initial challenges involve selection of algorithms for optimal filtering of genetic variants and are followed down the pipeline through selection of appropriate algorithms at all subsequent informatic stages necessary to identify meaningful variants (15). Furthermore, the very large number of loci interrogated in such discovery research represent individual tests for clinically relevant genetic variants, posing the statistical challenge inherent in multiple testing and concerns about identifying false positives. The quality of the data generated at the end of this genomic pipeline, i.e., the data on which clinical associations will be based, must be carefully monitored throughout. Bias and variable thresholds for calling individual genetic variants as clinically relevant can feed into erroneous conclusions drawn from data. Scrutiny of the findings at each stage of the pipeline is essential

to maximize the chance of identifying true positive variants and avoid missing false negatives. Furthermore, impediments to generation of accurate, meaningful data are not limited to technical decisions but are subject as well to inconsistent communications among researchers with differing expertise at each stage of the genomic pipeline (15). Cautionary approaches are therefore necessary if the users of the genomic findings in the healthcare setting can trust the quality of the underlying data.

Addressing the Limitations of Genomic Technologies: Analytic Validity and Probabilistic Outcomes

The laboratory technologies allied with their follow-up bioinformatic analyses provide the venue through which "big data" are generated, and then funneled down into clinically interpretable genetic information, where "humanization" of the "big data" needs to be emphasized. This stage is where the "variety" attribute of generated data must be sifted through to glean out irrelevant findings and select for meaningful outcomes that are potentially pertinent to clinical interpretation. Key to humanizing the data is communicating to the patient the limitations at the clinical level of the transmitted information, both technical and genetic.

The platforms most commonly used to identify pathogenic variants in the clinical setting are single nucleotide polymorphism (SNP) chip (microarray)-based and next generation sequencing (next gen sequencing)-based technologies. Although they are used in standard clinical practice, caution must be exercised in interpreting the results of these analytic tools. They are not perfect, and the limitations of the diagnostic accuracy, or analytic validity (16), of a given platform must be considered when communicating results to a patient. This is particularly true of SNP chips. When juxtaposed against results obtained from next gen sequencing, the diagnostic accuracy of SNP chips has been shown to be uncertain when used to detect rare pathogenic variants in the general population (17). The analytic validity of such rare variants is poor, leading to a very high false discovery rate. Thus, although SNP chips are useful for assessing the presence of common variants in a given population, such as polymorphisms, this does not translate into the rare variants relevant to clinical genetic diagnoses. Similar limitations exist for SNP chips from different manufacturers. This contrasts with sequencing platforms which are not affected by the same technical issues as chips and are therefore more accurate in genotyping rare variants (17).

Even in a setting of strong analytic validity, as seen with sequencing, many uncertainties remain. An accurately identified variant may have questionable clinical validity, the strength of its association with the phenotypic outcome of interest (16, 18), i.e., disease, being uncertain. These unknowns are inherent in the probabilistic nature of phenotypic expression of genetic variants. Patients may assume that identification of a pathogenic variant equates to certain development of the associated disease, whereas incomplete penetrance is generally the rule in heritable diseases such as adult cancers. Nevertheless, the actual penetrance of rare alleles is uncertain and can be over-estimated by clinical ascertainment methods (19). Even greater uncertainty exists for variants with unknown pathogenicity, namely "variants of uncertain significance," or VUSs. Without humanizing such findings by communicating the absence of documented clinical relevance to the patient, unnecessary anxiety may be provoked and avoidable invasive treatment interventions undertaken. Finally, documentation of analytic and clinical validity is not sufficient to make a genetic test truly useful to the patient. The test must have clinical utility in that it lays the groundwork for beneficial interventions, whether pharmaceutical, surgical, or behavioral, without overriding risks (16). By establishing that a genetic test can lead to a clinically actionable intervention, the role played by big data in performance of the test becomes humanized.

