frontiers
in Cell and Developmental Biology

PERSPECTIVE
published: 07 April 2021
doi: 10.3389/fcell.2021.631011

Topic Evolution Analysis for Omics Data Integration in Cancers

Li Ning[1,2] *and He Huixin*[3]*

[1] Business School of Huaqiao University, Quan Zhou, China, [2] Management Science and Engineering Department, Management School, Xiamen University, Xiamen, China, [3] Computer Science and Engineering Department, Computer Science and Engineering School, Huaqiao University, Quanzhou, China

One of the vital challenges for cancer diseases is efficient biomarkers monitoring formation and development are limited. Omics data integration plays a crucial role in the mining of biomarkers in the human condition. As the link between omics study on biomarkers discovery and cancer diseases is deepened, defining the principal technologies applied in the field is a must not only for the current period but also for the future. We utilize topic modeling to extract topics (or themes) as a probabilistic distribution of latent topics from the dataset. To predict the future trend of related cases, we utilize the Prophet neural network to perform a prediction correction model for existing topics. A total of 2,318 pieces of literature (from 2006 to 2020) were retrieved from MEDLINE with the query on "omics" and "cancer." Our study found 20 topics covering current research types. The topic extraction results indicate that, with the rapid development of omics data integration research, multi-omics analysis (Topic 11) and genomics of colorectal cancer (Topic 10) have more studies reported last 15 years. From the topic prediction view, research findings in multi-omics data processing and novel biomarker discovery for cancer prediction (Topic 2, 3, 10, 11) will be heavily focused in the future. From the topic visuallization and evolution trends, metabolomics of breast cancer (Topic 9), pharmacogenomics (Topic 15), genome-guided therapy regimens (Topic 16), and microRNAs target genes (Topic 17) could have more rapidly developed in the study of cancer treatment effect and recurrence prediction.

Keywords: omics, cancers, topic modeling, prophet neural network, evolution trend

OPEN ACCESS

Edited by:
Mingxiang Teng,
Moffitt Cancer Center, United States

Reviewed by:
Angela Di Giannatale,
Bambino Gesù Children Hospital (IRCCS), Italy
Donghui Yang,
Southeast University, China

***Correspondence:**
He Huixin
hehuixin@hqu.edu.cn

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *19 November 2020*
Accepted: *04 February 2021*
Published: *07 April 2021*

Citation:
Ning L and Huixin H (2021) Topic Evolution Analysis for Omics Data Integration in Cancers. Front. Cell Dev. Biol. 9:631011. doi: 10.3389/fcell.2021.631011

INTRODUCTION

Genomics, proteomics, metabolomics, transcriptomics, and other -omics studies involve comprehensive investigations (Saito et al., 2013). In recent years, advances in high-throughput technology have shown promise for discovering biomarkers (Njoku et al., 2020). Biomarkers are useful tools as indicators/predictors of disease severity and drug reactivity, and thus, are expected to be used for diagnostic or prognostic purposes for all different types of complex diseases. With the discovery and identification of HRAS and TP53, more proto-oncogenes, tumor suppressor genes, and susceptibility genes have been discovered (Hanahan and Weinberg, 2000, 2011). The essential characteristics of tumor cells have been elucidated at the molecular level. Combined with clinical information, the molecular mechanism and evolutionary dynamics of tumor development and cell heterogeneity can be observed (Urh and Kunej, 2016). Omics data integration and machine learning algorithm can be utilized to improve the predicting accuracy of familial tumor patients.

Meanwhile, biomarker discovery technology can also improve the sensitivity and accuracy of early diagnosis and provide more accurate molecular staging (Long et al., 2019). The combination of proteomics and metabolomics can reduce adverse reactions of targeted drugs and chemotherapy drugs (Ristori et al., 2020). Pharmacogenomics can be applied to detect new therapeutic targets and develop new drugs, and it can also turn old drugs into treasures (Shukla, 2017). Essential marks of metabolomics are biomarker development and its translation to the clinic that can do a favor to personalized diagnosis and deepen the understanding of disease pathogenesis (Nazifova-Tasinova et al., 2020). Despite the rapid development of omics data integration toward mining biomarkers in the academic medical world, few studies explored statistical relationships between literature text terms and their time-series features. Here, we present a topic modeling and predictive analysis for omics study in cancer diseases.

