

Biomarkers for Diagnosis, Therapy and Prognosis in Colorectal Cancer: a study from databases, machine learning predictions to laboratory confirmations

To my parents
献给我的父母

Örebro Studies in Medicine 214



XUELI ZHANG

**Biomarkers for Diagnosis, Therapy and Prognosis in
Colorectal Cancer: a study from databases, machine
learning predictions to laboratory confirmations**

© Xueli Zhang, 2020

Title: Biomarkers for Diagnosis, Therapy and Prognosis in Colorectal Cancer:
a study from databases, machine learning predictions to laboratory confirmations

Publisher: Örebro University 2020
www.oru.se/publikationer-avhandlingar

Print: Örebro University, Repro 05/2020

ISSN 1652-4063
ISBN 978-91-7529-341-7

Abstract

Xueli Zhang (2020): Biomarkers for Diagnosis, Therapy and Prognosis in Colorectal Cancer: a study from databases, machine learning predictions to laboratory confirmations. Örebro Studies in Medicine 214.

Colorectal cancer (CRC) is one of the leading causes of cancer death worldwide. Early diagnosis and better therapy response have been believed to be associated with better prognosis. CRC biomarkers are considered as precise indicators for the early diagnosis and better therapy response. It is, therefore, of importance to find out, analyze and evaluate the CRC biomarkers to further provide the more precise evidence for predicting novel potential biomarkers and eventually to improve early diagnosis, personalized therapy and prognosis for CRC.

In this study, we started with creating and establishing a CRC biomarker database. (CBD: <http://sys-bio.suda.edu.cn/CBD/index.html>) In the CBD database, there were 870 reported CRC biomarkers collected from the published articles in PubMed. In this version of the CBD, CRC biomarker data was carefully collected, sorted, displayed, and analyzed. The major applications of the CBD are to provide 1) the records of CRC biomarkers (DNA, RNA, protein and others) concerning diagnosis, treatment and prognosis; 2) the basic and clinical research information concerning the CRC biomarkers; 3) the primary results for bioinformatics and biostatistics analysis of the CRC biomarkers; 4) downloading/uploading the biomedicine information for CRC biomarkers.

Based on our CBD and other public databases, we further analyzed the presented CRC biomarkers (DNAs, RNAs, proteins) and predicted novel potential multiple biomarkers (the combination of single biomarkers) with biological networks and pathways analysis for diagnosis, therapy response and prognosis in CRC. We found several hub biomarkers and key pathways for the diagnosis, treatment and prognosis in CRC. Receiver operating characteristic (ROC) test and survival analysis by microarray data revealed that multiple biomarkers could be better biomarkers than the single biomarkers for the diagnosis and prognosis of CRC.

There are 62 diagnosis biomarkers for colon cancer in our CBD. In the previous studies, we found these present biomarkers were not enough to improve significantly the diagnosis of colon cancer. In order to find out novel biomarkers for the colon cancer diagnosis, we have performed /machine learning (ML) techniques such as support vector machine (SVM) and regression tree to predict candidate to discover diagnostic biomarkers for colon cancer. Based on the protein-protein interaction (PPI) network topology features of the identified biomarkers, we found 12 protein biomarkers which were considered as the candidate colon cancer diagnosis biomarkers. Among these protein biomarkers Chromogranin-A (CHGA) was the most powerful biomarker, which showed good performance in bioinformatics test and Immunohistochemistry (IHC). We are now expanding this study to CRC.

Expression of CHGA protein in colon cancer was further verified with a novel logistic regression based meta-analysis, and convinced as a valuable diagnostic biomarker as compared with the typical diagnostic biomarkers, such as TP53, KRAS and MKI67.

microRNAs (miRNAs/miRs) have been considered as potential biomarkers. A novel miRNA-mRNA interaction network-based model was used to predict miRNA biomarkers for CRC and found that miRNA-186-5p, miRNA-10b-5p and miRNA-30e-5p might be the novel biomarkers for CRC diagnosis. In conclusion, we have created a useful CBD database for CRC biomarkers and provided detailed information for how to use the CBD in CRC biomarker investigations. Our studies have been focusing on the biomarkers in diagnosis, therapy and prognosis. Based on our CBD and other powerful cancer associated databases, ML has been used to analyze the characteristics of the CRC biomarkers and predict novel potential CRC biomarkers. The predicted potential biomarkers were further confirmed at biomedical laboratory.

Keywords: biomarkers, diagnosis, therapy response, prognosis, database, machine learning, CRC

Xueli Zhang, School of Medical Sciences

Örebro University, SE-701 82 Örebro, Sweden, zhang.xueli@oru.se

Table of Contents

LIST OF PUBLICATIONS	9
OTHER PAPERS NOT IN THIS THESIS.....	10
LIST OF ABBREVIATIONS	11
1 INTRODUCTION.....	13
1.1 Colorectal cancer	14
1.1.1 Colorectal cancer diagnosis	15
1.1.2 Colorectal cancer treatment	16
1.1.3 Colorectal cancer prognosis	18
1.2 Biomarkers	19
1.2.1 Biomarkers in colorectal cancer	19
1.2.2 Biomarker detection	29
1.3 Bioinformatics approach	30
1.3.1 Biomedicine databases.....	30
1.3.2 Complex network	36
1.3.3 Machine learning	38
1.3.4 Novel meta-analysis	40
2 THE PRESENT INVESTIGATION	42
2.1 Paper I.....	42
2.1.1 Background and aims.....	42
2.1.2 Materials and methods.....	42
2.1.3 Results and discussions	42
2.2 Paper II.....	43
2.2.1 Background and aims.....	43
2.2.2 Materials and methods.....	43
2.2.3 Results and discussions	43
2.3 Paper III	44
2.3.1 Background and aims.....	44
2.3.2 Materials and methods.....	44
2.3.3 Results and discussions	44
2.4 Paper IV	45
2.4.1 Background and aims.....	45
2.4.2 Materials and methods.....	45
2.4.3 Results and discussions	45

2.5 Paper V	46
2.5.1 Background and aims	46
2.5.2 Materials and methods	46
2.5.3 Results and discussions	46
ACKNOWLEDGEMENTS	47
REFERENCES	48

List of publications

- I. **Zhang X**, Sun X-F, Cao Y, Ye B, Peng Q, Liu X, Shen B and Zhang H
CBD: a biomarker database for colorectal cancer.
Database 10.1093/database/bay046, 2018
- II. **Zhang X**, Sun X-F, Shen B and Zhang H
Potential applications of DNA, RNA and protein biomarkers in diagnosis, therapy and prognosis for colorectal cancer: a study from databases to AI-assisted verification.
Cancers 11:172, 2019
- III. **Zhang X**, Zhang H, Fan C-W, Shen B and Sun X-F
Loss of CHGA expression as a potential biomarker for colon cancer diagnosis: a study on biomarker discovery by machine learning and confirmation in colorectal cancer tissue microarrays.
Submitted, 2020
- IV. **Zhang X**, Zhang H, Shen B and Sun X-F
Chromogranin-A expression as a novel biomarker for early diagnosis of colon cancer patients.
Int J Mol Sci 20: 2919, 2019
- V. **Zhang X**, Zhang H, Shen B and Sun X-F
Novel microRNA biomarkers for colorectal cancer early diagnosis and 5-fluorouracil chemotherapy resistance but not prognosis: a study from databases to AI-assisted verifications.
Cancers 12:341, 2020

Other papers not in this thesis

Peng Q*, Zhang X*, Min M, Zou L, Shen P and Zhu Y

The clinical role of microRNA-21 as a promising biomarker in the diagnosis and prognosis of colorectal cancer: a systematic review and meta-analysis.

Oncotarget 8:44893-909, 2017

Liu X*, Zhang X*, Ye B, Lin Y, Sun X-F, Zhang H and Shen B

CRC-EBD: Epigenetic Biomarker Database for Colorectal Cancer.

Submitted, 2020

Fan C-W, Fang C, Zhang X, Lu Z-Y, Li Y, Zhang H, Wang C, Zhou Z-G and Sun X-F.

The identification and clinical significance of metastasis-related network based on a feature of the universal screening of deficient mismatch repair protein in colorectal cancer patients.

Submitted, 2020

Meng W-J, Pathak S, Adell G, Holmlund B, Wang Z-Q, Zhang X, Zhang H, Zhou Z-G and Sun X-F

Expression of miR-302a, miR-105 and miR-888 plays several critical roles in pathogenesis, radiotherapy response and prognosis in rectal cancer patients: a study from real-time PCR to big data analyses.

Submitted, 2020.

*Authors contributed equally to the work.

List of abbreviations

5-FU	fluorouracil
ABI2	Abl interactor 2
AI	artificial intelligence
AUC	area under the ROC curve
BBI	biomarker-biomarker interaction
CBD	colorectal cancer biomarker database
CEA	carcinoembryonic antigen
CHGA	chromogranin-A
CRC	colorectal cancer
CT	computed tomography
DBMS	database manger system
DEG	differentially expressed gene
DL	deep learning
DT	decision tree
EBM	evidence-based medicine
EGFR	epidermal growth factor receptor
ELISA	enzyme-linked immunosorbent assay
FIT	fecal immunochemical test
FOBT	fecal occult blood test
FP	false positivity
GE	gene expression
HR	hazard ratio
lncRNA	long noncoding RNA
miR/miRNA	microRNA
ML	machine learning
MMR	mismatch repair
MSI	microsatellite instability
MRI	magnetic resonance imaging
mRNA	massager RNA
ncRNA	non-coding RNA
NPV	negative predictive value
OR	odds ratio
PPI	protein-protein interaction

PPV	positive predictive value
qRT-PCR	quantitative reverse transcription polymerase chain reaction
ROC	receiver operating characteristic curve
RR	risk ratio
RRI	RNA-RNA interaction
SAGE	serial analysis of gene expression
SEER	surveillance, epidemiology, and end results
SVM	support vector machine
TCGA	the cancer genome atlas
TN	true negative
TNR	true negative rate
TP	true positivity
TPR	true positive rate

1 Introduction

Cancer is a “complex multigenic disease characterized by various types of epigenetic and genetic variations”¹⁻³, which is one of the leading causes of death worldwide⁴. Colorectal cancer (CRC) are the cancers derived from the colon or rectum, which is the third leading cancer and second cause of cancer death⁵. There are 1800977 new CRC cases and 861661 deaths of CRC in 2018, which occupies approximately 10% of all new diagnosed cancer patients and cancer caused deaths⁵. Taking advantage of the development of modern medicine, the mortality of CRC has been decreasing^{6,7}. However, the incidence of CRC is increasing^{6,7}. Colonoscope has been considered as the golden test for CRC diagnosis⁷. However, high-cost and body-intrusive are the significant disadvantage of colonoscope. Therefore, the development of diagnostic methods is still needed⁷. Surgery is the first choice in most cases of CRC treatment⁷. However, the high pain/cost of surgery is always concerned. The TNM stage system has been a typical way to guide the prognosis of CRC. However, it is still a challenge to divide CRC patients into the right stages accurately and quickly⁶.

Precision medicine (personalized medicine) is a new-developed medicine theory that uses the biomedicine information of a specific person to prevent, predict (before and after), diagnose and treat diseases⁸. The precision medicine in cancer is more focusing on the tumor information for a specific person, which can be used to make diagnosis, treatment and prognosis more accurately⁸. The development of precision medicine has drawn more attention and requirement for the discovery and research of biomarkers.

Biomarker is the specific biological indicator in the human body that can serve as a marker or indicator for diseases⁹. With the rapid development of modern science theory and technology, more and more biomarkers have been discovered and identified. The process of biomarker discovery and verification has developed into a system procedure¹⁰. As the progress of modern biomedicine and the advent of the era of big data, accurate pre-prediction and computational simulation verification require the participation and development of bioinformatics increasingly¹⁰.

Bioinformatics is a hybrid research field that use multiple dry-lab approach such as computer sciences, statistics and mathematics to solve biomedicine problems¹¹. With the development of precision medicine and other related subjects and technologies, bioinformatics is playing more and more critical role in the process of biomarker research. A majority of the bioinformatic researches have been focused on finding biosignature as

biomarkers, by gene expression (GE) data, which are considered with high heterogeneity among different studies. However, the biomarkers predicted by these classic models could not get enough evidences to be common biomarkers for most of patients.

Complex network theory has been an important component in bioinformatics study ^{10,12}. As the advent of the era of big data, huge amounts of biological network data have been generated and collected on related databases like String ¹³ and miRnet ¹⁴. Many studies support that many biomarkers have similar topology features on biological networks ^{15,16}. As such, it is possible to predict new biomarkers based on the topology features on networks.

Machine learning (ML) has been a popular method to predict biomarkers in bioinformatics, which is getting higher accurate since the development of computer science and the increasing amount of training data. Therefore, the application of ML on complex network to predict biomarkers is reasonable and worth waiting.

This thesis is focusing on 1) the background of CRC; 2) overview of the biomarkers in CRC diagnosis, treatment and prognosis; 3) the bioinformatics approach for CRC biomarker research; 4) the valuable biomarkers in CRC; 5) the further directions of biomarkers in cancer research.