CHALLENGES TO MANAGEMENT OF BIG DATA: GENOMIC AND CLINICAL DATA

Ethical Challenges

Ethical issues evolving from the amassing of genetic data should be addressed by researchers, health care providers and companies. Subsequent use of "big data" must consider the selective nature of the source of the data, i.e., the patient, and the generalizability as well as the absolute necessity to prevent data breaches and ensure data security (8). Informed consent is an essential part of this process. The sharing of information from big data accumulated from thousands of individuals, has long raised concerns about maintaining individual privacy while advancing our understanding of genetic associations that will promote public health (8, 20, 21). The potential disregard of maintaining genetic privacy has led to anxiety about sequelae involving discrimination in multiple aspects of life, including employment and health insurance (20). While the Genetic Information Nondiscrimination Act (GINA) was enacted to prohibit such discriminatory behavior, additional domains (e.g., life, disability, and long-term care insurance) have remained vulnerable to misuse of genetic information (20). The ethical issues arising from the need to optimize these two "goods"– health vs. privacy—while balancing the risks and benefits emerging from this process (22) constitute an essential part of humanizing the big data.

Security Challenges

Although security challenges overlap those inherent in the ethical concerns just described, a number of issues relating to security merit independent mention. Data needs to be accessible and at the same time secure. Security must guarantee privacy of data relating to the individual. An actual set of criteria, FISMA (Federal Information Security Management Act), provides a framework to guide protections of any information involving government activities. The private sector parallel is HIPAA (Health Insurance Portability and Accountability Act), which is widely adhered to in healthcare settings. These security concerns are becoming increasingly challenging due to the explosion of big data and their storage on multiple cloud resources (23).

Challenges to Management of Data Size and Data Storage (the Silo Problem)

The huge size of big data, exacerbated by its continuous growth in volume, poses challenges to storage (5, 24). Traditionally data have been generated and stored in isolated compartments that may even differ qualitatively from each other. As an example, different departments in the same organization may store data in their own data bases, resulting in "data silos." The content of siloed data in different departments may overlap but be encoded using differing terminology such that these data cannot "speak to each other." This creates a serious impediment to integrated analyses of healthcare-related data across siloes; such analyses are critical to understanding factors affecting health-directed outcomes, including genetics. Among critical siloed data sets are Electronic Health Records (EHRs) (23), valuable for generating trends and predictive models, including genomic and pharmacogenomic markers (5, 25). The huge size of certain types of data, i.e., genomic data, which must be integrated with other data types of smaller size but much greater complexity, i.e., phenotypic data as contained in the EHR, poses additional challenges, which will be discussed below.

Challenges to Management of Data in Unstructured Formats (26)

Frequently superimposed on the sheer size and ongoing growth of the data is the extreme architectural complexity of the data. The complexity of certain types of data (e.g., genomic) poses daunting challenges to being moved from home storage to an analytic environment. Unstructured data does not conform to a consistent accessible framework and language. Therefore, it needs to be converted into a structured readable format in order to identify useful information. In the clinical genomic setting, this conversion to a structured format is essential to teasing out genetic variants that are clinically meaningful and actionable. Historically, medical charting was entirely unstructured, comprising handwritten notes interspersed with machine-generated data, such as laboratory values. The EHR represents a first step at structuring such patient data by providing a consistent template for entries of medical information (23). However, data derived from the EHR are of multiple types (27). One estimate has 80% of data contained in EHRs as unstructured (26, 28). These varied entries in the EHR have value in that they can be used to formulate phenotypic classifications of patients. The technical challenges to this conversion process involve sophisticated algorithms using machine learning, natural language processing (NLP), and artificial intelligence (AI) (26). In the clinical genetic setting, examples of unstructured data that are difficult to convert to structured formats include EHRs, genomics, and other omic datasets. Commonly, for example, integration of the EHR with genomic and other types (e.g., biospecimen) of clinically relevant data results in questionable phenotypic diagnoses due to inaccurately determined correlations (29, 30). In essence, challenges to data quality, reliability, accuracy and integration must always be addressed. The ultimate goal is to discover associations between genetic/genomic variations and clinical phenotypes that are accurate and clinically meaningful in that they can be used to manage patient care, essentially creating predictive models (26).

Challenges to Data Sharing

Essential to gleaning meaningful, actionable information from large data sets, in any context, is sharing of data among data producers (31). Given the need for as much data as possible to deduce clinically meaningful genomic variants, sharing of data among source clinical sites is critical, especially for rare genetic diseases (32). A guideline known as FAIR (Findable, Accessible, Interoperable, and Reusable) has been developed to guide investigators in managing the sharing of big data (33). To optimize the quality and usefulness of shared data sets, regulatory policies governing all genomic-related data generated by NIH-funded research have been established. Such Genomic Data Sharing (GDS) policies are specific to given types of data (34).