In recent years, scientific production on omics data integration toward the mining of phenotype biomarkers have produced datasets of significant extreme interests and has expanded the physiology field of cell and developmental biology concepts. Research in this area is indeed kinds of. But how is this bunch of research evolved? Which directions are more valuable for future development? Unfortunately, few studies explore the underlying relationships in existing reports among title, abstract, and keywords. Practically, two methods, natural language or bibliometric analyses, can achieve this goal. For one thing, employing natural language processing cuts sentences into metadata; for another thing, using the bibliometric method statistically analyzes metadata based on the time and frequency of occurrence. Intuitively, combining the two would possibly make a valuable prediction of potential hot spots from the technical perspective. An acceptable way to accomplish that goal is to evaluate the frequency of emerged scientific terms and how the same words are aggregated in research. Furthermore, Latent Dirichlet Allocation (LDA, henceforth), as a statistical technique, is available to capture and explain potential relationships between the high-frequently-used terms in recent scientific products with high precision through a layered aggregation system of words Hence, compared with traditional bibliometric research paper, this article is characterized by natural language processing of many title texts and predictive analysis based on time series. From this perspective, we summarize the scientific terms presented on MEDLINE over the past 15 years. Research topics are extracted, and then theme intensity is calculated based on the time segment of the month. From the perspective of topic modeling and the predictive correction algorithm optimization, our study provides specific methods for scientific researchers in the cell development field. It furnishes more followers with intuitive understanding and disciplinary analysis in the study of omics data integration to biomarkers.

METHODS

This study evaluates the "Omics data integration" in "Cancers" to explore the research topic's evolution. The research framework contains seven steps, which are mainly embodied in two stages. The first phase is data preprocessing and topic extractions (Ning et al., 2020). Since the intensity of each subject has time-series features, so the second phase is the prediction analysis and display of the topic trend. **Figure 1** shows the research framework in Panel (A).

Dataset for our study includes titles, abstracts, and keywords from publications. Data sources come from the Web of Knowledge-MEDLINE database. We used the following keywords to extract literature: omics [TS] AND cancer [TS] AND "2006/01/01 [DATE]: "2020/12/31 [DATE]". 2,801 pieces of literature were obtained. Information retrieved includes title, author, abstract, keywords, references, and journal sources. We then filtered the literature to preserve only journal articles for downstream analysis, leaving 2,318 unduplicated items. All of the observed article titles, abstracts, and keywords were further processed by natural language routine.

Topic Extraction and Topic Intensity

LDA was first introduced by Blei's study in 2003 (Hofmann, 1999; Blei et al., 2003). Scholars in cell and developmental biology have applied the LDA model to identify scientific research topics (Li et al., 2015; Valle et al., 2020). Besides, perplexity is considered as a standard tool to evaluate the effectiveness of various natural language processing models (Rosen-Zvi et al., 2010). The lower the perplexity, the better the fitting effect of the training topic distribution model to the training set data. Meanwhile, the perplexity would also decrease gradually along with the increase of the topic number. Topic coherence tends to stabilize after reaching the optimal level. When N sets to be 20, the title and the abstract topic get their optimum and stabilize. Therefore, we set the number of topics N to 20. **Figure 1** shows the evaluation of the number of topics in Panel (B). **Table 1** lists the 20 topics and their high-frequency keywords.

LDAvis an interactive dynamic visualization tool for the LDA model. We use this tool to visualize the results of topic distribution. In Panel (C), the number of circles represents the number of topics, and its size means the corpus belonging to the topic. The quadrants of circle distribution represent the clustering situation. The distance between circles represents the semantic distance between topics. The farther space is, the higher the discrimination between topics is.

Topic intensity is a statistical attribute of topic, indicating the degree of concern of the topic. In our research, we employed the number of documents distributed to each topic to calculate intensity. The intensity index of each topic constitutes a series of time. Define in e time slice, in the document set, the number of documents is Ne, and the intensity of topic X, T_X means the number of articles attributed to topic X:

$$Intensity\_T_X = \sum_{d \in N_e} \theta_{dX}$$

Topic Trend Prediction

In our study, we use the time series model, the Prophet neural network, to predict the topic trend. Compared with the classical ARIMA model and LSTM model, the Prophet model can better predict the growth trend. It does not require extensive sample data for text training, so it is easier to achieve convergence than the LSTM method. Since the minimum unit for the journal