1.1 Colorectal cancer

CRC can be divided into colon cancer and rectal cancer, based on the location of occurrence. In 2018, 1096601 patients are diagnosed with colon cancer and 704376 with rectal cancer in the world ⁵. There are 551269 and 310394 died because of colon and rectal cancer, respectively ⁵.

CRC is a globe disease. Asia has the largest number of CRC patients among all the continents, which occupies half of CRC occurrence and death, since it has around 60% population of the world ⁵. The next is Europe, Americas, and Africa ⁵.

The common sign of CRC is the unexplained blood/bleeding appearance in the stool, continues change of bowl habit, stomach discomfort, unusual loss of weight, unreasonable felling of tired, weakness and vomiting ^{17,18}.

There are plenty of risk factors for CRC, which can be divided as hereditary factors, modifiable risk factors and other factors ⁶. There are around 20% cancer patients have been proven related with familiar genetic factors ¹⁹. Some modifiable risk factors like smoking, red meat and body have been wildly convinced as positive factors for CRC ⁶; other factor like

fish, whole grains and physical activity are known for their negative roles in the process of CRC, which are good for health ⁶. CRC risk factors can also be divided into sociodemographic factors, medical factors, lifestyle factors and diet factors. Figure 1 displays the risk level distribution for CRC risk factors. (Data from Hermann et al.²⁰)

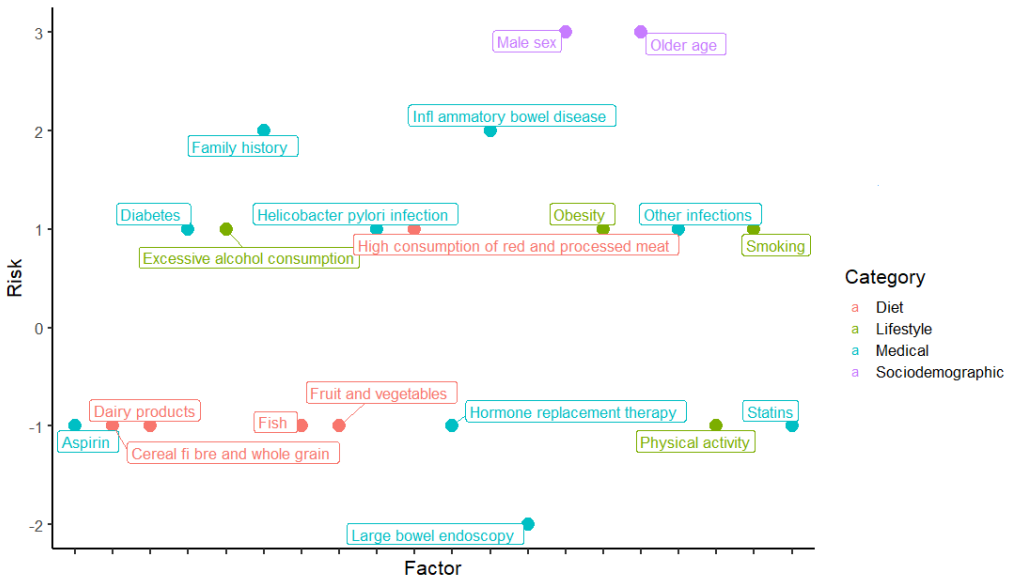


Figure 1. Distribution of risk factors for CRC. X-lab is the risk level for these factors, and the positive value represents the positive risk factor, which is ranged as 1, 2 and 3, according to its risk level. The negative risk factor is shown as negative value as -1, -2 and -3. The higher absolute value, the bigger risk level.

1.1.1 Colorectal cancer diagnosis

The outcome of CRC follows the specific rules: early-stage of CRC patients always have a better prognosis than late-stage. The stage I CRC patients have a 5-year survival rate for more than 90% ²⁰. However, if the CRC patients are diagnosed at stage IV, the 5-year survival rate will turn to 10% ²⁰. Unfortunately, more than 50% patients are already at late stage (stage III and IV) ²¹. Therefore, the accurate early screening and diagnosis of high-risk population and CRC early-stage patients are extremely important. The high-risk population like persons with family history is recommended to

make CRC screening regularly: colonoscopy every 10 years, computed tomography (CT) colonography every 5 years, fecal immunochemical test (FIT) 3 years, and flexible sigmoidoscopy every 5-10 years ⁵.

The conventional diagnostic tests for CRC diagnosis are physical exam and history, digital rectal exam, fecal occult blood test (FOBT), barium enema, sigmoidoscopy, colonoscopy, and biopsy ²², of which colonoscopy and biopsy test have been considered as golden test. One of the important advantage for colonoscope as golden test is that the diagnosis and treatment can be conducted at the same time ⁵.

However, some CRC patients do not show the typical performance in these tests, which make it even more difficult to diagnose these patients accurately. The common disadvantages of colonoscope are the procedural risks such as perforation, bleeding and aspiration, and the high cost ⁵. The precision medicine requires more accurate and personalized diagnosis for specific patients ²³. Benefits from its advantage in detection, calculation and stability, biomarker is considered as a suitable solution for the CRC diagnosis in precision medicine ²⁴.

1.1.2 Colorectal cancer treatment

Surgery, radiotherapy, chemotherapy, immunotherapy and targeted therapies are the main treatment methods for CRC. According to the guide of precision medicine, the treatment for CRC needs to be more precision to decrease the pain and cost of patients.

Surgery

Surgery is the basis for CRC treatment. The total mesorectal excision is the most important recent advance in rectal surgery, which significantly decreases the rate of local recurrence ²⁵. Based on right timing and good skill, surgery can reach good prognosis in rectal cancer, even without radiotherapy ²⁶. Fast-track surgery is an effective development for CRC surgery, since many procedures in traditional surgery can be saved, by which the cost and time for surgery, and the pain for patients both in body and mentality could be reduced ⁷. For late-stage CRC patients, the combination therapy of aggressive cytoreduction with hyper thermic intraperitoneal could be a new choice ^{7,27}.

Laparoscopic surgery is a safe choice for CRC patients. Long-term results do not show significant difference with common methods, but short-term outcomes are better ⁷.

Radiotherapy

Radiotherapy has become a mature treatment strategy for rectal cancer patients to reduce local recurrence and promote prognosis^{7,28}. For stage II and stage III rectal cancer patients, radiotherapy has been a standard treatment method⁷, which has been convinced to decrease the morbidity²⁹.

Preoperative radiotherapy has been suggested to use for rectal cancer patients. In a 700 patients prospective randomized trial published on 1975, five year survival rate for patients with preoperative radiotherapy is 48.5%, significantly higher than the rate of 38.8% for controls, which convinced that preoperative radiotherapy can improve the prognosis for rectal cancer patients³⁰. In the past years, more and more studies convinced this conclusion. The advantages for preoperative radiotherapy are: 1. Better effect in killing tumor cells, since they are much sensitivity for radiotherapy; 2. More precision in targeting tumor sessions; 3. Less treatment time and cost³¹.

However, the use of radiotherapy in colon cancer is still concerned, since the colon is moveable, which make it hard for radiotherapy to find the right target³¹. Further, the dose-limiting structures around the colon is another question for radiotherapy in colon cancer³¹.

Chemotherapy

Fluorouracil-based adjuvant chemotherapy has been recommended as standard treatment strategy for colon cancer patients in stage III but not stage II^{28,32}. Fluorouracil (5-FU) is a pyrimidine analog, which has been used in cancer therapy for more than 40 years, especially in CRC and prostate cancer³³. 5-FU belongs to the antimetabolite and has often been used in clinical together with leucovorin³⁴. As a thymidylate synthase inhibitor, 5-FU can block the synthesis of thymine (an essential material of DNA replication) to inhibit the tumor cell division³⁵. Since it can also act on the normal rapidly dividing cells like gastrointestinal epithelial cells and germ cells, 5-FU may cause some side effects like severe dehydration, enteritis and renal impairment³⁶. Therefore, it is still a challenge to develop precise target-guided therapy for 5-FU.

The effect of combination of chemotherapy and radiotherapy is still in discussion. A study enrolled 1011 patients suggested that adding chemotherapy to the rectal cancer patients with preoperative radiotherapy cannot increase survival significantly³⁷.

Immunotherapy

Immunotherapy (biotherapy) is to use human immune system to against cancers, by the substances made by human or lab to guide or restore the body immune ability²². Nowadays, immunotherapy has been an important treatment for CRC³⁸, which shows good performance especially in tumors with high microsatellite instability³⁹. A combination of immunotherapy with nivolumab and ipilimumab has received the approval by the US food and Drug Administration²⁸.

Targeted therapy

Targeted therapy is a drug therapy that prevents the growth of specific molecular to prevent and weak the tumor development, instead of the traditional chemotherapy that interferes with all the rapidly dividing cells⁴⁰.

Targeted therapy is more used in stage IV colon cancer patients. Targeted therapy could improve the treatment effect in the patients that do not show sensitivity in chemotherapy. The choice of targeted therapy in clinical depends on whether the cancer is resectable⁴¹.

1.1.3 Colorectal cancer prognosis

The prognosis of CRC has been improved steadily in the past years²⁰. The 5-year survival rate has been used as a typical method to measure the CRC prognosis in clinical. The prognosis of CRC is highly related to the medicine environment. In some developed countries like Canada and Australia, 5-year survival rate have reached around 65%²⁰. However, in most of developing countries, it is still less than 50%⁴²⁻⁴⁴.

Patient, therapy and tumor related factors are the three major prognosis factors for CRC⁷. The situation of body and mental of patient is high related to the CRC prognosis. The quality of surgery is one of the most crucial factors for CRC prognosis.

Diagnosis stage is the most important prognosis factor for CRC. The TNM stage system, as the most widely used cancer stage system, has been used in CRC for many years. In TNM stage system, “T” represents the primary tumor, “N” is the regional lymph nodes, and “M” reflects the distant metastasis. All these three effects are defined with different level in clinical, and the final stage of CRC (I, II, III, and IV) are decided based on the combined situation of them. The prognosis of CRC is highly related with the stage of patients. The 5-year overall survival rate for stage I is around 80-95%, for stage II is 65%-75%, for stage III is 35%-60%, and for stage V is 0%-7%⁴⁵⁻⁴⁹.

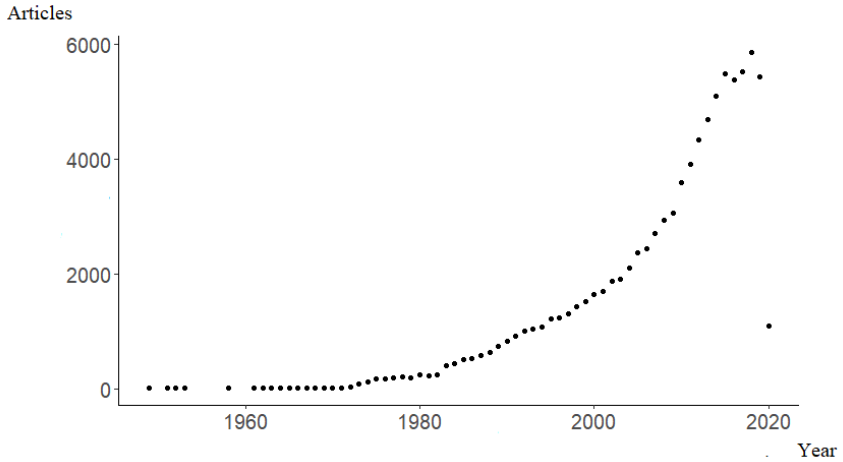
1.2 Biomarkers

Biomarkers are specific biological indicators in the human body that can serve as a marker or indicator for diseases⁹, which have been widely used to improve the diagnosis, therapy and prognosis in many human diseases. According to their biological components, biomarkers are categorized into three main groups: DNA, RNA and protein biomarkers. Proteins are the main executor for biological functions, which have been studied most among all kinds of biomarkers⁵⁰. Recently, owing to their stable structure, specific detectability and altered expression, some non-coding RNAs like microRNAs (miRNAs) and long noncoding RNAs (lncRNAs) have become new sources for biomarker discovery⁵¹⁻⁵³.

According to their applications in clinical, biomarkers are divided into three categories: diagnosis, treatment and prognosis biomarkers. Since the development of disease is a continuous process, the relationships for diagnosis, treatment and prognosis are close. Many studies demonstrate that some biomarkers can be applied in several aspects in diagnosis, treatment or prognosis⁵⁴⁻⁵⁶.

1.2.1 Biomarkers in colorectal cancer

Until March 7th, 2020, there are 82577 papers in PubMed concerning CRC biomarker researches, and the paper amounts have a significant increasing tendency followed by year. Figure 2 shows the search result for biomarker in CRC from PubMed.