Challenges to Testing of the Individual, i.e., the Data Source

Sources for large analytic data sets, i.e., "big data," include data from clinical settings as well as genetic testing companies. Thus, potential selection factors for who gets testing will affect the results and interpretation. Until recently, the ordering of cancer genetic tests for cancer susceptibility syndromes for those diagnosed with cancer or with a strong family history of cancer was done in the clinical setting, after genetic counseling by a qualified health care provider. More recently, cancer genetic testing, as well as other health-related genetic testing, has expanded beyond the clinical setting, with companies advertising and offering testing directly to consumers without the need for involving a health care provider, or offering the test with a company-provided physician to order the test. The benefit of direct-to-consumer testing is potentially improved accessibility through convenience of in-home testing, bypassing requirements for health care provider visits, and lower cost tests. Data sets with a preponderance of clinically sourced data are likely to have higher risk individuals than direct-to-consumer or consumer-driven genetic testing. Also, in contrast to direct-to-consumer generated data, clinical settings are more likely to have extensive family history information, which is critical for interpreting test results. However, the extensive family history documentation may or may not be adequately or accurately transmitted to the "big data" compilation.

The individual who is the source of the data, the researcher analyzing it, and the clinicians who use the results of analyses should have a broad understanding of the process of consent, genetic testing, its benefits, harms and limitations, the potential implications of data sharing and with whom genetic testing results are shared. Immersed in the massive amounts of information and issues surrounding the use of genetic/genomic data at the clinical level, the input source of these data—the patient/individual—and the process of generating the data may be overlooked.

Pre-test counseling prior to proceeding with genetic testing is recommended because of the complexity of genetic information, and the need to anticipate how that information will be used for

subsequent management of risk. Counseling includes several key components: medical and family history, risk assessment, risk perception, discussion of the most appropriate test, benefits and limitations of testing, communication with family members, and follow-up management (35, 36). This patient-centered approach espouses shared decision making, a process by which the patient has an informed discussion with the health care provider about the above issues, taking into consideration their personal values and whether or not to pursue genetic testing. Pre-test genetic counseling informs the individual and facilitates shared decision making while ensuring patient autonomy in the process (37) and is recommended by the U. S. Preventive Services Task Force (USPSTF) (38) and the National Comprehensive Cancer Network (NCCN) (39) in appropriate situations. Unlike other medical tests, genetic testing has implications for the family members, leading to issues such as how to communicate test results to family members as well as how the data may be shared. These downstream components of the genetic pipeline illustrate the strong human element with which the process culminates. Those using big data should ensure that the individual's preferences are respected and that they are informed of the potential broad sharing of data. Similarly, when applying information gathered from analyses of "Big Data," the uncertainty that may be introduced by the methodologic issues in data generating activities as noted previously should be considered. Progress in technical and computational methodologies has simplified the generation of massive genomic analyses but limitations still exist.

SUMMARY

The application of technologies to generate and interpret big data related to genetic testing holds promise for the future of cancer medicine. The practice of "precision medicine," in which the diagnostic and therapeutic interactions are tailored to a given patient, should benefit considerably from modern genomic technologies. Unquestionably genetic understanding is a key component of this approach to patient care, given the foundational role played by cumulative somatic mutations in carcinogenesis (40). Precision medicine must be built on precision data. The sources of the data used in "big data" should be stated along with the characterization of the population source, specimen source and preparation, assays used and analytic methods and algorithms employed. At the application and interpretation of data, the "precision" of precision medicine derives as much from an understanding of the psychological and social setting and needs of the patient and from the standard clinical attributes that brought the individual to the medical system as from the genetic underpinnings of the cancer or cancer risk. The composite of all these attributes makes the focus on a given patient truly precise, humanizing the process of incorporating genetic content into the practice of cancer medicine.

The potential of technology to improve the public health is unquestionable. However, understanding how technical platforms that analyze large-scale data feed into clinically relevant information can be daunting for patients and healthcare providers without specific genomic training. In this paper we have drawn attention to the many challenges and limitations as well as benefits associated with analyzing and applying big data to clinical applications. Our goal has been to point the way to demystifying the complexity of "big data" so that recipients of its benefits, patients and providers, will be in a better position to make appropriate clinical decisions. In this sense, we have attempted to "humanize big data," by unraveling its many components in an effort to make its meaning, if not all its details, more accessible to non-specialists.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. KH, BD, and DM contributed most of the writing.