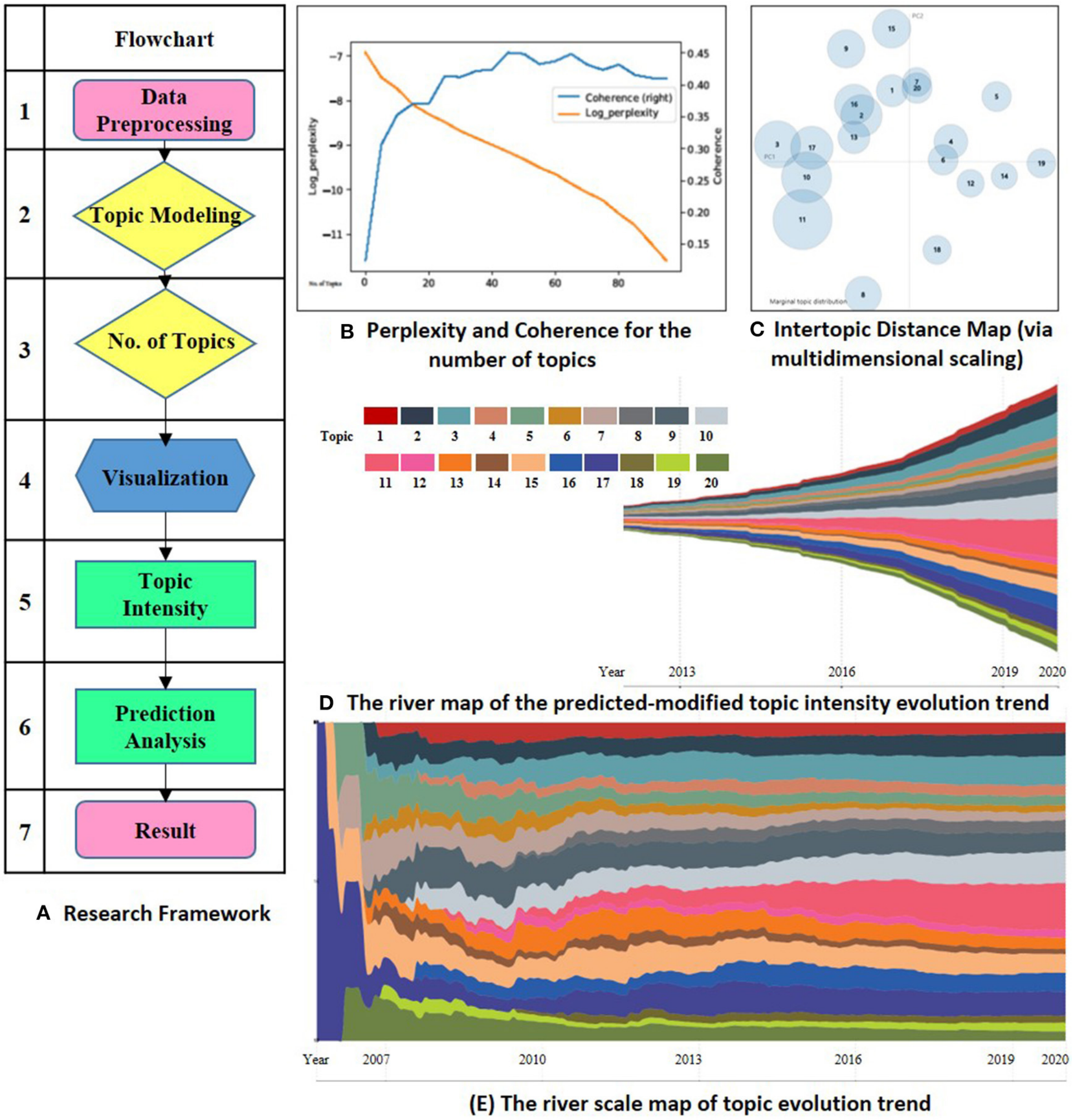

FIGURE 1 | (A) is the research framework. **(B)** shows the perplexity and coherence for the number of topics. **(C)** visualizes the Intertopic distance for the 20 topics. **(D,E)** is the river map and river scale map of the predicted-modified evolution. Topic 1 biomarker discovery in early diagnostics, 2 machine learning for cancer prediction, 3 novel biomarker identify technology, 4 multi-omics in Hepatocellular Carcinoma, 5 genomics in clinical application, 6 genomics in tumor heterogeneity discovery, 7 proteomics in post-translational modification, 8 transcriptional and metabolic processes, 9 metabolomics of breast cancer, 10 genomics of colorectal cancer, 11 multi-omics analysis, 12 cancer vaccines, 13 tumor immunotherapy, 14 metabolomics in prostate cancer, 15 pharmacogenomics, 16 genome-guided therapy regimens, 17 non-coding RNA target genes, 18 microbial metabolomics, 19 metagenomics, 20 genomics of colorectal cancer.

publication cycle in the data sample is monthly. When predicting the size trend of each topic, we take it monthly, ignoring the influence of cycle factors and holiday factors.

R2_score represents the deviation between the observed value and the real value, ranging from 0 to 1. This is used to evaluate the quality of the trend predictions. The case that R2_score is closer to 1 means the fitting effect is ideal. Besides, we compare Prophet-Liner and Prophet-Logistic results to examine whether the topic trend is linear or not. Intuitively, the Prophet-Liner result represents the direct prediction on linear trend, while the other is the direct prediction on logistic trend.

The ECharts ThemeRriver diagrams are used in the series to represent changes in events or topics over time. The different colored river branches in the theme river encode various topics. The width of the river branch encodes the value in the original dataset. **Figure 1** shows the evolution trend for omics data integration in cancer research in Panel (D).

TABLE 1 | Results for topic extraction and topic prediction.

	Topic name	Keywords	R2_score (Prophet-logistic, Prophet-liner)	Topic %	Predicted %	Up/Down
Topic 1	Biomarker discovery in early diagnostics	biomarker, discovery, diagnostics, omics	(0.9542, 0.9982)	3.70	3.21	↓
Topic 2	Machine learning for cancer prediction	prediction, classification, learning, information	(0.9774, 0.9985)	6.40	7.10	↑
Topic 3	Novel biomarker identify technology	profiling, expression, novel, signature	(0.9821, 0.9994)	7.90	9.14	↑
Topic 4	Multi-omics in Hepatocellular Carcinoma	carcinoma, protein, hepatocellular, genomics	(0.9530, 0.9980)	4.20	3.39	↓
Topic 5	Genomics in clinical application	clinical, application, genomics, functional	(0.9434, 0.9956)	3.20	2.81	↓
Topic 6	Genomics in tumor heterogeneity discovery	single, inhibitor, plasma, heterogeneity	(0.9187, 0.9962)	3.30	2.08	↓
Topic 7	Proteomics in post-translational modification	translational, selection, oncology, proteomics	(0.9392, 0.9968)	3.00	2.62	↓
Topic 8	Transcriptional and metabolic processes	pathway, transcriptional, miRNA, metabolic	(0.9571, 0.9975)	4.90	3.80	↓
Topic 9	Metabolomics of breast cancer	metabolomics, breast, colorectal, sequence	(0.9745, 0.9988)	5.20	6.06	↑
Topic 10	Genomics of colorectal cancer	colorectal, pancreatic, genomic, clustering	(0.9836, 0.9992)	9.20	10.10	↑
Topic 11	Multi-omics analysis	genomic, proteomic, metabolomics, multi-omics	(0.9890, 0.9995)	12.90	14.80	↑
Topic 12	Cancer vaccines	clinical, trial, vaccine, alcoholic	(0.9339, 0.9934)	2.70	2.53	↓
Topic 13	Tumor immunotherapy	biology, epidemiology, immunotherapy, regulator	(0.9565, 0.9976)	3.80	3.48	↓
Topic 14	Metabolomics in prostate cancer	prostate, miRNA, metabolic, liver	(0.9065, 0.9924)	2.50	1.63	↓
Topic 15	Pharmacogenomics	medicine, personalize, precision, oncology	(0.9740, 0.9993)	5.40	5.80	↑
Topic 16	Genome-guided therapy regimens	genome, carcinoma, target, treatment	(0.9728, 0.9975)	5.70	5.84	↑
Topic 17	MicroRNAs target genes	breast, target, mediate, microRNA	(0.9788, 0.9989)	6.60	7.65	↑
Topic 18	Microbial metabolomics	mutation, pathway, microenvironment, immune	(0.9299, 0.9945)	3.00	2.26	↓
Topic 19	Metagenomics	environmental, epigenetic, cycle, fingerprinting	(0.9374, 0.9942)	3.10	2.76	↓
Topic 20	Disease gene localization cloning	disease, colon, model, gene	(0.9474, 0.9979)	3.30	2.94	↓

RESULTS

Topic Extractions

In **Table 1**, the results for topic extractions, we only list the top four words with the highest frequency in each topic, excluding terms like "cancer" or "integration" et al., which are less informative in our analysis. In summary, these topics include: biomarker discovery in early diagnostics (86 studies); machine learning for cancer prediction (148 studies); novel biomarker identify technology (183 studies); multi-omics in Hepatocellular Carcinoma (97 studies); genomics in clinical application (74 studies); genomics in tumor heterogeneity discovery (76 studies); proteomics in post-translational modification (70 studies); transcriptional and metabolic processes (114 studies); metabolomics of breast cancer (121 studies); genomics of colorectal cancer (213 studies); multi-omics analysis (299 studies); cancer vaccines (63 studies); tumor immunotherapy (88 studies); metabolomics in prostate cancer (58 studies); pharmacogenomics (125 studies); genome-guided therapy regimens (132 studies); non-coding RNA target genes (153 studies); microbial metabolomics (70 studies); metagenomics (71 studies); genomics of colorectal cancer (77 studies). In **Table 1**, Column 5 shows the weights of each topic, indicating the numbers of documents that are assigned to the corresponding themes. Topic 11 and Topic 10 hold the most assigned documents, indicating the increased research trends in omics data integration in the multi-omics study.

Column 3 in **Table 1** demonstrates the contradistinctive results of two trend prediction methods: The Prophet model (logistic trend) and the Prophet model (linear trend). For each research topic, the R2_score value of the Prophet model exceeds 0.90, indicating that the Prophet model can well fit the evolution trend of the research topic. The reason is that each topic's distribution has a noticeable growth trend, the sequence is non-stationary, so the Prophet model's effect fits well. Besides, most of the research topics are more in line with the Liner trend, for the R2_score is higher than 0.99 on average because most of the topics in this field are in a period of rapid growth and have not yet reached saturation growth.

From the overall development trend of the subject, the research achievements of omics analysis in cancer research increase year by year. **Figure 1** Panel (D) shows the result. In recent 5–10 years, omics technology has been widely used. More in-depth exploration has been obtained from various cancer research perspectives, indicating that the integration of omics data has been more recognized by the academic. The application of biomarker discovery technology in the study of cancers is supported by medical and biological research.

Trend Prediction Analysis

Column 5 in **Table 1** shows the predicted results based on the Prophet modified model. It can be seen that the research achievements exhibit a growing trend on the whole. However, the growth trends are slightly different in each topic field. Compared between Column 4 and Column 5 in **Table 1**, we can see that the weights of Topic 2, 3, 9, 10, 11, 15, 16, 17 are higher than their current proportions. The river scale map of the predicted-modified topic evolution trend also shows the emergence and the development of various topics. **Figure 1** Panel (E) shows the topic evolution results in the past 15 years. In terms of published journal articles, these topics maintain a high degree of research interest.

Machine learning is an intelligent scientific tool to improve concrete algorithms in experiential learning (Smith et al., 2020). Machine learning has many cancer applications (Han and Li, 2011; Swan et al., 2013; Kim et al., 2016). It can study tumor subtypes' classification and predict tumor patients' phenotype (Hanczar et al., 2020; Smith et al., 2020). For example, to indicate the treatment effect or to predict the recurrence. Machine learning can be combined with molecular networks to study the molecular mechanisms of tumors (Zhu et al., 2019). The development of tumor genomes has promoted the development of machine learning. Meanwhile, the optimization of machine learning algorithms has also announced the research on phenotypic biomarkers of cancer. Hence, Topic 2 machine learning techniques will be more widely used in study refers to cancer prediction.

Topic 3 and Topic 11 refer to novel biomarker identifies technology and multi-omics analysis. By employing Genomics and proteomics technologies, an immense amount of genomic data is being generated on clinical tumors, which has transformed the cancer landscape and can transform cancer diagnosis and prognosis (Shukla, 2017). The future of metabolomics and other omics approaches rests with their ability to monitor subtle changes that occur before the detection of a gross phenotypic change reflecting disease (Kim et al., 2008, 2015). Triple-negative breast cancer (TNBC) represents ∼15% of breast cancers. In light of the complexity of TNBC, by applying transcriptional regulatory and protein-protein interaction networks and tumor necrosis factor signaling pathways can be identified (Karagoz et al., 2015). The statistical clustering approach and the application of the omics methods, both phenotypes, and endotypes, can better illuminate mechanisms and processes that lead to the complexity of asthma (Perlikos et al., 2016). Advances in multi-omics technology have allowed for the delineation of pathways, which will be particularly significant in TH2 low eosinophilic asthma, and also in pauci-inflammatory disease (Abdel-Aziz et al., 2020). Therefore, the multi-omics approach can offer ways forward on novel diagnostics and potentially help to design personalized therapeutics for cancer. In the topic prediction results (see Column 5 and Column 6 in **Table 1**), there will be more research studies focusing on novel biomarker discovery and multi-omics analysis in the future.

Topic 9 describes the rapid development of the metabolomics approach in breast cancer study. A growing literature reports the use of metabolites to modulate diverse processes, such as stem cell differentiation, oligodendrocyte maturation, insulin signaling, T-cell survival, and macrophage immune responses (Zanni et al., 2017; Guijas et al., 2018; Procopet et al., 2018; Njoku et al., 2020). The links between metabolomics and breast cancer keywords show more than 200 strength links. Metabolomics has enabled researchers to complement genomic and protein level analysis of disease with both semi-quantitative and quantitative metabolite levels, which are the chemical mediators that constitute a given phenotype (Njoku et al., 2020). Breast cancer is associated with significantly lowered plasma aspartate levels in a training group comprising 1:1 breast cancer patients and controls (Xie et al., 2015). Researchers find that lowed circulating aspartate is a crucial metabolic feature of human breast cancer (Huang et al., 2016). Another breakthrough in metabolomics analysis has led to the discovery of new targets for cancer therapy. Unlike genes or proteins, metabolites are often readily available, which means that MAS is broadly amenable to high-throughput screening of virtually any biological system (Guijas et al., 2018). Therefore, increasing clinical research will be discovered by applying metabolomics in breast cancer and other types of cancer diseases.

Topic 10 focuses on the importance of genomics study in colorectal cancer. Colorectal cancer (CRC) is a common and lethal disease with a high therapeutic need. In a range of protein, DNA, and RNA-based biomarkers under investigation for CRC, long non-coding RNAs (lncRNA) plasmacytoma variant translocation have been evaluated as a diagnostic, prognostic, and therapeutic biomarker in colorectal cancer(Ogunwobi et al., 2020). Researchers also identified a unique subclass of colorectal cancer characterized by hypermutation associated with the POLE mutation. 7.2% of Early-onset colorectal cancers (EOCRCs) had the POLE P286R mutation, which was not found in late-onset CRCs (LOCRCs) (Ahn et al., 2016). Biomarker analysis supported the functional equivalence of weekly and every 2nd-week administration of cetuximab. It provided further confirmation that patients with KRAS wild-type metastatic colorectal cancer (mCRC) were those most likely to benefit from cetuximab treatment (Tabernero et al., 2010).

Topic 15 and Topic 16 refer to cancer pharmacogenomics and genome-guided therapy regimens. A key aspect of precision medicine is identifying biomarkers that predict the response to medications (i.e., pharmacogenetics) (Kranzler et al., 2017). The primary pharmacogenomics research strategy is to select candidate genes related to drug absorption, transport, activation, metabolism, initiation, and excretion (Shukla, 2017). Then the relationship between gene variation and drug efficacy is analyzed by statistical principle. Multiple genes with specific types of alterations have now been identified associated with improved response to chemotherapy and radiotherapy (Ostrom et al., 2013). Although many academic achievements on biomarkers have been reported, a few biomarkers are used in cancer drug development and clinical settings (Saito et al., 2013). Thus, to enable the optimal selection of drug(s) for a cancer patient, more research finding for critical biomarkers discovery is of urgent demand.