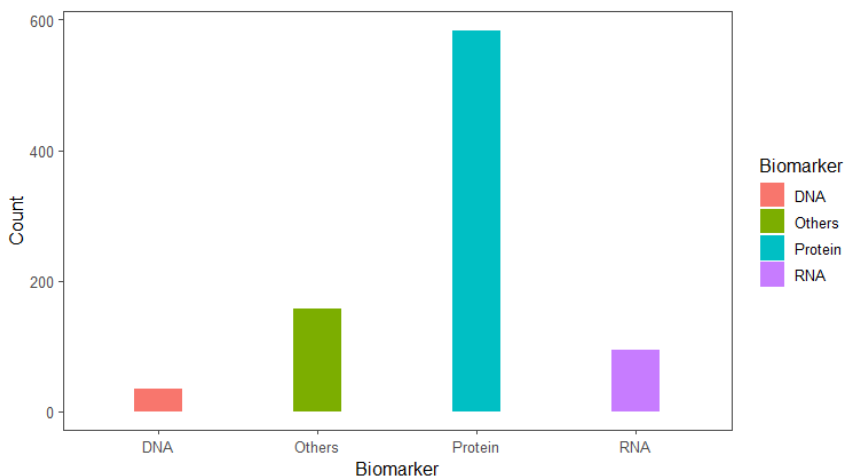


Figure 3. CRC biomarker distribution.

DNA biomarkers

Fecal DNA biomarkers combined with FIT has been a common method to diagnose CRC, which is recommended by the US Multi-Society Task Force on Colorectal Cancer ⁵⁸. DNA-FIT shows better sensitivity, but lower specificity compared with FIT alone ⁵. High cost is the most significant shortage of DNA-FIT.

Cell-free DNA (cfDNA) is the fragment of degraded DNA that detected in the blood plasma or sera, including circulating tumor DNA (ctDNA) and cell-free fetal DNA (cffDNA) ^{59,60}. cfDNA has been considered as “liquid biopsy” in cancer study ⁸. Further, many researchers report that cfDNA can be effective universal biomarker in other complex diseases sepsis, diabetes and stroke.

Recently, mutations in DNA mismatch repair (MMR) genes have been reported as novel biomarkers for CRC. DNA MMR is a system to detect and repair the consecutive erroneous insertions, deletions, and erroneous merges that may occur during DNA replication and recombination ⁶¹. Microsatellite instability (MSI) is the genetic hypermutability related condition caused by impaired MMR, which has been used as biomarkers since its high correlation with cancer prognosis ⁶².

RNA biomarkers

RNA is the second leading component of CRC biomarker. There are plenty of RNAs have been reported as CRC biomarkers, and Figure 4 shows the distribution of RNA biomarkers. There are 72 miRNAs have been reported as biomarkers in the diagnosis, treatment and prognosis ⁵⁰.

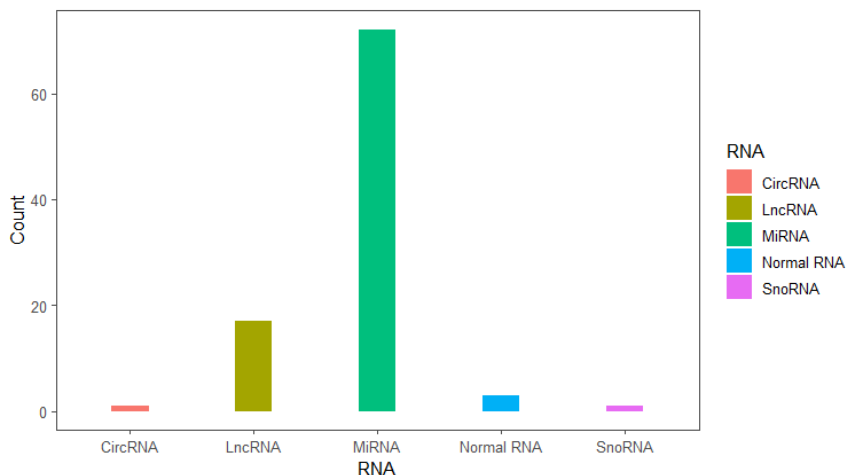


Figure 4. Distribution of RNA biomarkers.

MicroRNA-21 (miR-21), encoded by IR21 gene, is one of the earliest miRNAs that have been identified, which has been reported as downregulated biomarker in many cancers ^{63,64}. MiR-21 has been used in the CRC biomarker research for many years, and there are several publications reflect that miRNA-21 could be a promising biomarker in the diagnosis, treatment and prognosis of CRC ^{50,65}. In 2017, Peng et al. reported that circulating miR-21 was better at diagnosis, and tissue miR-21 could be a better prognosis biomarker in CRC ⁶⁶. Further, the combination of miR-21 with other miRNAs as multiple biomarkers could reach better performance than single miRNA biomarkers ⁶⁶.

Circular RNA (circRNA) is a new direction for RNA biomarker discovery. Different from normal linear RNAs, circRNA presents a continuous loop since the lack of 3' and 5' end ⁶⁷. CircRNA has been reported corrected with many complex diseases, such as cancer, diabetes and heart diseases ⁶⁸⁻⁷². Some studies suggest that circRNA is related to the growth of

cancer cell, the occurrence of cancer metastasis, and the resistance for cancer drug ⁷³. Recently many studies report that circRNA could be optional biomarker for CRC. The reason for circRNA to be CRC biomarker is that 1) circRNA has been detected in many human environments like blood, gastric fluid and saliva; 2) the expression of circRNA has high specificity, stability and universality ⁷⁴⁻⁷⁷. For the biological reason of circRNA as biomarker, the relationship between circRNA and cancer still needs to be further detected ⁶⁷.

Protein biomarkers

Protein biomarker is the major component of CRC biomarker. Many scientists believe that protein biomarker is the most accurate biomarker since protein is the main executor for human life activities. There are 583 protein biomarkers are recorded in the CRC biomarker database (CBD), including different kinds of proteins.

TP53

TP53 (protein name: p53) is a typical tumor suppressor gene, locates on the short arm of chromosome 17 ⁶⁰. TP 53 plays regulating rule in many activities, such as cell growth, apoptosis, and genetic stability ⁷⁸. As the most mutated gene in cancer, TP53 was first discovered mutated in CRC in 20th ⁷⁹. Around 50% CRC patients were found with significant TP53 mutation ^{80,81}. Many studies have convinced that TP53 plays an essential role in the development of tumor cell ⁸²⁻⁸⁴. It is clear that TP53 can detect the tumor cell and guide them to apoptosis ⁸⁵. The ability of TP53 would sometimes lose when mutations happening, which will increase the possibility of cancer occurrence ^{84,86}. There are 20 researches reported that TP53 could serve as biomarker for CRC ⁸⁷. TP53 has been reported as biomarker in the diagnosis, treatment and prognosis of CRC ^{55,56,88}.

KRAS

As a typical oncogene, around 60% CRC patients were detected with KRAS mutation ⁸⁹. In the CBD database, there are some studies reported that KRAS could be useful biomarker for CRC, of which supposed that KRAS could be prognostic biomarker ⁵⁰.

MKI67

MKI67 (ki-67) has been convinced as biomarker for cellular proliferation, since it's essential role in cell growth ⁹⁰. Several studies have shown that

MKI67 could be prognosis or treatment biomarker for CRC ⁵⁰, since the prognosis and therapy is highly related to the tumor cell situation ⁹¹.

CEA

Carcinoembryonic antigen (CEA) is one of the most being researched biomarkers for CRC, which has been widely used in clinical, especially in the detection of liver metastasis. The earliest research recorded in the CRC biomarker database (CBD) is in 1987, Davey et al. reported that CEA could be prognosis biomarker in the radiotherapy and recurrence of rectal cancer ⁹². CEA has been suggested as a useful diagnostic biomarker for CRC ⁵⁰. However, the sensitivity of CEA test to detect CRC is still questioned ⁹³.

CHGA

Chromogranin-A (gene name: CHGA) is a 439-residue-long protein in neuroendocrine cells, which plays a crucial role in the co-stored and co-released of protein. CHGA has been convinced as an important biomarker for neuroendocrine neoplasms. Several studies have revealed that CHGA is related to human cancers: Yang et al. reported that CHGA could be promising biomarker for gastric cancer ⁹⁴; Ma et al. found that CHGA could be used as prognosis biomarker for prostate cancer ⁹⁵; Weisbrod et al. suggested that CHGA could serve as prognosis biomarker in pancreatic neuroendocrine tumors ⁹⁶.

ABI2

ABI2 (Abl Interactor 2) is a gene focusing on the protein coding ⁹⁷. Some studies suppose that the ABI2 protein may contribute in the regulation of cell growth and transformation ⁹⁸. A recent study by Meng et al. identified three novel miRNA biomarkers (miR-302a, miR-105 and miR-888) by PCR and bioinformatics analysis ⁹⁹. Further, they found that these three miRNAs all had strong relationships with ABI2, on the miRNA-gene interaction network. Therefore, ABI2 is supposed as a future novel biomarker for rectal cancer, and the followed verification from hundreds of rectal cancer patients and normal controls convinced that ABI2 could be promising biomarker for the diagnosis but not prognosis of rectal cancer ⁹⁹.

Other biomarkers

Besides the traditional biomarkers, scientists also discovered other kinds of biomarkers in different components. Image biomarker is the biomarker as image form ¹⁰⁰, such as X-ray, computed tomography (CT), magnetic

resonance imaging (MRI). With the development of artificial intelligence (AI) technology, many medicine fields have benefited from it. This happens in cancer research too, especially the image biomarker detection and improvement. Recently, a study from the University of Michigan shows that using AI technology together with imaging detection, the accurate diagnosis time has decreased in less than 3 minutes in brain cancer surgery ¹⁰¹.

Biomarkers in colorectal cancer diagnosis

It has been wildly convinced that biomarker can improve the diagnosis accuracy of CRC ¹⁰². There are plenty of diagnosis biomarkers for CRC ¹⁰³. The change of biomarker expression level could be a guider for CRC diagnosis. Now, some biomarkers like TP53 and CEA have been used in clinical, as the assistant for diagnosis. Scientists are still trying to find the “perfect” biomarker to reach the ideal level of accurate diagnosis of CRC.

Sensitivity (true positive rate (TPR)) and specificity (true negative rate (TNR)) have been wildly used as the statistics effect value for the measurement of CRC biomarker diagnostic accuracy. Sensitivity is used to measure the proportion of the true diagnosed patients by the biomarker in the total patients. On the other hand, specificity is used to calculate the rate of true diagnosed non-patients by the biomarker in the total healthy people. Normally, 0.6 is a cut off for sensitivity and specificity. A biomarker with a sensitivity higher than 0.8 or even 0.9 is considered with good ability in diagnosing patients, and for specificity, same cut off indicates good ability in diagnosing non-patients. The ideal biomarker should have both good sensitivity and specificity. However, there is always a phenomenon that most of biomarker can only occupy one good value in sensitivity or specificity. For example, TP53 has been used in clinical since its high sensitivity. But the specificity of it is always been concerned.

As shown in Figure 5: The percentage of true diagnosed patients by the biomarker in the real total patients is the True positivity (TP), and the rate of false diagnosed patients by the biomarker in the real total patients is the False positivity (FP). Same rates in healthy people is called True negative (TN) and False negative (FN). The formulas of sensitivity and specificity are as following:

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (TN + FP)$$

		Real condition	
		Patient	Normal
Prediction condition	Patient	TP	FP
	Normal	FN	TN

Figure 5. 2x2 table in biomarker diagnosis test

In clinical, people are more likely to use another pair of statistics effect size: positive predictive value (PPV) and negative predictive value (NPV). In machine learning, PPV is more called as precision, which is an important value to judge the accuracy of prediction model. The formulas for PPV and NPV are as following:

$$PPV = TP / (TP + FP)$$

$$NPV = TN / (TN + FN)$$

In order to give a systemic view of diagnosis accuracy, receiver operating characteristic curve (ROC) test has been created to combine the sensitivity and specificity together in a plot. The Y-axis in ROC curve is sensitivity, and the X-axis is 1-specificity. The area under the ROC curve (AUC) is considered as a digitized value to evaluate the diagnosis accuracy of biomarker. AUC is between 0 to 1. The higher AUC, the better accuracy of the biomarker.

Biomarkers in colorectal cancer treatment

The biomarkers in the treatment of CRC is usually to guide the drug selection and dosage in clinical, according to the expression of them. Many biomarkers can serve both in the treatment and prognosis of CRC, since the treatment and prognosis are highly related to the development of tumor. There are 152 treatment biomarkers for CRC, which occupies the last amounts in the distribution of CRC biomarkers¹⁰³. The reason may be that because the treatment of CRC is a complex process, which is highly different from in different patients. Therefore, it is a challenge to find common biomarkers to guide CRC treatment.

The common treatment biomarkers for CRC are mismatch-repair deficiency, epidermal growth factor receptor (EGFR), BRAF, PIK3CA and PTEN etc⁹.

Biomarkers in colorectal cancer prognosis

Prognosis biomarkers occupies most in the CRC biomarker distribution (707)⁵⁰. The most significant application of biomarker in CRC prognosis is to help doctors divide CRC patients into right stage more accurately and quickly. Some famous biomarkers like TP53 and KRAS have been used in clinical prognosis of CRC for many years, which reflects the importance of biomarker in prognosis. However, it is still needed to detect new key biomarkers for CRC prognosis, according to the theory of precision medicine, which aims to provide personalized medicine for specific patients¹⁰⁴.

The main prognosis biomarker in CRC prognosis is CEA⁵⁷. TP53 and RAS family have been investigated as the prognosis biomarker in CRC^{105,106}. More and more different kinds of prognosis biomarker for CRC prognosis have been reported¹⁰⁷⁻¹¹⁰.

The effects of treatment and prognosis biomarker can be assessed with hazard ratio (HR), odds ratio (OR) and risk ratio (RR). If the p value of HR, OR or RR less than 0.05, the biomarker is considered with significant effect.

Multiple-functional biomarker

Recently more and more studies reported that specific biomarkers could be served in CRC in more than one aspect. We call them ‘multiple-functional biomarker’. For example, TP53 could be used as CRC biomarker in diagnosis, treatment and prognosis^{54,55}. Figure 6 shows the distribution of multiple-functional biomarkers in CRC. There were 64 CRC biomarkers that have been convinced in treatment & prognosis, 11 of them have been

used in diagnosis and treatment, and 38 have been reported in diagnosis and prognosis. Three biomarkers can be served in all three aspects of CRC: diagnosis, treatment and prognosis ¹⁰³.

The reason of the appearance of multiple-functional biomarkers is that the progression of CRC is a systemic continues process, and the diagnosis, treatment and prognosis are with close relations. On the other hand, some key genes in CRC like TP53 are highly associated with the whole process of CRC.

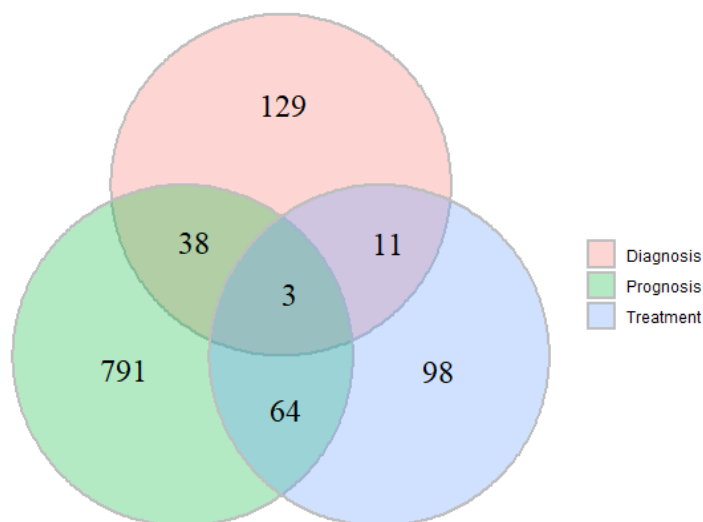


Figure 6. Venn plot for CRC biomarker distribution in diagnosis, treatment and prognosis.

Multiple biomarkers

Although more and more biomarkers have been discovered and reported, the effects of these biomarkers are still questioned. One of the possible reasons is that CRC is a multigene disease. Hence, many scientists suggest that combining different combining different single biomarkers together as “multiple biomarkers” could be a strategy ¹¹¹. Many studies have convinced that multiple biomarkers could improve the diagnosis, treatment and prognosis value significantly ^{112,113}. The methods to combine biomarkers can be generally categorized into tow aspects: 1. Measure different biomarkers in different timing according to their expression regulation, then calculate

point or percentage of positive expressional biomarkers, which will be considered the final result for diagnosis decision; 2. Use algorithm to combine the expression level of different biomarkers to get a specific point, as the final evidence for diagnosis. The most common and simple way is logistic regression: using the expression of biomarkers as independent variables and the sample situation (patients or not) as dependent variable. Then input the variables collected from known samples as train data into logistic regression model to train the model, and finally we can get a final model including the coefficient of the expression of each biomarker. Using the final model, further diagnosis test can be conducted. The first measure has been used commonly both in clinical and lab now. However, as the accumulation of more and more biomedicine data for CRC and the development of computational methods, the second method would be more accurate and popular.

1.2.2 Biomarker detection

Since the discovery of biomarkers, various detection methods for them have been created and applied in biomedicine field. This section will display the popular detection approaches in wet lab, and the dry-lab approaches would be introduced in section 1.3.

Genomic technologies

The genomic approach for biomarker discovery including genome wide methods like microarrays, splicing expression profiling, and serial analysis of gene expression (SAGE); and individual gene sequences like quantitative reverse transcription polymerase chain reaction (qRT-PCR) ¹¹⁴.

Proteomic technologies

Proteomic technology has always been a common way to detect biomarker. The traditional methods in proteomic biomarker discovery are gel electrophoresis, protein array, enzyme-linked immunosorbent assay (ELISA) and liquid chromatography ¹¹⁴.

Imaging technologies

Imaging technology is the most direct way to detect new biomarkers, of which microscope is the most used method, including light microscope and electron microscope. Other imaging approach like X-ray and MRI have also been used as a common way for biomarker discovery ¹¹⁴.

1.3 Bioinformatics approach

With the development of computer technology and the accumulation of huge biomedicine data, bioinformatics has been playing a crucial role in the biomarker discovery ¹¹⁵.

It has been convinced that the biomarker discovery is a comprehensive and continues system, in which bioinformatics should be both the beginning and ending of biomarker discovery. In the beginning, precious prediction based on large omics data by bioinformatics could provide specific target for biomarker discovery. After the verification and investigation by traditional and novel wet-lab experiment, bioinformatics can be used to further verify the result and guide future direction.

GE approach has been widely used to predict new biomarkers and verify identified biomarkers in CRC. For personalized medicine, GE based biomarker discovery is a good way since it can find suitable biomarkers for specific samples. However, it is questioned that the biomarkers found by GE data can be serve as common biomarker for worldwide population.

Network theory has been applied in bioinformatics for many years, since the biological system is a big network consisted by different biological components, and each component plays specific role in the biological network. Many studies report that biological network shares some common rules with other networks like human society ^{116,117}. Further, some researches declare that biomarkers occupy specific position on some biological networks such as protein-protein interaction (PPI) network and miRNA-mRNA interaction network, which inspires scientists to predict and verify CRC biomarkers on these networks according to the topology features ^{15,16}.

1.3.1 Biomedicine databases

Biomedicine database is the internet-based library storing the scientific information for biological issues, which is one of the most important foundations for bioinformatics or biomedicine study. The information stored in biomedicine databases could be sequencing and structure data, clinical disease and sample information, as well as other biological data generated from dry or wet experiment in genomics, metabolomics and proteomics. There are different kinds of databases widely been used in biomedicine studies.

PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>)

PubMed, created and managed by the American national library of medicine, has been the most popular and authoritative text-mining based database for scientific paper searching. PubMed is the initial door for most biomedicine students to make research. In CRC biomarker field, PubMed has recorded more than 80000 related papers. (Figure 1)

SEER (<https://seer.cancer.gov/>)

The Surveillance, Epidemiology, and End Results (SEER) Program collects and displays huge amounts of cancer statistics data from American populations, which has been a popular platform for cancer researchers to search and analysis cancer information. Until March 9, 2020, there are 64,600 papers related to SEER recorded on PubMed.

TCGA

(<https://www.cancer.gov/aboutnci/organization/ccg/research/structural-genomics/tcga>)

The Cancer Genome Atlas (TCGA) is one of the most popular cancer databases which contains huge amounts of omics data for cancer patients collected by the American National Cancer Institute. TCGA contains the omics data from more than 20000 cancer patients, which has been a powerful data foundation for cancer related studies. Every year, there are more and more studies use the data downloaded from TCGA. Using “TCGA” as key words searching in PubMed, more than 7000 related research records will appear.

Xena (<http://xena.ucsc.edu/>)

The UCSC Xena is a public platform integrated with multiple genomic data collected from popular databases like TCGA and GTEx (contains the omics data of healthy population) and other individual experiments, which is developed by the UCSC Computational Genomics Laboratory from The Regents of the University of California ¹¹⁸. Comparing with other similar databases, Xena occupies the outstanding position benefiting for its powerful visualization and analysis function. Xena used a novel pipeline to combine the data from different source as same format, which makes it possible to analysis these data together. What’s more, users can also submit and analysis their own data on Xena.

GEPIA database (<http://gepia.cancer-pku.cn/>)

The GEPIA database is an interactive database established by Zhang et al. from Peking University ¹¹⁹. On the foundation of standard data from Xena, GEPIA provides more functions for cancer RNA-seq data analysis, such as differential expressional gene (DEG) analysis for specific cancer, and survival and expression analysis for specific gene. Table 1 presents the top 10 DEGs for colon cancer and rectal cancer in GEPIA. (calculated by ANOVA algorithm) In 2019, CEPIA2 has been published, which adding the function of analyzing users' own data ¹²⁰.

Table 1. DEGs in colon (A)/rectal (B) cancer patients with normal controls.

A.

Gene Symbol	Median (Tumor)	Median (Normal)	Log2 (FC)	P value
RP11-40C6.2	1090.972	1.620	8.703	1.01e-151
CEACAM6	488.359	4.060	6.596	4.87e-77
DPEP1	111.113	0.480	6.243	1.75e-142
S100P	516.275	6.310	6.145	5.08e-123
LCN2	489.918	6.530	6.027	2.75e-65
CEACAM5	1586.904	27.971	5.776	5.53e-49
CLDN2	45.972	0.090	5.429	1.34e-124
ETV4	67.798	0.750	5.297	7.54e-267
CDH3	47.609	0.270	5.258	9.29e-301
MMP7	37.139	0.090	5.129	2.81e-136

B.

Gene Symbol	Median (Tumor)	Median (Normal)	Log2 (FC)	P value
RP11-40C6.2	1090.972	1.620	8.703	1.01e-151
CEACAM6	488.359	4.060	6.596	4.87e-77
DPEP1	111.113	0.480	6.243	1.75e-142
S100P	516.275	6.310	6.145	5.08e-123
LCN2	489.918	6.530	6.027	2.75e-65
CEACAM5	1586.904	27.971	5.776	5.53e-49
CLDN2	45.972	0.090	5.429	1.34e-124
ETV4	67.798	0.750	5.297	7.54e-267
CDH3	47.609	0.270	5.258	9.29e-301
MMP7	37.139	0.090	5.129	2.81e-136

P value has been adjusted.

String database (<https://string-db.org/>)

The string database is the most popular database for protein-protein interaction network¹³. The powerful visualization function is the important reason that makes String outstanding among various PPI databases. Figure 7 displays the PPI network for the proteins including in this thesis. Further, it keeps regular updating since it was first developed at the year of 2000¹³. The most important part for a database is the quality and amount of data containing in it: more than 2000 million interactions of 24.6 million proteins collected from 5090 organisms have been recorded and displayed in the newest version of String (v11)¹²¹.

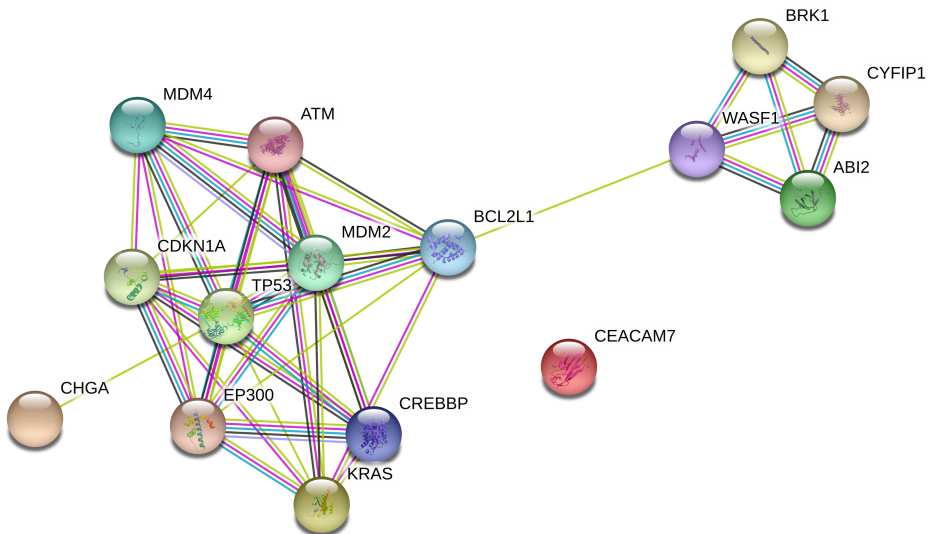


Figure 7. PPI network for the protein biomarker in this introduction (TP53, GEA, KRAS, MKI67, CHGA and ABI2), generated by String. (Accessed 2020/03/13)

miRNAnet (<https://www.mirnet.ca/>)

The miRNAnet database is a powerful tool that collects multiple information for miRNAs, focusing on the interaction networks for miRNAs¹⁴. miRNAnet not only contains the information for human but also the information for other species like mouse, rat, cattle, pig, and zebrafish, etc.

An attractive advantage for miRNAnet is its regular and frequent updating¹²². (almost every week) miRNAnet integrates multiple miRNA related interaction knowledge for diverse biological components such as genes, non-coding RNAs (ncRNAs), epigenetic modifiers, transcription factors, diseases, and

small biological compounds. Meanwhile, miRNet also contains different types of miRNA data generated from multiple experiments like RT-qPCR and next generation sequencing. As a network-based database, miRNet also has integrative and user-friendly visualization function. Further, miRNet provides the functions of pathway enrichment analysis and Gene ontology annotation for related genes of target miRNAs. Figure 8 shows the miRNA related interaction networks for the miRNA biomarkers involving in this study, which also integrated the PPI network for related genes.

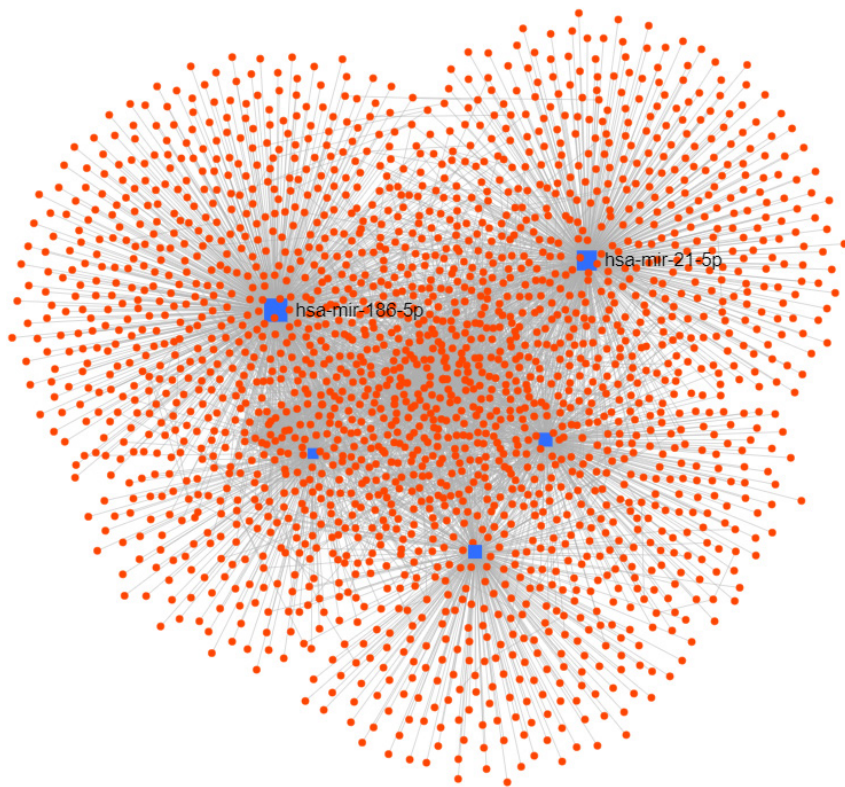


Figure 8. miRNA-gene interaction network for miR-21-5p, miR-186-5p, miR-30e-5p, miR-31-5p and miR-10b-5p.

Biomarker databases

Biomarker database is a kind of biological database specifically concerning on biomarkers. There are some biomarker databases have been created, such as the Global Online Biomarker Database ¹²³, which has collected >130000 biomarkers for 18 different therapeutic areas, and the America National Cancer Institute has established the early detection research network, which collects 13 CRC biomarkers ¹²⁴. There are some specific biomarker databases for various diseases, such as the US Environmental Protection Agency Biomarker which has collected the biomarkers for children diseases ¹²⁵; the Tuberculosis Biomarker Database includes the tuberculosis associated biomarkers ¹²⁶; the Infectious Disease Biomarker Database with the biomarkers for the infection diseases ¹²⁷.

OncoMx is a compressive knowledge which contains cancer biomarker information collected from different sources ¹²⁸. However, the detention of biomarkers in OncoMx is questioned. Gastric Cancer (Biomarkers) Knowledgebase ¹²⁹, and the Liver Cancer bioMarker Reference Into Function database (LiverCancerMarkerRIF) ¹³⁰.

Database construction process

Normally the construction of a biomedicine database follows the common process of establishing a basic database ¹⁷:

1. Requirement analysis: collect and analysis the requirement for database, from candidate users and related specialists;
2. Conceptual design: transfer the realistic requirement to computer conceptual;
3. Logical design: set the logical concept in database, with primary and foreign keys;
4. Normalization process: normalize the data with standard format and remove the redundancy;
5. Physical design: design the database in real and chose the right database manger system (DBMS). The are many popular DBMS, such as Oracle, MySQL, Access, and Microsoft SQL Server, etc.

Database construction tools

There are plenty of tools for database construction, among which, “XAMPP” software package has been wildly used. “X” is the operation system for database construction and management, which usually can be Windows, Linux, Mac OS X, or Solaris. “A” represents to Apache, which is used as HTTP server. “M” refers to MySQL DBMS. “P” is the PHP and

Perl languages, which are used to connect the DBMS with HTML web page. Benefiting with its outstanding advantages like free, user-friendly, and open source, XAMPP has been one of the most common tools for database construction and management.

1.3.2 Complex network

Network graph theory was proposed in the Eighteenth century. After years of development, random graph theory, small world theory and scale free theory have been the three breakthroughs in network graph theory. Before the introduction of these three theories, we should know what is “Regular network”. The most obvious characteristic of regular network is that each nodes has the same number of lines on it ¹³¹. To the opposite of regular network, the numbers of lines connected to points are randomly distributed on random network, which is definitely a disorder system ¹³². Scale free network is a highly connected network. Comparing with random network, the degree distribution on scale free network follows a power law distribution: there exists some hub points, whose connected lines are more than other points, significantly. On small world network, most of points are not connected directly, but they can reach to another by small neighbors, which have been used in human society network. For example, Six Degrees of Separation supposes that one person can contact to anyone in the world in six people.

Biological environment is a connecting complex system, which is constructed by multiple interaction relationships, i.e. PPI, RNA-RNA interaction (RRI), protein-DNA interaction etc.¹³³ These interaction relationships can be transferred to complex networks in network theory, where the nodes (points, vertices) on the network could represent the biological components and the lines (edges, links) represent the interaction relationship ¹³³. The shape, color and size of nodes could reflect different properties of biological object. The direction of lines could represent the relationships of objects, and the size and color could reflect the level and category of relationships. Systems biology supposes that by studying these interaction networks could help to explain and further investigate the rules of biological components ¹³³.

Complex network as an important research field has been developed in biomedicine during the last 30 years. Adequate evidences have shown that biological complex networks share similar features with human social networks. Several studies have reported that biomarkers occupy special position on some biological networks such as PPI networks and RRI

network^{15,103}. In the biomarker related network research field, many studies focus on the detection of network biomarkers¹². However, few studies work on the specific biomarker-biomarker interaction (BBI) network. If the amounts of biomarkers is big enough, the topology features on the BBI networks would be robust, by which we may get some common rules for biomarkers to further detect new better biomarkers. Table 2 presents the common topology features in the complex network.

Table 2. Common topology features

Feature	Definition
Average Shortest Path Length	$C(v) = \frac{\sum_w d(v, w)}{n - 1}$
Betweenness Centrality	$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$
Closeness Centrality	$C(x) = \frac{1}{\sum_y d(y, x)}$
Clustering Coefficient	$C = \frac{\text{number of closed triplets}}{\text{number of all triplets}}$
Degree	the number of edges incident to the point
Eccentricity	the biggest geodesic distance between the point and any other point
Radiality	$C_{rad}(v) = \frac{\sum_w d - d(v, w) + 1}{n - 1}$
Stress	$C_s(v) = \sum_{s \neq t \neq v} \sigma_{st}(v)$
Topological Coefficient	$T_n = \frac{avg(J(n, m))}{k_n}$

n: the number of the points

d: the diameter of the network.

$d(v_i, v_j)$: the shortest distance between v_i and v_j .

σ_{st} : the total number of shortest paths from node s to node t.

$\sigma_{st}(v)$: the total number of shortest paths from node s to node t that pass through v.

J (n, m): the number of neighbors shared between the nodes n and m, plus one if there is a direct link between n and m.

k_n : the neighbors of n

PPI networks

Benefits from the development of experimental technology and the accumulation of related data and knowledge, PPI network has been an important source for protein research.

It has been proven that biomarkers always occupy specific position in the PPI network^{23,103}. In the previous introduction, we introduced that some biomarkers could be served as multiple-functional biomarkers. These biomarkers always occupy hub position in the PPI network.

MiRNA-mRNA network

miRNA can regulate the messenger RNA (mRNA) in human cellular networks. As such, the regulation relationships between miRNAs and mRNAs have constructed multiple miRNA-mRNA interaction networks. Different from the PPI network, the direction of miRNA-mRNA network is single: only from miRNA to mRNA. In the previous work, our research group has created an algorithm to detect miRNA biomarker based on two topology features on miRNA-mRNA network¹³⁴.

Other networks

Other networks such as drug-target network and virus-host network have also been important tools in different biomedicine fields. For example, based on the existed drug-target networks, researchers can predict new drugs with similar network features with reported drugs. On the other hand, other targets could also be predicted.

1.3.3 Machine learning

AI is a technology to let computer detect and react for some situations like human intelligence system. Programmers tell AI model how to learn and act for related situations by design and perform related algorithm¹³⁵. As the development of AI technology and the accumulation of big data, AI is changing our daily life, including the medicine health care. In CRC research, AI can: improve the screen and diagnosis accuracy to support the clinical diagnosis; help to find the drug targets with less time and better precious, and find suitable drug based on target molecular, in order to promote the treatment¹³⁶; use the huge clinical patients data stored in related databases like TCGA and SEER and control data from databases like GTEX, to construct model to predict the outcome for CRC patients, by which to improve the prognosis. What's more, AI technology can simulate the

process of CRC patient and predict the possibility of CRC occurrence for healthy samples on dry lab, which can save both time and money.

As a subset of AI, machine learning (ML) is more focusing on to make prediction or decision rather than to conduct tasks¹³⁵. ML use the knowledge and regulation learned from sample data (training data) to build model to predict or classify target data. As the advent of big data era and the development of computer science, ML has developed to a mature method. Along with the development of biomedicine experiment equipment and technology, huge amounts of biomedicine data have been published and stored in plenty of databases, which could be good sources for training data in ML. Therefore, ML has been used for the new biomarker discovery in many bioinformatics studies. However, most of them use gene expression data to detect biosignature to find new biomarkers^{137,138}, which are not robust enough because of the heterogeneity among different population. Recently, many biological network databases have been established, such as String database for PPI¹³, miRNet database for miRNA related networks¹⁴, which all contain huge network data collected from different sources worldwide. As such, using biological network to predict new biomarkers could be a new sight in biomarker discovery.

Support vector machine

Support vector machine (SVM) has developed to a popular ML method since it was first reported at 1963. It is a supervised learning algorithm that analysis data in regression and classification model. Supervised learning algorithm and unsupervised learning are two major methods in ML. The difference of these two methods is that before establishing of model, supervised learning has clear expectation for output of model, but not unsupervised. Therefore, supervised learning is more used in regression and classification, and unsupervised learning is good at cluster.

The famous advantage of SVM is the use of “kernel” in construction of model, which can transfer the classifications of points into higher dimension to increase the prediction accuracy. Although this transformation may increase the generalization error, but it can still get good accuracy on the foundation of enough data.

Decision tree

Decision tree (DT) learning is one of the most popular ML method in the era of “AI”. DT is to pretend the structure of tree to divide the space of input data with a hyperplane. After enough dividing process, the candidate

data would be divided to the corresponding “leaf”. There are two parts for DT: classification DT for discrete data, and regression DT (regression tree) for continuous data.

Deeping learning

Deeping Learning (DL) is a subset of ML, which uses artificial neural network to learn and make prediction based on big data. Artificial Neural Network is a typical method in DL. Obviously, this method tries to learn and pretend the process like human neural systems. The application of DL has been widely used in cancer research. A recent study use DL to predict tumor mutation using pathology images from TCGA ¹³⁹.

Tools for ML

Popular software like R language, Python and matlab all have powerful packages for ML specifically. Here, some famous R packages will be introduced:

“CARAT” is a R package focusing on the regression and classification. With 217 powerful functions, “CARAT” has been considered as one of the best R package for ML ¹⁴⁰. “RandomForest” is a package for Random Forest, which could train missing values. “e1071” is a famous R package that shows high practicality especially in SVM.

1.3.4 Novel meta-analysis

Evidence-based medicine (EBM) is a new direction in medicine developed in the 1990s ¹⁴¹. The purpose of EBM is to use the results from high-quality researches, as evidence, to guide the clinical decision. EBM is considered as a research paradigm with highest credibility, and meta-analysis is the most credible EBM. Now the development of interdisciplinary research has got great improvement, which has been convinced with the ability to provide more comprehensive and accurate information for specific questions. Recently, the combinations of bioinformatics and EBM has generate some new types of meta-analysis.

Logistic regression based meta-analysis

How to combine GE data from different datasets has always been a key problem in bioinformatics analysis. Zhang et al. proposed a novel method to use logistic regression to deal with GE data to get the 2×2 diagnostic table ²³. Then input the table to diagnosis meta-analysis model to make traditional meta-analysis.

Bayesian meta-analysis of diagnostic test

Recently, combining the Bayesian theory into diagnosis meta-analysis has been a new direction in bioinformatics and EMB. The so called “Bayesian meta-analysis of diagnostic test can compare the diagnosis accuracy between two tests directly, and compare two not related tests by the common third test, which could provide a prediction result on the foundation of real result, based on the Bayesian theory. “bamdit” package in R language is professional in perform this novel meta-analysis.

2 The present investigation

2.1 Paper I

Zhang X, Sun X-F, Cao Y, Ye B, Peng Q, Liu X, Shen B and Zhang H
CBD: a biomarker database for colorectal cancer.

Database 10.1093/database/bay046, 2018

2.1.1 Background and aims

With more and more CRC biomarkers been discovered and reported, there is an emergency need to establish a compressive database to collect, normalize and display these biomarkers information. This paper aimed to construct a CRC biomarker database to record all the reported CRC biomarkers published on PubMed, which would provide users a platform for searching, downloading, analyzing and submitting the CRC biomarkers with their biomedicine information.

2.1.2 Materials and methods

PubMed provided the data source, and the key words for searching is “(((biomarker OR marker) OR indicator) OR predictor) AND ((colorectal cancer OR rectal cancer) OR bowel cancer).” “WAMP” software package was used to construct the database: Windows system as operation system; Apache as database server; MySQL as DBMS; PHP as the bridge to connect the database with HTML webpage.

2.1.3 Results and discussions

870 biomarkers selected from 8753 CRC biomarker related articles were collected and included in our CRC biomarker database (CBD). These biomarkers were categorized into DNA, RNA, protein and other biomarkers according to their biological components. Further, the CBD provided multiple searching strategies and user-friendly webpage for users to search and download target data. Further system analysis is needed to find hub biomarkers and some common rules for these biomarkers to guide future biomarker prediction.

2.2 Paper II

Zhang X Sun X-F, Shen B and Zhang H

Potential applications of DNA, RNA and protein biomarkers in diagnosis, therapy and prognosis for colorectal cancer: a study from databases to AI-assisted verification.

Cancers 11, 172, 2019

2.2.1 Background and aims

In the CBD database, we collected all the 870 reported biomarkers for CRC, and further system analysis is needed. Many researches suppose that the performance of single biomarker is concerned and combining different biomarkers as multiple biomarkers could be a new direction to improve the diagnosis and prognosis of CRC.

2.2.2 Materials and methods

The 870 CRC biomarkers were analyzed on biological networks and pathways. The overlap of miRNA and protein biomarkers related genes were used to predict new multiple biomarkers based on the PPI network.

2.2.3 Results and discussions

CRC biomarkers were clustered to diagnosis, treatment and prognosis networks, and some multiple-functional biomarkers like TP53 were found. Some hub biomarkers like TP53 and KRAS were convinced on the PPI networks of protein biomarkers. KEGG pathway enrichment analysis and Gene ontology was utilized to find the common pathways for CRC biomarkers, and different-functional biomarkers were mapped on the enriched pathways. Several novel multiple biomarkers like KRAS-PTEN-STAT3-CD44-ZEB1-ZEB2-S1PR1 were predicted based on the PPI network. This work guides us to make further prediction of biomarkers based on complex networks.

2.3 Paper III

Zhang X, Zhang H, Fan C-W, Shen B Sun X-F

Loss of CHGA protein as a potential biomarker for colon cancer diagnosis: a study on biomarker discovery by machine learning and confirmation by immunohistochemistry in colorectal cancer tissue microarrays. *Submitted*, 2020.

2.3.1 Background and aims

New key biomarkers are still needed for colon cancer diagnosis. Bioinformatics approach has been a new direction for biomarker discovery. As the advent of the era of big data, huge amounts of PPI data have been accumulated, which gives a foundation for studying and predicting protein biomarkers. ML has been a mature way to predict biomarker in cancer research. The aim of this study was to using ML to predict colon cancer diagnosis biomarkers based on the PPI network.

2.3.2 Materials and methods

RNA-seq data from TCGA and GTEx was used to perform DEA, and the DEGs were mapped on the human PPI from String to get the colon cancer specific PPI (CCS-PPI) network, as the foundation for biomarker prediction. Reported diagnosis biomarkers collected from the CBD database was mapped on the CCS-PPI, which provided the network features for biomarkers. The non-DEGs were also mapped to the CCS-PPI to provide the network features for non-biomarkers. Regression tree was used select the best features to further be used in SVM to predict biomarkers from the DEGs. Diagnostic ROC test performed by microarray data from GEO was utilized to make verification for the predicted results and further select final candidates. Tissue arrays was further conducted to verify the clinical value for biomarker candidate.

2.3.3 Results and discussions

12 novel colon cancer diagnosis biomarkers were predicted, of which CHGA showed the best performance in the AUC and tissue arrays. We supposed that it is a good try to use network topology features of reported biomarkers to predict new biomarkers, since it is logical and the predicted biomarker (CHGA) had good performance in the verification test in both dry and wet lab. However, there is still something needed to be further improved: 1. The negative group in the predict model was non-DEGs. If the real non-biomarkers selected from scientific experiments could be used as negative group, the credibility and even accuracy of the prediction model could be improved; 2. The number of reported biomarkers was concerned.

2.4 Paper IV

Zhang X, Zhang H, Shen B and Sun X-F

Chromogranin-A expression as a novel biomarker for early diagnosis of colon cancer patients. *Int J Mol Sci*, 20, 2919.

2.4.1 Background and aims

We predicted that CHGA could be a promising biomarker for colon cancer diagnosis in the previous work, and it is needed to make verification in more datasets. Meta-analysis is an EBM with high credibility, which could combine different results from various studies for a specific topic. The aim of this study was to use meta-analysis to verify the diagnostic value for CHGA in colon cancer, based on the GE data from GEO.

2.4.2 Materials and methods

4 datasets containing both colon cancer patients and healthy controls from GEO was selected to make logistic regression to get the 2×2 tables, which were then been used to conduct the diagnostic meta-analysis. Further biological functional and network analyses were conducted for CHGA, and CHGA related single and multiple biomarkers were predicted based on the PPI network and expression level.

2.4.3 Results and discussions

CHGA showed good diagnostic performance (sensitivity 0.89; specificity 0.89) compared with some typical biomarkers. (TP53, KRAS, and MKI67) Some CHGA related genes were predicted as future biomarkers for colon cancer. Future verification in animal model and clinical is still needed.

2.5 Paper V

Zhang X, Zhang H, Shen B and Sun X-F

Novel microRNA biomarkers for colorectal cancer early diagnosis and 5-fluorouracil chemotherapy resistance but not prognosis: a study from databases to AI-assisted verifications. *Cancers* 12(2), 341, 2020.

2.5.1 Background and aims

miRNA is a new direction for CRC biomarker discovery, and 72 miRNAs have been reported as CRC biomarkers. In our previous study, a novel network-based miRNA biomarker prediction model (MiRNA-BD) has been created and developed into software. The aim of this study was to use different source of GE data to predict new miRNA biomarkers for CRC, based on the miRNA-mRNA interaction network, via MiRNA-BD.

2.5.2 Materials and methods

The GEPIA database and the GEO database provided the DEG data from RNA-seq and microarray, separately. DEGs from these two sources together with CRC related miRNAs from miRANet were input into MiRNA-BD to construct CRC specific miRNA-mRNA network. Two topology features were used to predict novel biomarkers based on the network.

2.5.3 Results and discussions

Based on the specific network, 3 novel miRNA biomarkers (miR-10B, miR-31, and miR-30e) were predicted based on two topology features. Further meta-analysis and biological network and functional analysis verified the finding. Further clinical verification is still needed.

Acknowledgements

The earliest Chinese teacher, Confucius, once said: “time is like water, never stop”. Now I quite understand this sentence. Four years PhD study period, or even the past twenty years student period, is going to an end. Now many feelings and words are in my mind, of which, “thank you” is the first sentence I would like to express:

First and foremost, thank you, my dear supervisor, prof. Hong Zhang. I still remember the first time we met, you wore a red Nike shoes, with white hair. You are not like a traditional professor I ever met, and “cool” is the best word to describe you. I still remember the first sentence you told me: “Let’s do something interesting in the next four years”. I think we did it. Together we published four papers and now we have several manuscripts to be submitted. The most important thing you teaches me is the true love to science, and you are the one who let me decide the future direction as a scientist.

I also want to express my gratitude to my co-supervisors: prof. Xiao-Feng Sun, who helps me regain confidence in science; prof. Dirk Replibber, who gives me a lot of help in bioinformatics; prof. Bairong Shen, who guides my study in China. All of you are important in my PhD study, and I can not imagine if I can still finish my study without you.

Of course, I need to thank to my colleges and friends in Örebro, Linköping and Suzhou, especially the guys in Functional bioinformatics group, and I will never forget the good time we spent in Fika!

Finally, I want to say thank you to my parents. Thank you for all the support you give to me. Work hard and keep glad, which is our family slogan, and I will continue to follow it!

References

- 1 Vogelstein, B. & Kinzler, K. W. The multistep nature of cancer. *Trends Genet* **9**, 138-141, doi:10.1016/0168-9525(93)90209-z (1993).
- 2 Chin, L., Hahn, W. C., Getz, G. & Meyerson, M. Making sense of cancer genomic data. *Genes Dev* **25**, 534-555, doi:10.1101/gad.2017311 (2011).
- 3 Weir, B., Zhao, X. & Meyerson, M. Somatic alterations in the human cancer genome. *Cancer Cell* **6**, 433-438, doi:10.1016/j.ccr.2004.11.004 (2004).
- 4 <http://gco.iarc.fr/today/home> (Accessed 2020/03/11).
- 5 Wolf, A. M. D. *et al.* Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J Clin* **68**, 250-281, doi:10.3322/caac.21457 (2018).
- 6 Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *Lancet* **394**, 1467-1480, doi:10.1016/S0140-6736(19)32319-0 (2019).
- 7 Weitz, J. *et al.* Colorectal cancer. *Lancet* **365**, 153-165, doi:10.1016/S0140-6736(05)17706-X (2005).
- 8 <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/precision-medicine> (Accessed 2020/03/14).
- 9 Yiu, A. J. & Yiu, C. Y. Biomarkers in Colorectal Cancer. *Anticancer Res* **36**, 1093-1102 (2016).
- 10 Lin, Y. *et al.* Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Brief Bioinform* **20**, 952-975, doi:10.1093/bib/bbx158 (2019).
- 11 Lesk, A. M. Bioinformatics. *Encyclopaedia Britannica* (2013).
- 12 Yuan, X. *et al.* Network Biomarkers Constructed from Gene Expression and Protein-Protein Interaction Data for Accurate Prediction of Leukemia. *J Cancer* **8**, 278-286, doi:10.7150/jca.17302 (2017).
- 13 Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**, 3442-3444, doi:10.1093/nar/28.18.3442 (2000).
- 14 Fan, Y. *et al.* miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res* **44**, W135-141, doi:10.1093/nar/gkw288 (2016).
- 15 Lin, Y. *et al.* Biomarker microRNAs for prostate cancer metastasis: screened with a network vulnerability analysis model. *J Transl Med* **16**, 134, doi:10.1186/s12967-018-1506-7 (2018).

- 16 Qi, X., Lin, Y., Chen, J. & Shen, B. Decoding competing
endogenous RNA networks for cancer biomarker discovery. *Brief*
17 *Bioinform*, doi:10.1093/bib/bbz006 (2019).
[https://www.mayoclinic.org/diseases-conditions/colon-](https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669)
18 [cancer/symptoms-causes/syc-20353669](https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669) (accessed 2020/03/06) .
[https://www.cancer.gov/types/colorectal/patient/colon-treatment-](https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq)
19 [pdq](https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq) (Accessed 2020/03/06).
Lynch, H. T. & de la Chapelle, A. Hereditary colorectal cancer. *N*
20 *Engl J Med* 348, 919-932, doi:10.1056/NEJMra012242 (2003).
Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet*
21 383, 1490-1502, doi:10.1016/S0140-6736(13)61649-9 (2014).
Siegel, R. L. *et al.* Colorectal cancer statistics, 2017. *CA Cancer J*
22 *Clin* 67, 177-193, doi:10.3322/caac.21395 (2017).
[https://www.cancer.gov/types/colorectal/patient/colon-treatment-](https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq#_112)
23 [pdq# 112](https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq#_112) (Accessed 2020/03/16).
Zhang, X., Zhang, H., Shen, B. & Sun, X. F. Chromogranin-A
Expression as a Novel Biomarker for Early Diagnosis of Colon
Cancer Patients. *Int J Mol Sci* 20, doi:10.3390/ijms20122919
(2019).
24 Newton, K. F., Newman, W. & Hill, J. Review of biomarkers in
colorectal cancer. *Colorectal Dis* 14, 3-17, doi:10.1111/j.1463-
1318.2010.02439.x (2012).
25 Cecil, T. D., Sexton, R., Moran, B. J. & Heald, R. J. Total
mesorectal excision results in low local recurrence rates in lymph
node-positive rectal cancer. *Dis Colon Rectum* 47, 1145-1149;
discussion 1149-1150, doi:10.1007/s10350-004-0086-6 (2004).
26 Simunovic, M., Sexton, R., Rempel, E., Moran, B. J. & Heald, R.
J. Optimal preoperative assessment and surgery for rectal cancer
may greatly limit the need for radiotherapy. *Br J Surg* 90, 999-
1003, doi:10.1002/bjs.4210 (2003).
27 Glehen, O. *et al.* Intraperitoneal chemohyperthermia and
attempted cytoreductive surgery in patients with peritoneal
carcinomatosis of colorectal origin. *Br J Surg* 91, 747-754,
doi:10.1002/bjs.4473 (2004).
28 Figueredo, A., Charette, M. L., Maroun, J., Brouwers, M. C. &
Zuraw, L. Adjuvant therapy for stage II colon cancer: a systematic
review from the Cancer Care Ontario Program in evidence-based
care's gastrointestinal cancer disease site group. *J Clin Oncol* 22,
3395-3407, doi:10.1200/JCO.2004.03.087 (2004).
29 Gunderson, L. L., Haddock, M. G. & Schild, S. E. Rectal cancer:
Preoperative versus postoperative irradiation as a component of
adjuvant treatment. *Semin Radiat Oncol* 13, 419-432,
doi:10.1016/S1053-4296(03)00073-0 (2003).

- 30 Higgins, G. A., Jr. *et al.* Preoperative radiotherapy for colorectal
cancer. *Ann Surg* **181**, 624-631, doi:10.1097/00000658-
197505000-00017 (1975).
- 31 <https://www.bcmj.org/articles/radiotherapy-colorectal-cancer>
(Accessed 2020/04/14).
- 32 Benson, A. B., 3rd *et al.* American Society of Clinical Oncology
recommendations on adjuvant chemotherapy for stage II colon
cancer. *J Clin Oncol* **22**, 3408-3419,
doi:10.1200/JCO.2004.05.063 (2004).
- 33 Wei, Y., Yang, P., Cao, S. & Zhao, L. The combination of
curcumin and 5-fluorouracil in cancer therapy. *Arch Pharm Res*
41, 1-13, doi:10.1007/s12272-017-0979-x (2018).
- 34 <http://www.merriam-webster.com/dictionary/fluorouracil>
(Accessed 2020/03/11).
- 35 <https://www.drugs.com/monograph/fluorouracil.html> (Accessed
2020/03/11).
- 36 <https://www.sciencedaily.com/releases/2008/04/080421191425.htm>
(Accessed 2020/03/11).
- 37 Bosset, J. F. *et al.* Chemotherapy with preoperative radiotherapy
in rectal cancer. *N Engl J Med* **355**, 1114-1123,
doi:10.1056/NEJMoa060829 (2006).
- 38 Ganesh, K. *et al.* Immunotherapy in colorectal cancer: rationale,
challenges and potential. *Nat Rev Gastroenterol Hepatol* **16**, 361-
375, doi:10.1038/s41575-019-0126-x (2019).
- 39 [https://www.cancerresearch.org/immunotherapy/cancer-](https://www.cancerresearch.org/immunotherapy/cancer-types/colorectal-cancer)
[types/colorectal-cancer](https://www.cancerresearch.org/immunotherapy/cancer-types/colorectal-cancer) (Accessed 2020/04/14).
- 40 [https://www.cancer.gov/publications/dictionaries/cancer-](https://www.cancer.gov/publications/dictionaries/cancer-terms/def/targeted-therapy?redirect=true)
[terms/def/targeted-therapy?redirect=true](https://www.cancer.gov/publications/dictionaries/cancer-terms/def/targeted-therapy?redirect=true) (Accessed 2020/03/14).
- 41 [https://www.targetedonc.com/publications/targeted-therapy-](https://www.targetedonc.com/publications/targeted-therapy-news/2017/november-2017/guidelines-consider-use-of-targeted-therapies-in-colorectal-cancer)
[news/2017/november-2017/guidelines-consider-use-of-targeted-](https://www.targetedonc.com/publications/targeted-therapy-news/2017/november-2017/guidelines-consider-use-of-targeted-therapies-in-colorectal-cancer)
[therapies-in-colorectal-cancer](https://www.targetedonc.com/publications/targeted-therapy-news/2017/november-2017/guidelines-consider-use-of-targeted-therapies-in-colorectal-cancer) (Accessed 2020/04/14)
- 42 Siegel, R. *et al.* Cancer treatment and survivorship statistics, 2012.
CA Cancer J Clin **62**, 220-241, doi:10.3322/caac.21149 (2012).
- 43 Brenner, H. *et al.* Progress in colorectal cancer survival in Europe
from the late 1980s to the early 21st century: the EURO CARE
study. *Int J Cancer* **131**, 1649-1658, doi:10.1002/ijc.26192
(2012).
- 44 Sankaranarayanan, R. *et al.* Cancer survival in Africa, Asia, and
Central America: a population-based study. *Lancet Oncol* **11**,
165-173, doi:10.1016/S1470-2045(09)70335-3 (2010).
- 45 Gill, S. *et al.* Pooled analysis of fluorouracil-based adjuvant
therapy for stage II and III colon cancer: who benefits and by how

- much? *J Clin Oncol* **22**, 1797-1806, doi:10.1200/JCO.2004.09.059 (2004).
- 46 Greene, F. L., Stewart, A. K. & Norton, H. J. A new TNM staging strategy for node-positive (stage III) colon cancer: an analysis of 50,042 patients. *Ann Surg* **236**, 416-421; discussion 421, doi:10.1097/00000658-200210000-00003 (2002).
- 47 Greene, F. L., Stewart, A. K. & Norton, H. J. New tumor-node-metastasis staging strategy for node-positive (stage III) rectal cancer: an analysis. *J Clin Oncol* **22**, 1778-1784, doi:10.1200/JCO.2004.07.015 (2004).
- 48 Gunderson, L. L. *et al.* Impact of T and N stage and treatment on survival and relapse in adjuvant rectal cancer: a pooled analysis. *J Clin Oncol* **22**, 1785-1796, doi:10.1200/JCO.2004.08.173 (2004).
- 49 Staib, L., Link, K. H., Blatz, A. & Beger, H. G. Surgery of colorectal cancer: surgical morbidity and five- and ten-year results in 2400 patients--monoinstitutional experience. *World J Surg* **26**, 59-66, doi:10.1007/s00268-001-0182-5 (2002).
- 50 Zhang, X. *et al.* CBD: a biomarker database for colorectal cancer. *Database (Oxford)* **2018**, doi:10.1093/database/bay046 (2018).
- 51 <https://www.abcam.com/kits/micrnas-as-biomarkers-in-cancer> (Accessed 2020/04/01).
- 52 Mehra, M. & Chauhan, R. Long Noncoding RNAs as a Key Player in Hepatocellular Carcinoma. *Biomark Cancer* **9**, 1179299X17737301, doi:10.1177/1179299X17737301 (2017).
- 53 Lan, H., Lu, H., Wang, X. & Jin, H. MicroRNAs as potential biomarkers in cancer: opportunities and challenges. *Biomed Res Int* **2015**, 125094, doi:10.1155/2015/125094 (2015).
- 54 Bouzourene, H. *et al.* p53 and Ki-ras as prognostic factors for Dukes' stage B colorectal cancer. *Eur J Cancer* **36**, 1008-1015, doi:10.1016/s0959-8049(00)00036-8 (2000).
- 55 Takeda, A. *et al.* Serum p53 antibody as a useful marker for monitoring of treatment of superficial colorectal adenocarcinoma after endoscopic resection. *Int J Clin Oncol* **6**, 45-49, doi:10.1007/pl00012079 (2001).
- 56 Adell, G. *et al.* p53 status: an indicator for the effect of preoperative radiotherapy of rectal cancer. *Radiother Oncol* **51**, 169-174, doi:10.1016/s0167-8140(99)00041-9 (1999).
- 57 Vacante, M., Borzi, A. M., Basile, F. & Biondi, A. Biomarkers in colorectal cancer: Current clinical utility and future perspectives. *World J Clin Cases* **6**, 869-881, doi:10.12998/wjcc.v6.i15.869 (2018).

- 58 Imperiale, T. F., Ransohoff, D. F. & Itzkowitz, S. H. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med* 371, 187-188, doi:10.1056/NEJMc1405215 (2014).
- 59 Shaw, J. A. & Stebbing, J. Circulating free DNA in the management of breast cancer. *Ann Transl Med* 2, 3, doi:10.3978/j.issn.2305-5839.2013.06.06 (2014).
- 60 Zarkavelis, G. *et al.* Current and future biomarkers in colorectal cancer. *Ann Gastroenterol* 30, 613-621, doi:10.20524/aog.2017.0191 (2017).
- 61 Iyer, R. R., Pluciennik, A., Burdett, V. & Modrich, P. L. DNA mismatch repair: functions and mechanisms. *Chem Rev* 106, 302-323, doi:10.1021/cr0404794 (2006).
- 62 <https://www.ajmc.com/newsroom/biomarker-use-in-colorectal-cancer> (Accessed 2020/04/19).
- 63 Kumarswamy, R., Volkmann, I. & Thum, T. Regulation and function of miRNA-21 in health and disease. *RNA Biol* 8, 706-713, doi:10.4161/rna.8.5.16154 (2011).
- 64 Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* 294, 853-858, doi:10.1126/science.1064921 (2001).
- 65 Moridikia, A., Mirzaei, H., Sahebkar, A. & Salimian, J. MicroRNAs: Potential candidates for diagnosis and treatment of colorectal cancer. *J Cell Physiol* 233, 901-913, doi:10.1002/jcp.25801 (2018).
- 66 Peng, Q. *et al.* The clinical role of microRNA-21 as a promising biomarker in the diagnosis and prognosis of colorectal cancer: a systematic review and meta-analysis. *Oncotarget* 8, 44893-44909, doi:10.18632/oncotarget.16488 (2017).
- 67 Zhang, Y. *et al.* Circular RNAs: emerging cancer biomarkers and targets. *J Exp Clin Cancer Res* 36, 152, doi:10.1186/s13046-017-0624-z (2017).
- 68 Liu, Q. *et al.* Circular RNA Related to the Chondrocyte ECM Regulates MMP13 Expression by Functioning as a MiR-136 'Sponge' in Human Cartilage Degradation. *Sci Rep* 6, 22572, doi:10.1038/srep22572 (2016).
- 69 Xu, H., Guo, S., Li, W. & Yu, P. The circular RNA Cdr1as, via miR-7 and its targets, regulates insulin transcription and secretion in islet cells. *Sci Rep* 5, 12453, doi:10.1038/srep12453 (2015).
- 70 Zhao, Y., Alexandrov, P. N., Jaber, V. & Lukiw, W. J. Deficiency in the Ubiquitin Conjugating Enzyme UBE2A in Alzheimer's Disease (AD) is Linked to Deficits in a Natural Circular miRNA-7 Sponge (circRNA; ciRS-7). *Genes (Basel)* 7, doi:10.3390/genes7120116 (2016).

- 71 Wang, K. *et al.* A circular RNA protects the heart from
pathological hypertrophy and heart failure by targeting miR-223.
Eur Heart J **37**, 2602-2611, doi:10.1093/eurheartj/ehv713 (2016).
- 72 Li, J. *et al.* Circular RNAs in cancer: novel insights into origins,
properties, functions and implications. *Am J Cancer Res* **5**, 472-
480 (2015).
- 73 Guarnerio, J. *et al.* Oncogenic Role of Fusion-circRNAs Derived
from Cancer-Associated Chromosomal Translocations. *Cell* **165**,
289-302, doi:10.1016/j.cell.2016.03.020 (2016).
- 74 Zheng, Q. *et al.* Circular RNA profiling reveals an abundant
circHIPK3 that regulates cell growth by sponging multiple
miRNAs. *Nat Commun* **7**, 11215, doi:10.1038/ncomms11215
(2016).
- 75 Xia, S. *et al.* Comprehensive characterization of tissue-specific
circular RNAs in the human and mouse genomes. *Brief Bioinform*
18, 984-992, doi:10.1093/bib/bbw081 (2017).
- 76 Suzuki, H. & Tsukahara, T. A view of pre-mRNA splicing from
RNase R resistant RNAs. *Int J Mol Sci* **15**, 9331-9342,
doi:10.3390/ijms15069331 (2014).
- 77 Barrett, S. P. & Salzman, J. Circular RNAs: analysis, expression
and potential functions. *Development* **143**, 1838-1847,
doi:10.1242/dev.128074 (2016).
- 78 Chen, J. The Cell-Cycle Arrest and Apoptotic Functions of p53 in
Tumor Initiation and Progression. *Cold Spring Harb Perspect*
Med **6**, a026104, doi:10.1101/cshperspect.a026104 (2016).
- 79 Bodmer, W. F. *et al.* Genetic analysis of colorectal cancer.
Princess Takamatsu Symp **20**, 49-59 (1989).
- 80 Wanebo, H. J. *et al.* Meeting the biologic challenge of colorectal
metastases. *Clin Exp Metastasis* **29**, 821-839,
doi:10.1007/s10585-012-9517-x (2012).
- 81 Rodrigues, N. R. *et al.* p53 mutations in colorectal cancer. *Proc*
Natl Acad Sci U S A **87**, 7555-7559,
doi:10.1073/pnas.87.19.7555 (1990).
- 82 Levine, A. J., Reich, N. & Thomas, R. The regulation of a cellular
protein, p53, in normal and transformed cells. *Prog Clin Biol Res*
119, 159-169 (1983).
- 83 Rotter, V. & Wolf, D. Biological and molecular analysis of p53
cellular-encoded tumor antigen. *Adv Cancer Res* **43**, 113-141,
doi:10.1016/s0065-230x(08)60944-6 (1985).
- 84 Levine, A. J. p53, the cellular gatekeeper for growth and division.
Cell **88**, 323-331, doi:10.1016/s0092-8674(00)81871-1 (1997).
- 85 Gottlieb, T. M. & Oren, M. p53 and apoptosis. *Semin Cancer*
Biol **8**, 359-368, doi:10.1006/scbi.1998.0098 (1998).

- 86 Sabapathy, K. & Lane, D. P. Therapeutic targeting of p53: all mutants are equal, but some mutants are more equal than others. *Nat Rev Clin Oncol* **15**, 13-30, doi:10.1038/nrclinonc.2017.151 (2018).
- 87 <http://sysbio.suda.edu.cn/CBD/inf.php?id=p53> (Accessed 2020/03/31).
- 88 Yamaguchi, A. *et al.* Expression of p53 protein in colorectal cancer and its relationship to short-term prognosis. *Cancer* **70**, 2778-2784, doi:10.1002/1097-0142(19921215)70:12<2778::aid-cncr2820701209>3.0.co;2-1 (1992).
- 89 Karapetis, C. S. *et al.* K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* **359**, 1757-1765, doi:10.1056/NEJMoa0804385 (2008).
- 90 Yang, Y. *et al.* Independent Correlation Between Ki67 Index and Circulating Tumor Cells in the Diagnosis of Colorectal Cancer. *Anticancer Res* **37**, 4693-4700, doi:10.21873/anticancer.11874 (2017).
- 91 Zhang, Y., Zhang, Y. & Zhang, L. Expression of cancer-testis antigens in esophageal cancer and their progress in immunotherapy. *J Cancer Res Clin Oncol* **145**, 281-291, doi:10.1007/s00432-019-02840-3 (2019).
- 92 Davey, P., Arnott, S. J. & Sturgeon, C. M. Carcinoembryonic antigen as a prognostic indicator in the radiotherapeutic management of rectal cancer. *Eur J Surg Oncol* **13**, 17-20 (1987).
- 93 Su, B. B., Shi, H. & Wan, J. Role of serum carcinoembryonic antigen in the detection of colorectal cancer before and after surgical resection. *World J Gastroenterol* **18**, 2121-2126, doi:10.3748/wjg.v18.i17.2121 (2012).
- 94 Yang, S. & Chung, H. C. Novel biomarker candidates for gastric cancer. *Oncol Rep* **19**, 675-680 (2008).
- 95 Ma, Z. *et al.* Clinical significance of polymorphism and expression of chromogranin a and endothelin-1 in prostate cancer. *J Urol* **184**, 1182-1188, doi:10.1016/j.juro.2010.04.063 (2010).
- 96 Weisbrod, A. B. *et al.* Altered PTEN, ATRX, CHGA, CHGB, and TP53 expression are associated with aggressive VHL-associated pancreatic neuroendocrine tumors. *Horm Cancer* **4**, 165-175, doi:10.1007/s12672-013-0134-1 (2013).
- 97 <https://www.genecards.org/cgi-bin/carddisp.pl?gene=ABI2> (Accessed 2020/03/17).
- 98 Dai, Z. & Pendergast, A. M. Abi-2, a novel SH3-containing protein interacts with the c-Abl tyrosine kinase and modulates c-

- Abl transforming activity. *Genes Dev* **9**, 2569-2582, doi:10.1101/gad.9.21.2569 (1995).
- 99 Meng, W.-J. *Expression of miR-302a, miR-105 and miR-888 Plays Several Critical Roles in Pathogenesis, Radiotherapy Response and Prognosis in Rectal Cancer Patients: A Study from Real-time PCR to Big Data Analyses.*
- 100 Smith, J. J., Sorensen, A. G. & Thrall, J. H. Biomarkers in imaging: realizing radiology's future. *Radiology* **227**, 633-638, doi:10.1148/radiol.2273020518 (2003).
- 101 <https://www.cancer.gov/news-events/cancer-currents-blog/2020/artificial-intelligence-brain-tumor-diagnosis-surgery> (Accessed 2020/03/08).
- 102 Ni, Y., Xie, G. & Jia, W. Metabonomics of human colorectal cancer: new approaches for early diagnosis and biomarker discovery. *J Proteome Res* **13**, 3857-3870, doi:10.1021/pr500443c (2014).
- 103 Zhang, X., Sun, X. F., Shen, B. & Zhang, H. Potential Applications of DNA, RNA and Protein Biomarkers in Diagnosis, Therapy and Prognosis for Colorectal Cancer: A Study from Databases to AI-Assisted Verification. *Cancers (Basel)* **11**, doi:10.3390/cancers11020172 (2019).
- 104 Goetz, L. H. & Schork, N. J. Personalized medicine: motivation, challenges, and progress. *Fertil Steril* **109**, 952-963, doi:10.1016/j.fertnstert.2018.05.006 (2018).
- 105 Huh, J. W., Kim, H. R. & Kim, Y. J. Prognostic role of p53 messenger ribonucleic acid expression in patients after curative resection for stage I to III colorectal cancer: association with colon cancer stem cell markers. *J Am Coll Surg* **216**, 1063-1069, doi:10.1016/j.jamcollsurg.2013.01.058 (2013).
- 106 Nakanishi, R. *et al.* Prognostic relevance of KRAS and BRAF mutations in Japanese patients with colorectal cancer. *Int J Clin Oncol* **18**, 1042-1048, doi:10.1007/s10147-012-0501-x (2013).
- 107 Zeng, Z. S., Huang, Y., Cohen, A. M. & Guillem, J. G. Prediction of colorectal cancer relapse and survival via tissue RNA levels of matrix metalloproteinase-9. *J Clin Oncol* **14**, 3133-3140, doi:10.1200/JCO.1996.14.12.3133 (1996).
- 108 Tomoda, H., Kakeji, Y. & Furusawa, M. Prognostic significance of flow cytometric analysis of DNA content in colorectal cancer: a prospective study. *J Surg Oncol* **53**, 144-148, doi:10.1002/jso.2930530303 (1993).
- 109 Rodel, F. *et al.* High survivin expression is associated with reduced apoptosis in rectal cancer and may predict disease-free survival after preoperative radiochemotherapy and surgical

- resection. *Strahlenther Onkol* **178**, 426-435, doi:10.1007/s00066-002-1003-y (2002).
- 110 Boulay, J. L. *et al.* SMAD7 is a prognostic marker in patients with colorectal cancer. *Int J Cancer* **104**, 446-449, doi:10.1002/ijc.10908 (2003).
- 111 Hisada, Y. & Mackman, N. Cancer-associated pathways and biomarkers of venous thrombosis. *Blood* **130**, 1499-1506, doi:10.1182/blood-2017-03-743211 (2017).
- 112 Wang, H., Li, X., Zhou, D. & Huang, J. Autoantibodies as biomarkers for colorectal cancer: A systematic review, meta-analysis, and bioinformatics analysis. *Int J Biol Markers* **34**, 334-347, doi:10.1177/1724600819880906 (2019).
- 113 Rotte, A. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. *J Exp Clin Cancer Res* **38**, 255, doi:10.1186/s13046-019-1259-z (2019).
- 114 https://fac.ksu.edu.sa/sites/default/files/10_biomarker_detection_tech_niques_0.pdf (Accessed 2020/03/09).
- 115 Baumgartner, C., Osl, M., Netzer, M. & Baumgartner, D. Bioinformatic-driven search for metabolic biomarkers in disease. *J Clin Bioinforma* **1**, 2, doi:10.1186/2043-9113-1-2 (2011). Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212-1226, doi:10.1016/j.cell.2014.10.050 (2014).
- 117 Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601, doi:10.1126/science.1257601 (2015).
- 118 Goldman, M. *et al.* The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. 326470, doi:10.1101/326470 %J bioRxiv (2019).
- 119 Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* **45**, W98-W102, doi:10.1093/nar/gkx247 (2017).
- 120 Tang, Z., Kang, B., Li, C., Chen, T. & Zhang, Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* **47**, W556-W560, doi:10.1093/nar/gkz430 (2019).
- 121 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).
- 122 Fan, Y. & Xia, J. miRNet-Functional Analysis and Visual Exploration of miRNA-Target Interactions in a Network Context.

- Methods Mol Biol* **1819**, 215-233, doi:10.1007/978-1-4939-8618-7_10 (2018).
- 123 <https://gobiomdb.com/> (Accessed 2020/04/03).
- 124 <https://edrn.nci.nih.gov/> (Accessed 2020/04/03).
- 125 Lobdell, D. T. & Mendola, P. Development of a biomarkers database for the National Children's Study. *Toxicol Appl Pharmacol* **206**, 269-273, doi:10.1016/j.taap.2004.07.016 (2005).
- 126 Yerlikaya, S., Broger, T., MacLean, E., Pai, M. & Denking, C. M. A tuberculosis biomarker database: the key to novel TB diagnostics. *Int J Infect Dis* **56**, 253-257, doi:10.1016/j.ijid.2017.01.025 (2017).
- 127 Yang, I. S. *et al.* IDBD: infectious disease biomarker database. *Nucleic Acids Res* **36**, D455-460, doi:10.1093/nar/gkm925 (2008).
- 128 Dingerdissen, H. M. *et al.* OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data. *JCO Clin Cancer Inform* **4**, 210-220, doi:10.1200/CCI.19.00117 (2020).
- 129 Lee, B. T. *et al.* Gastric Cancer (Biomarkers) Knowledgebase (GCBKB): A Curated and Fully Integrated Knowledgebase of Putative Biomarkers Related to Gastric Cancer. *Biomark Insights* **1**, 135-141 (2007).
- 130 Dai, H. J. *et al.* LiverCancerMarkerRIF: a liver cancer biomarker interactive curation system combining text mining and expert annotations. *Database (Oxford)* **2014**, doi:10.1093/database/bau085 (2014).
- 131 Wai-Kai, C. Graph Theory and its Engineering Applications. *World Scientific* pp. 29 (1997).
- 132 <http://www.patternsinnature.org/Book/RandomNetworks.html> (Accessed 2020/03/15).
- 133 Cho, D. Y., Kim, Y. A. & Przytycka, T. M. Chapter 5: Network biology approach to complex diseases. *PLoS Comput Biol* **8**, e1002820, doi:10.1371/journal.pcbi.1002820 (2012).
- 134 Lin, Y., Wu, W., Sun, Z., Shen, L. & Shen, B. MiRNA-BD: an evidence-based bioinformatics model and software tool for microRNA biomarker discovery. *RNA Biol* **15**, 1093-1105, doi:10.1080/15476286.2018.1502590 (2018).
- 135 <https://www.cancer.gov/about-nci/budget/plan/artificial-intelligence> (Accessed 2020/03/08).
- 136 <https://datascience.cancer.gov/collaborations/atom> (Accessed 2020/03/09).

- 137 Yerukala Sathipati, S. & Ho, S. Y. Identifying a miRNA signature
for predicting the stage of breast cancer. *Sci Rep* 8, 16138,
doi:10.1038/s41598-018-34604-3 (2018).
- 138 Wang, H. Predicting MicroRNA Biomarkers for Cancer Using
Phylogenetic Tree and Microarray Analysis. *Int J Mol Sci* 17,
doi:10.3390/ijms17050773 (2016).
- 139 [https://www.cancer.gov/news-events/cancer-currents-
blog/2018/artificial-intelligence-lung-cancer-classification](https://www.cancer.gov/news-events/cancer-currents-
blog/2018/artificial-intelligence-lung-cancer-classification)
(Accessed 2020/03/09).
- 140 www.ubuntupit.com > ML & AI (Accessed 2020/03/12).
- 141 Evidence-Based Medicine Working, G. Evidence-based medicine.
A new approach to teaching the practice of medicine. *JAMA* 268,
2420-2425, doi:10.1001/jama.1992.03490170092032 (1992).