REFERENCES

1. *Precision Medicine in Cancer Treatment 2019*. Available online at: https://www.cancer.gov/about-cancer/treatment/types/precision-medicine (accessed October 3, 2017).
2. Su P. Direct-to-consumer genetic testing: a comprehensive view. *Yale J Biol Med*. (2013) 86:359–65.
3. Wikipedia. *Big Data 2019*. Available online at: https://en.wikipedia.org/wiki/Big_data; https://webcache.googleusercontent.com/search?q=cache:ISElu6k9ARwJe; https://en.wikipedia.org/wiki/Big__data$+$&cd=1&hl=en&ct=clnk&gl=us (accessed February 25, 2019).
4. Na KS, Han C, Kim YK. Big data and discovery sciences in psychiatry. *Adv Exp Med Biol*. (2019) 1192:3–15. doi: 10.1007/978-981-32-9721-0_1
5. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. *IEEE J Biomed Health Inform*. (2015) 19:1193–208. doi: 10.1109/JBHI.2015.2450362
6. Dictionaries O. *Oxford Dictionaries*. Oxford: Lexico.com (2019). Available online at: https://www.lexico.com/en/definition/big_data
7. Fuller D, Buote R, Stanley K. A glossary for big data in population and public health: discussion and commentary on terminology and research methods. *J Epidemiol Community Health*. (2017) 71:1113–7. doi: 10.1136/jech-2017-209608
8. Herschel R, Miori VM. Ethics & big data. *Technol Soc*. (2017) 49:31–6. doi: 10.1016/j.techsoc.2017.03.003
9. *Genetic Testing Market Surge to Cross $22 Billion By 2024: MarketWatch*. (2019). Available online at: https://www.marketwatch.com/press-release/genetic-testing-market-will-register-116-growth-to-cross-usd-225-billion-by-2024--2019-02--26 (accessed February 22, 2019).
10. Zhang S. *Big Pharma Would Like Your dna 23 and meś $300 Million Deal With Glaxosmithkline is Just the Tip of the Iceberg*. The Atlantic (2018).
11. Glossary of genomics terms. *JAMA*. (2013) 309:1533–5. doi: 10.1001/jama.2013.2950
12. Meerzaman D, Dunn BK, Lee M, Chen Q, Yan C, Ross S. The promise of omics-based approaches to cancer prevention. *Semin Oncol*. (2016) 43:36–48. doi: 10.1053/j.seminoncol.2015.09.004
13. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. (2010) 11:31–46. doi: 10.1038/nrg2626
14. Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders. *N Engl J Med*. (2018) 379:1353–62. doi: 10.1056/NEJMra1711801