Topic 17 refers to the MicroRNAs target genes study. MicroRNAs (miRNAs) are short, non-coding RNA that negatively regulates gene expression and are differentially expressed in human cancers (i.e., breast cancer, prostate cancer, lung cancer, or CRC) (Gwak et al., 2014; Li et al., 2016). In the last decade, microRNAs have emerged as a new class of gene regulators. To date, about 2,000 human miRNAs have been reported in miRBase (v22) (Kozomara et al., 2019). By employing an improved algorithm for miRNA target prediction, now miRDB can present transcriptome-wide target prediction data, including 3.5 million predicted targets regulated by 7,000 miRNAs in five species (Chen and Wang, 2020). It has been demonstrated that miRNAs control major cellular processes, including metabolism, developmental timing, stem cell division, cell growth and differentiation, and apoptosis (Hannafon et al., 2011). At both post-transcriptional and translational levels, miRNAs regulate most known protein-coding genes (Konno et al., 2014; Kim and Naisbitt, 2016). Given this expansive role, the discovery of miRNAs contributes to the pathogenesis of many cancer diseases (Xi et al., 2016).

To sum up, the recent development of omics and biomarker discovery technology in the past two decades has brought a burst inflection point in the above fields. The above eight topics will keep a rapid development in the future. Even though the remaining topics trends show decreases, it does not mean the related research are getting less. In **Table 1** Column 4, the Prophet-logistic results are lower than Prophet-liner results, indicating that research related to other topics will keep a steady development in the short term.

DISCUSSION AND CONCLUSION

Conclusions

The scientific panorama involved in studying the omics data integration toward the mining of biomarkers in cancers is described in the 20 extracted topics. Among the 2,318 sample studies, the core element from the current scientific discussion is the multi-omics analysis. For each topic, the Prophet model can better adapt to the evolution trend. In addition, all of the research topics are in line with the Liner trend, indicating a development stage for omics analysis in cancer studies. The research methodology proposed in this study is hoped to promote a different approach to conceptualizing and treating omics data integration study based on existing methods.

From the perspective of the distance between topics and the evolution of distribution, the innovative technologies of biomarker discovery represented by genomics, metabolomics, and proteomics are the actual contents of integrating omics data in cancer research. Genomics has done extensive research on cancer cells and cancer patients in different treatment stages. With the continuous development and update of international genome-related databases, traditional machine learning algorithms and deep neural networks are essential for data integration and disease prediction. In the future, innovative algorithms will be explored for biomarker discovery technology in different stages of different types of cancer diseases. There will be more research performing clustering optimizations in the development of multi-omics data integration.

Metabolomics has made many achievements in the research of genitourinary tumors, lung cancer, and breast cancer, mainly reflected in the study of metabolic phenotype biomarkers in the metabolic pathway. It is also used in the research of cancer treatment and prognosis prediction. Vertically omics data integration among proteomics, genomics, and metabolomics is becoming more common in the study of cancers. To obtain biological information through omics technology and to discover the molecular mechanism related to cancers are the primary purpose of proteomics research. To detect new marker molecules and to improve the clinical treatment effect, the development of a multi-omics data platform and schema database together with the exploration of a multi-omics feature expression algorithm will have more practical significance.

Discussions

In omics research toward the mining of biomarkers in cancers, our study summarizes the existing evolution and predicts the future trends of researches. Specifically, the adopted topic model describes and forecasts the change of time segments, which helps track the research development. Additionally, as our results go beyond the original thinking of theory and framework, this study also helps scholars and readers to understand the hot spots and future opportunities regarding omics data integration in cancer studies.

It is noted that there also exist some limitations. On the one hand, although we described the hot research issues and predicted the frontier research issues, there is still a shortage in natural language processing. Due to the variety of text language rules in different stages and different cancer diseases, the same substance can be expounded in several ways. In the process of document-level natural language processing, the expression differences of specific terms cannot be avoided. On the other hand, this paper's conclusion does not involve finding out the research deficiencies, technical challenges, and other omics research problems toward the mining of phenotype-specific biomarkers. Expert systems development and optimizing algorithms will be performed in the next study to provide a more comprehensive analysis of the research problems.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

LN wrote the manuscript. HH contributed to the modification of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by Internet Innovation Research Center Fund (Grant no: IIRC20200104) and by Huaqiao University Research Fund (Grant no: 605-50Y17031).

REFERENCES

Abdel-Aziz, M. I., Neerincx, A. H., Vijverberg, S. J., Kraneveld, A. D., and Maitland-van der Zee, A. H. (2020). Omics for the future in asthma. *Semin. Immunopathol*. 42, 111–126. doi: 10.1007/s00281-019-00776-x

Ahn, S.-M., Ansari, A. A., Kim, J., Kim, D., Chun, S.-M., Kim, J., et al. (2016). The somatic POLE P286R mutation defines a unique subclass of colorectal cancer featuring hypermutation, representing a potential genomic biomarker for immunotherapy. *Oncotarget* 7, 68638–68649. doi: 10.18632/oncotarget.11862

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.1371/journal.pone.0245393

Chen, Y. H., and Wang, X. W. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res*. 48, D127–D131. doi: 10.1093/nar/gkz757

Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E., and Siuzdak, G. (2018). Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat. Biotechnol*. 36, 316–320. doi: 10.1038/nbt.4101

Gwak, J. M., Kim, H. J., Kim, E. J., Chung, Y. R., Yun, S., Seo, A. N., et al. (2014). MicroRNA-9 is associated with epithelial-mesenchymal transition, breast cancer stem cell phenotype, and tumor progression in breast cancer. *Breast Cancer Res. Treat*. 147, 39–49. doi: 10.1007/s10549-014-3069-5

Han, H., and Li, X.-L. (2011). Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. *BMC Bioinform*. 12:S7. doi: 10.1186/1471-2105-12-S1-S7

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hanczar, B., Zehraoui, F., Issa, T., and Arles, M. (2020). Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinform*. 21:501. doi: 10.1186/s12859-020-03836-4

Hannafon, B. N., Sebastiani, P., de las Morenas, A., Lu, J., and Rosenberg, C. L. (2011). Expression of microRNA and their gene targets are dysregulated in preinvasive breast cancer. *Breast Cancer Res*. 13:R24. doi: 10.1186/bcr2839

Hofmann, T. (1999). "Probabilistic latent semantic indexing," in *Proceedings of 22nd International Conference on Research and Development in Information Retrieval* (Berkeley, CA), 55–57. doi: 10.1145/312624.312649

Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G., and Garmire, L. X. (2016). Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome Med*. 8:34. doi: 10.1186/s13073-016-0289-9

Karagoz, K., Sinha, R., and Arga, K. Y. (2015). Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *OMICS* 19, 115–130. doi: 10.1089/omi.2014.0135

Kim, S., Lin, C.-W., and Tseng, G. C. (2016). MetaKTSP: a meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis. *Bioinformatics* 32, 1966–1973. doi: 10.1093/bioinformatics/btw115

Kim, S. H., and Naisbitt, D. J. (2016). Update on advances in research on idiosyncratic drug-induced liver injury. *Allergy Asthma Immunol. Res*. 8, 3–11. doi: 10.4168/aair.2016.8.1.3

Kim, Y. J., Sertamo, K., Pierrard, M.-A., Mesmin, C., Kim, S. Y., Schlesser, M., et al. (2015). Verification of the biomarker candidates for non-small-cell lung cancer using a targeted proteomics approach. *J. Proteome Res*. 14, 1412–1419. doi: 10.1021/pr5010828

Kim, Y. S., Maruvada, P., and Milner, J. A. (2008). Metabolomics in biomarker discovery: future uses for cancer prevention. *Future Oncol*. 4, 93–102. doi: 10.2217/14796694.4.1.93

Konno, Y., Dong, P., Xiong, Y., Suzuki, F., Lu, J., Cai, M., et al. (2014). MicroRNA-101 targets EZH2, MCL-1 and FOS to suppress proliferation, invasion and stem cell-like phenotype of aggressive endometrial cancer cells. *Oncotarget* 5, 6049–6062. doi: 10.18632/oncotarget.2157

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 47, D155–D162. doi: 10.1093/nar/gky1141

Kranzler, H. R., Smith, R. V., Schnoll, R., Moustafa, A., and Greenstreet-Akman, E. (2017). Precision medicine and pharmacogenetics: what does oncology have that addiction medicine does not? *Addiction* 112, 2086–2094. doi: 10.1111/add.13818

Li, D.-C., Rastegar-Mojarad, M., Okamoto, J., Liu, H., and Leichow, S. (2015). A bibliometric analysis on cancer population science with topic modeling. *AMIA Joint Summits on Translational Science proceedings*. *AMIA Joint Summits on Translational Science, 2015* (San Francisco, CA), 102–106.

Li, T., Lai, Q., Wang, S., Cai, J., Xiao, Z., Deng, D., et al. (2016). MicroRNA-224 sustains Wnt/beta-catenin signaling and promotes aggressive phenotype of colorectal cancer. *J. Exp. Clin. Cancer Res.* 35:21. doi: 10.1186/s13046-016-0287-1

Long, N. P., Jung, K. H., Anh, N. H., Yan, H. H., Nghi, T. D., Park, S., et al. (2019). An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. *Cancers* 11:20155. doi: 10.3390/cancers11020155

Nazifova-Tasinova, N., Radeva, M., Galunska, B., and Grupcheva, C. (2020). Metabolomic analysis in ophthalmology. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech. Repub*. 164, 236–246. doi: 10.5507/bp.2020.028

Ning, L., Peng, L. F., and He, H. X. (2020). Prediction correction topic evolution research for metabolic pathways of the gut microbiota. *Front. Mol. Biosci*. 7:720. doi: 10.3389/fmolb.2020.600720

Njoku, K., Sutton, C. J., Whetton, A. D., and Crosbie, E. J. (2020). Metabolomic biomarkers for detection, prognosis and identifying recurrence in endometrial cancer. *Metabolites* 10:314. doi: 10.3390/metabo10080314

Ogunwobi, O. O., Mahmood, F., and Akingboye, A. (2020). Biomarkers in colorectal cancer: current research and future prospects. *Int. J. Mol. Sci.* 21:5311. doi: 10.3390/ijms21155311

Ostrom, Q., Cohen, M. L., Ondracek, A., Sloan, A., and Barnholtz-Sloan, J. (2013). Gene markers in brain tumors: what the epileptologist should know. *Epilepsia* 54, 25–29. doi: 10.1111/epi.12439

Perlikos, F., Hillas, G., and Loukides, S. (2016). Phenotyping and endotyping asthma based on biomarkers. *Curr. Top. Med. Chem*. 16, 1582–1586. doi: 10.2174/1568026616666150930120803

Procopet, B., Fischer, P., Farcau, O., and Stefanescu, H. (2018). Metabolomics: from liver chiromancy to personalized precision medicine in advanced chronic liver disease. *World J. Hepatol*. 10, 371–378. doi: 10.4254/wjh.v10.i3.371

Ristori, M. V., Mortera, S. L., Marzano, V., Guerrera, S., Vernocchi, P., Ianiro, G., et al. (2020). Proteomics and metabolomics approaches towards a functional insight onto AUTISM spectrum disorders: phenotype stratification and biomarker discovery. *Int. J. Mol. Sci*. 21:6274. doi: 10.3390/ijms21176274

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. (2010). Learning author-topic models from text corpora. *Acm Trans. Inform. Syst*. 28:4. doi: 10.1145/1658377.1658381

Saito, Y., Sai, K., Kaniwa, N., Tajima, Y., Ishikawa, M., Nishimaki-Mogami, T., et al. (2013). Biomarker exploration and its clinical use. *Yakugaku Zasshi* 133, 1373–1379. doi: 10.1248/yakushi.13-00232-2

Shukla, H. D. (2017). Comprehensive analysis of cancer-proteogenome to identify biomarkers for the early diagnosis and prognosis of cancer. *Proteomes* 5:28. doi: 10.3390/proteomes5040028

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., et al. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinform*. 21:119. doi: 10.1186/s12859-020-3427-8

Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S., and Bacardit, J. (2013). Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS* 17, 595–610. doi: 10.1089/omi.2013.0017

Tabernero, J., Cervantes, A., Rivera, F., Martinelli, E., Rojo, F., von Heydebreck, A., et al. (2010). Pharmacogenomic and pharmacoproteomic studies of cetuximab in metastatic colorectal cancer: biomarker analysis of a phase I dose-escalation study. *J. Clin. Oncol*. 28, 1181–1189. doi: 10.1200/JCO.2009.22.6043

Urh, K., and Kunej, T. (2016). Molecular mechanisms of cryptorchidism development: update of the database, disease comorbidity, and initiative for standardization of reporting in scientific literature. *Andrology* 4, 894–902. doi: 10.1111/andr.12217

Valle, F., Osella, M., and Caselle, M. (2020). A topic modeling analysis of tcga breast and lung cancer transcriptomic data. *Cancers* 12:3799. doi: 10.3390/cancers12123799

Xi, X.-P., Zhuang, J., Teng, M.-J., Xia, L.-J., Yang, M.-Y., Liu, Q.-G., et al. (2016). MicroRNA-17 induces epithelial-mesenchymal transition consistent with the cancer stem cell phenotype by regulating CYP7B1 expression in colon cancer. *Int. J. Mol. Med*. 38, 499–506. doi: 10.3892/ijmm.2016.2624

Xie, G., Zhou, B., Zhao, A., Qiu, Y., Zhao, X., Garmire, L., et al. (2015). Lowered circulating aspartate is a metabolic feature of human breast cancer. *Oncotarget* 6, 33369–33381. doi: 10.18632/oncotarget.5409

Zanni, E., Schifano, E., Motta, S., Sciubba, F., Palleschi, C., Mauri, P., et al. (2017). Combination of metabolomic and proteomic analysis revealed different features among lactobacillus delbrueckii subspecies bulgaricus and lactis strains while *in vivo* testing in the model organism caenorhabditis elegans highlighted probiotic properties. *Front. Microbiol.* 8:1206. doi: 10.3389/fmicb.2017.01206

Zhu, Z., Albadawy, E., Saha, A., Zhang, J., Harowicz, M. R., and Mazurowski, M. A. (2019). Deep learning for identifying radiogenomic associations in breast cancer. *Comput. Biol. Med.* 109, 85–90. doi: 10.1016/j.compbiomed.2019.04.018

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