15. Meerzaman D, Dunn BK. Value of collaboration among multi-domain experts in analysis of high-throughput genomics data. *Cancer Res*. (2019) 79:5140–5. doi: 10.1158/0008-5472.CAN-19-0769
16. Burke W. Genetic tests: clinical validity and clinical utility. *Curr Protoc Hum Genet*. (2014) 81:9.15.1–8. doi: 10.1002/0471142905.hg0915s81
17. Weedon M, Jackson L, Harrison J, Ruth K, Hattersley A, Wright C. *Very Rare Pathogenic Genetic Variants Detected by SNP-Chips Are Usually False Positives: Implications for Direct-to-Consumer Genetic Testing Online*. Cold Spring Harbor Laboratory (2019).
18. Zion TN, Wayburn B, Darabi S, Lamb Thrush D, Smith ED, Johnston T, et al. Clinical validity assessment of genes for inclusion in multi-gene panel testing: a systematic approach. *Mol Genet Genomic Med*. (2019) 7:e630. doi: 10.1002/mgg3.630
19. Wright CF, West B, Tuke M, Jones SE, Patel K, Laver TW, et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am J Hum Genet*. (2019) 104:275–86. doi: 10.1101/407981
20. Green RC, Lautenbach D, McGuire AL. GINA, genetic discrimination, and genomic medicine. *N Engl J Med*. (2015) 372:397–9. doi: 10.1056/NEJMp1404776
21. Salerno J, Knoppers BM, Lee LM, Hlaing WM, Goodman KW. Ethics, big data and computing in epidemiology and public health. *Ann Epidemiol*. (2017) 27:297–301. doi: 10.1016/j.annepidem.2017.05.002
22. Knoppers BM, Thorogood A. Ethics and big data in Health. *Curr Opin Syst Biol*. (2017) 4:53–7. doi: 10.1016/j.coisb.2017.07.001
23. Rodrigues JJ, de la Torre I, Fernandez G, Lopez-Coronado M. Analysis of the security and privacy requirements of cloud-based electronic health records systems. *J Med Internet Res*. (2013) 15:e186. doi: 10.2196/jmir.2494
24. Harvey C. *Big Data Management Datamation Daily Newsletter*. (2017). Available online at: https://www.datamation.com/big-data/big-data-management.html (accessed June 20, 2017).
25. Barrot CC, Woillard JB, Picard N. Big data in pharmacogenomics: current applications, perspectives and pitfalls. *Pharmacogenomics*. (2019) 20:609–20. doi: 10.2217/pgs-2018-0184
26. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. the revival of the notes field: leveraging the unstructured content in electronic health records. *Front Med*. (2019) 6:66. doi: 10.3389/fmed.2019.00066
27. He KY, Ge D, He MM. Big data analytics for genomic medicine. *Int J Mol Sci*. (2017) 18:E412. doi: 10.3390/ijms18020412
28. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. (2013) 309:1351–2. doi: 10.1001/jama.2013.393
29. Hughey JJ, Rhoades SD, Fu DY, Bastarache L, Denny JC, Chen Q. Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genomics*. (2019) 20:805. doi: 10.1186/s12864-019-6192-1
30. Lee J, Hamideh D, Nebeker C. Qualifying and quantifying the precision medicine rhetoric. *BMC Genomics*. (2019) 20:868. doi: 10.1186/s12864-019-6242-8
31. Stuart D, Allin K, Penny D, Lucraft M, Astell M. *Practical Challenges for Researchers in Data Sharing*. Springer Nature; Springer; NatureResearch; BMC; Palgrave MacMillan (2018). doi: 10.6084/m9.figshare.5971387. (accessed March 21, 2018).
32. Raza S, Hall A. Genomic medicine and data sharing. *Br Med Bull*. (2017) 123:35–45. doi: 10.1093/bmb/ldx024
33. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. (2016) 3:160018. doi: 10.1038/sdata.2016.18
34. Shabani M, Dove ES, Murtagh M, Knoppers BM, Borry P. Oversight of genomic data sharing: what roles for ethics and data access committees? *Biopreserv Biobank*. (2017) 15:469–74. doi: 10.1089/bio.2017.0045
35. Trepanier A, Ahrens M, McKinnon W, Peters J, Stopfer J, Grumet SC, et al. Genetic cancer risk assessment and counseling: recommendations of the national society of genetic counselors. *J Genet Couns*. (2004) 13:83–114. doi: 10.1023/B:JOGC.0000018821.48330.77
36. Trepanier A, Ahrens M, McKinnon W, Peters J, Stopfer J, Grumet SC, et al. Cancer *Genetics Risk Assessment and Counseling (PDQ®)–Health Professional Version*. National Cancer Institute PDQ. Available online at: https://www.cancer.gov/about-cancer/causes-prevention/genetics/risk-assessment-pdq#_1004 (accessed March 1, 2019).
37. The SHARE Approach. *The SHARE Approach*. (2014). Available online at: https://www.ahrq.gov/health-literacy/curriculum-tools/shareddecisionmaking/index.html (accessed August 2018).
38. U.S. PSTF. *BRCA-Related Cancer: Risk Assessment, Genetic Counseling, and Genetic Testing USPSTF Website*. (2019). Available online at: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/brca-relatedcancer-risk-assessment-genetic-counseling-andgenetic-testing1?ds=1&s=BRCA-related%20cancer (accessed February 19, 2020).
39. National Comprehensive Cancer Network IU. *Genetic/Familial High-Risk Assessment*. Available online at: https://www.nccn.org/professionals/physician_gls/pdf/genetics_screening.pdf
40. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. *Lancet*. (2010) 375:1525–35. doi: 10.1016/S0140-6736(10)60452-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership