Novel Biomarker Candidates for Colorectal Cancer Metastasis: A Meta-analysis of *In Vitro* Studies

Nguyen Phuoc Long, Wun Jun Lee, Nguyen Truong Huy, Seul Ji Lee, Jeong Hill Park and Sung Won Kwon

College of Pharmacy and Research Institute of Pharmaceutical Sciences, Seoul National University, Seoul, Korea.
The first two authors contributed equally to this work.

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes (A)

ABSTRACT: Colorectal cancer (CRC) is one of the most common and lethal cancers. Although numerous studies have evaluated potential biomarkers for early diagnosis, current biomarkers have failed to reach an acceptable level of accuracy for distant metastasis. In this paper, we performed a gene set meta-analysis of in vitro microarray studies and combined the results from this study with previously published proteomic data to validate and suggest prognostic candidates for CRC metastasis. Two microarray data sets included found 21 significant genes. Of these significant genes, ALDOA, IL8 (CXCL8), and PARP4 had strong potential as prognostic candidates. LAMB2, MCM7, CXCL23A, SERPINA3, ABCA3, ALDH3A2, and POLR2I also have potential. Other candidates were more controversial, possibly because of the biologic heterogeneity of tumor cells, which is a major obstacle to predicting metastasis. In conclusion, we demonstrated a meta-analysis approach and successfully suggested ten biomarker candidates for future investigation.

KEYWORDS: colorectal cancer, biomarker candidate, microarray analysis, proteomics

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes (A)

CITATION: Long et al. Novel Biomarker Candidates for Colorectal Cancer Metastasis: A Meta-analysis of *In Vitro* Studies. *Cancer Informatics* 2016:15(S4) 11–17
doi: 10.4137/CIN.S40301.

TYPE: Original Research

RECEIVED: June 22, 2016. **RESUBMITTED:** August 15, 2016. **ACCEPTED FOR PUBLICATION:** August 16, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 881 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by the Rural Development Administration of Korea (PJ011646) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0023057, NRF-2013R1A1A1A05005753). This work was supported by BK21 Plus Program in 2016.The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: swkwon@snu.ac.kr

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and the fourth leading cause of cancer-related mortality and affects both genders equally.[1] It is also of interest to note the geographic differences in the incidence and survival rates among different regions around the world.[2] Currently, CRC is believed to be a disease that mainly occurs in developed countries. However, a sudden increase in the incidence of CRC with industrialization is becoming a major concern in developing countries.[3,4]

Although the prognosis of CRC patients depends on many factors, such as patient anthropometric and biochemical characteristics, therapeutic options, and personal care, early diagnosis is a key factor for reducing the overall mortality.[5,6] For instance, when CRC patients are diagnosed early and treated appropriately, more than 90% of patients survive longer than 5 years. In contrast, the survival rate dramatically decreases to 10% of patients with distant metastases before diagnosis.[5] Taking into consideration the need for novel biomarkers, numerous studies have been conducted to evaluate potential gene and protein biomarkers. However, current biomarkers in clinical use have failed to reach an acceptable sensitivity and specificity for appropriate disease diagnosis.[7] For other candidates, additional evidence from large, well-designed studies is needed before they can be used in clinical practice.[8] Moreover, a pressing need for finding, validating, and combining biomarkers for prognosis is undeniable.[9] From this perspective, microarray gene-based profiling technology has shown strong potential because of its capacity to provide substantial data in a short time, which eventually enhances the possibilities for discovering new candidate biomarkers.[10]

It is worth noting that individual gene expression analysis does not provide adequate information for translating biological processes.[11] On the other hand, Gene set enrichment analysis (GSEA) is a powerful method that focuses on the gene groups, which share common functions, instead of focusing on the correlation between the gene expression and a given phenotype.[12] However, the uncertain robustness in identifying gene expression profiles among the comparative groups and limitations of the GSEA method should be carefully considered.[13] Important methods for improving the data quality include increasing the sample size, using a better statistical analysis

algorithm, integrating multi-omics platforms, performing dependent validation studies, and combining the results of several studies with similar experimental designs.[14,15]

In this study, we conducted a meta-analysis of available microarray data sets on human colorectal cell lines using a generally applicable gene set enrichment (GAGE) approach to detect gene candidates for metastatic cancer. GAGE was chosen because it expands the applicability of gene set analysis in several aspects and overcomes the limitations of GSEA.[16,17] Then, we validated the consistency of the gene expression results and combined the findings of this study with the previously published results from a proteogenomic analysis on three different cell lines.[18] The results of this study demonstrate the importance of combining and crosschecking the results from different studies because a single study or platform alone does not provide adequate reliability.[19] Finally, we suggested a list of potential biomarker candidates for future investigation.

Methods

Microarray analysis. *Data collection*. Figure 1 shows the workflow of the data collection. In short, we retrieved microarray data from the Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) using "colorectal cancer" and "colon cancer" as search terms. From a total of 18,235 results, we used the DataSets option, which included 61 available data sets on CRC. Next, only data sets including human cell lines were considered based on the following criteria: (1) the study was conducted on human colorectal cell lines and (2) the study compared differentially expressed gene data in at least two cell lines that represent different aggressive properties. As a result, three data sets, namely, GSE1323,[20] GSE14733,[21] and GSE15102,[22] met the criteria. However, the GSE15102 data set contains only one sample for each comparative group, which is not suitable for data analysis. Eventually, two data sets (GSE1323 and GSE14773) were chosen for the next evaluation.

Data preprocessing. The data sets from GSE1323 and GSE14773 were created based on the Affymetrix Human Genome U133A Array and Affymetrix Human Genome U133 Plus 2.0 Array, respectively. First, we applied a robust multi-array average algorithm using the affy package for background adjustment, normalization, and summarization of the data sets.[23]

Because GSE14773 contains two comparisons of different cell lines, the ComBat method was used to remove the batch effect and combine the data a single data set.[24] Initially, the GSE14773 samples were clustered into two main branches for the cell lines (Fig. 2A). After applying the ComBat method to adjust the batch effect, the samples were regrouped according to their aggressive properties (Fig. 2B). Finally, according to the data sets provided by Fanayan et al.[18], we extracted 2,476, 2,455, and 1,866 associated proteins/genes from LIM1215, LIM1899, and LIM2405, respectively. LIM1899 and LIM2405 represent the more aggressive cancer cell lines compared to LIM1215.

Gene set analysis. Prior to data analysis, we converted gene labels to Entrez IDs by Database for Annotation, Visualization, and Integrated Discovery (DAVID).[25] The GAGE

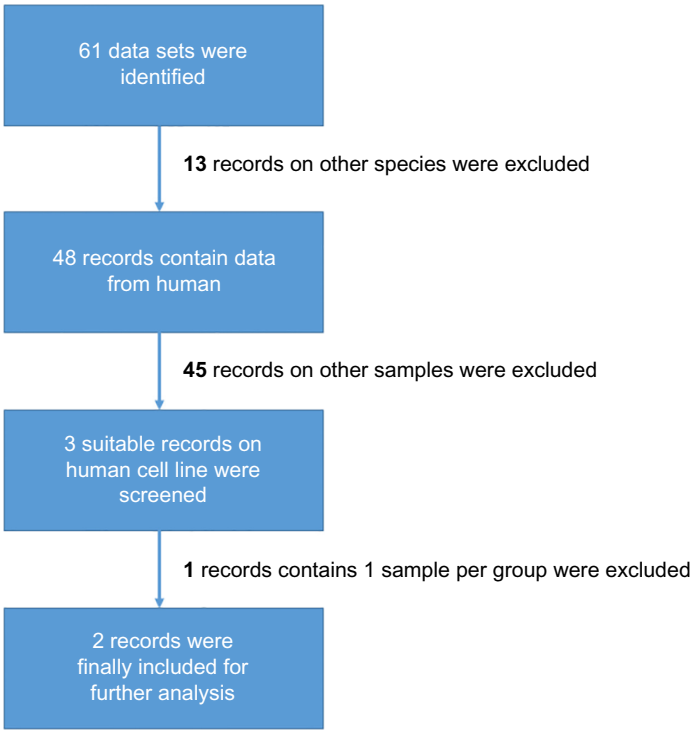

Figure 1. Data collection flowchart. Of 61 data sets, two data sets were included for further investigation.

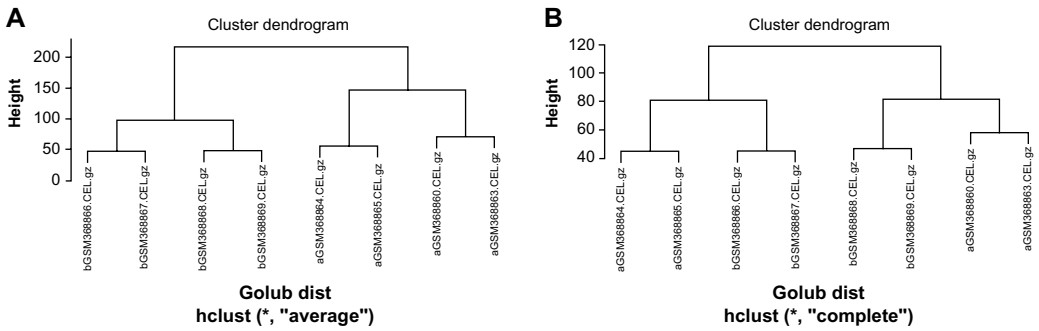

Figure 2. Adjusting for batch effects. Batch effects from different GSE14773 groups before (**A**) and after (**B**) applying the ComBat method. aGSM368860 and aGSM368863: HT29 parental control. aGSM368864 and aGSM368865: HT29 colonospheres. bGSM368866 and bGSM368867: SW480 Vector. bGSM368868 and bGSM368869: CRC SW480 with SNAIL overexpression.

Bioconductor package was then used to perform gene set analysis.[17] *P*-values with a cutoff value of $P < 0.01$ were used to choose the differentially expressed gene sets from two final data sets (GSE1323 and GSE14773 combined).

Validation and selection of candidates. For validation, the leave-one-out cross-validation (LOOCV) method was used to obtain the accuracy of significant gene sets by using prediction analysis for microarrays (PAM), as previously described.[14] PAM uses "the nearest shrunken centroid classification" method to predict the category of a sample with respect to its gene expression profile.[26,27] We conducted the LOOCV for each data set. Iteratively, each sample in the data set was removed, and the remaining samples were utilized to develop a prediction model with PAM.[26] The model was then applied to predict the categorization of the removed sample. After selecting significant gene sets, we performed a global test using the globaltest R package for all individual genes of each gene set.[28] Each individual gene with a *P*-value < 0.05 and exhibiting the same expressed direction between two microarray data sets was selected. The positive and negative associations represent the upregulation (Up) and downregulation (Down), respectively, of selected genes.[28] In our study, for example, a positive association indicates that the expression of the current gene was upregulated in more aggressive cancer cells and vice versa. Finally, peptide spectral count (PSC), a semi-quantitative parameter of protein abundance, from proteomic data sets provided by Fanayan et al.[18] was extracted for every significant gene in our meta-analysis. When the PSC value of a given gene could not be found, the value was set as "0".

The gene candidates were initially classified into the following three groups: (1) good candidates include all genes with compatibility between the expressed direction and PSC, (2) candidates include all genes with PSC = 0, and (3) controversial candidates include all genes with the opposite trend between the expressed direction and PSC. It is useful to emphasize that the controversial results might be due to the biologic heterogeneity of the tumor cell lines because this study combined different cell lines in the statistical analysis.[29]

Results

Data characteristics. In this paper, we included two different microarray experiments, GSE1323 and GSE14773. GSE1323 contains the gene expression information for two cancer cell lines, SW480 and SW620. In detail, SW480 and SW620 represent primary tumor and lymph node metastasis, respectively. By using this model, changes in the gene expression in late progression could be properly analyzed.[20] Additionally, GSE14773 contains gene expression comparisons between HT29 colonospheres versus HT29 "parental controls" and SW480 SNAIL versus SW480 vector. CRC HT29 colonospheres strongly expressed CD44 and CD166, which exhibit more aggressive malignant properties.[30] In addition, CRC SW480 with SNAIL overexpression has epithelial–mesenchymal transition properties, which enhance invasion and chemoresistance.[31] Using meta-analysis to evaluate the gene expression changes in the aforementioned models may reveal potential gene candidates for cancer metastasis. Finally, two data sets (GSE1323 and GSE14773-combined) were included for GAGE.

Gene set analysis and validation. After using gene set analysis with the cutoff of $P < 0.01$, we obtained 12 significant pathways from GSE1323 and 28 significant pathways from GSE14773-combined. Based on the results, we chose the overlapped significant gene sets between two data sets (Fig. 3, Supplementary File 1). As a result, five gene sets (hsa03420 nucleotide excision repair, hsa03030 DNA replication, hsa04060 cytokine–cytokine receptor interaction, hsa01430 cell junctions, and hsa00240 pyrimidine metabolism) were selected for further analysis and validation. Table 1 shows the *P*-value, *Q*-value, and accuracy of the five gene set candidates, and Supplementary File 2 presents the results of LOOCV.

Candidate genes. The significant genes from the microarray data sets and proteomic data set were combined to select potential candidates (Table 2). In addition to performing statistical analysis, we considered individual candidates along the line of action mechanisms for the corresponding gene sets in cancer cells.

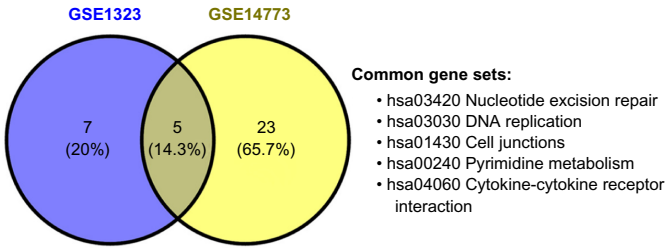

Figure 3. Venn diagram of the common gene sets between two data sets.
Notes: Five gene sets (hsa03420 nucleotide excision repair, hsa03030 DNA replication, hsa04060 cytokine–cytokine receptor interaction, hsa01430 cell junctions, and hsa00240 pyrimidine metabolism) were selected.

The nucleotide excision repair (NER) pathway plays an essential role in eliminating DNA lesions caused by carcinogens.[32] The mechanisms of action of NER are well established.[33] In a previous study, Berndt et al.[34] suggested that the genetic variations in this pathway might be associated with a higher risk of CRC. The gene expression of ABCA3, POLD1, and SERPINA3, all of which belong to NER pathway, satisfied *P*-value <0.05 in the same direction in both microarray data sets. Oncogenes might affect DNA replication, a crucial phenomenon of biological inheritance, via several mechanisms. In addition, dysregulation of the related genes of DNA replication may be a prognostic factor.[35] Surprisingly, our gene expression data indicated that all ABCA3, POLD1, SERPINA3, and MCM7 are involved in DNA replication and have downregulated expression. The cytokine–cytokine receptor interaction pathway has been reported to be involved in metastasis and treatment resistance.[36] Its roles were demonstrated in our study in which relatively abundant candidates were found (AKR1B1, ABCA3, TNFSF12-TNFSF13, IL8, ALDH3A2, IL23A, PARP4, ALDOA, ABCD1, SERPINA3, AHCY, and ACTN1). The relationship between cell junctions and cancer metastasis is well established.[37,38] The potential for using cell junction molecules as prognostic markers has also achieved good results, making the cell junction pathway an interesting target.[39] Among the genes in the cell junctions gene set, ABCA3, LAMB2, LAMA5, SERPINA3, and ACTN1 satisfied *P*-value < 0.05 in the same direction in both microarray data sets. Pyrimidine metabolism is one of the major pathways involved in DNA synthesis, a fundamental process for the survival of both normal and cancer cells. The inhibitors of the purine and pyrimidine metabolism have been the main targets for cancer treatments. Several studies found dysregulation of pyrimidine metabolism in CRC and suggested potential metabolite markers.[40] Similarly, our gene set analysis suggested 10 candidates related to this pathway (ABCA3, POLR1D, UCKL1, POLR2I, POLR2B, POLD1, NT5E, SERPINA3, ACTN1, and ENTPD5). In the next paragraph, for convenience, we divided the significant genes into three initial groups: (1) good candidate, (2) candidate, and (3) controversial candidate.

Initially, there were six genes, ALDH3A2, ALDOA, LAMB2, MCM7, PARP4, and POLR2I, belonging to the "good candidate" group. ALDOA was the only gene showing the "up" expressed direction. In addition, the PSCs of ALDOA were 3.5-fold and 3.1-fold higher in LIM1899 and LIM2405, respectively, than in LIM1215. ALDOA seems to be a reasonably good candidate based on the consistency in the expressed direction between two microarray data sets and the high protein expression level in LIM cell lines. In contrast, ALDH3A2, LAMB2, MCM7, PARP4, and POLR2I exhibited the "down" expressed direction. With the exception of POLR2I, the corresponding proteins in those genes were only expressed in the LIM1215 cell line.

Five genes were upregulated (ABCD1, IL8, IL23, POLR2B, and TNFSF12–TNFSF13) and seven genes were downregulated (ABCA3, ENTPD5, LAMA5, POLD1, POLR1D, SERPINA3, and UCKL1) in the "candidate" group. No corresponding proteins could be detected in the LIM cell lines. Among the significant genes, ABCA3 and SERPINA3 were downregulated in all five gene sets, while POLD1 was downregulated in three significant gene sets (hsa03420 nucleotide excision repair, hsa03030 DNA replication, and hsa00240 pyrimidine metabolism).

Table 1. *P*-value, *Q*-value, and the accuracy of the five gene set candidates.

DATA SET	GENE SET	*P*-VALUE	*Q*-VALUE	ACCURACY (%)
GSE1323	hsa00240 Pyrimidine metabolism	2.59E-3	1.12E-1	100
	hsa01430 Cell junctions	1.21E-3	6.96E-2	100
	hsa03030 DNA replication	4.70E-5	4.04E-3	100
	hsa03420 Nucleotide excision repair	1.96E-5	3.38E-3	100
	hsa04060 Cytokine-cytokine receptor interaction	5.00E-3	1.23E-1	100
GSE14773	hsa00240 Pyrimidine metabolism	3.50E-6	8.70E-5	100
	hsa01430 Cell junctions	1.30E-3	3.76E-2	100
	hsa03030 DNA replication	1.66E-14	2.89E-12	100
	hsa03420 Nucleotide excision repair	3.88E-5	8.44E-4	100
	hsa04060 Cytokine–cytokine receptor interaction	1.42E-8	1.92E-6	100

Table 2. Differentially expressed genes and corresponding peptide spectral counts (with a correction) from Fanayan et al.[18]

CLASSIFICATION	CANDIDATE	GSE1323		GSE14773		PEPTIDE SPECTRAL COUNTS*		
		P-VALUE	DIRECTION	*P*-VALUE	DIRECTION	LIM1215	LIM1899	LIM2405
Good candidate	ALDH3A2[3]	3.25E-03	Down	4.77E-03	Down	1	0	0
	ALDOA[3]	1.89E-02	Up	9.27E-03	Up	37	131	115
	LAMB2[4]	8.44E-04	Down	1.22E-02	Down	1	0	0
	MCM7[2]	2.80E-02	Down	1.33E-02	Down	5	0	0
	PARP4[3]	1.37E-02	Down	1.43E-03	Down	2	0	0
	POLR2I[5]	1.12E-03	Down	3.05E-02	Down	2	1	0
Candidate	ABCA3[^,1,2,3,4,5]	1.06E-04	Down	2.54E-02	Down	0	0	0
	ABCD1[3]	2.27E-02	Up	8.21E-03	Up	0	0	0
	ENTPD5[5]	4.36E-02	Down	1.23E-02	Down	0	0	0
	IL8 (CXCL8)[3]	2.35E-03	Up	8.78E-04	Up	0	0	0
	IL23A[3] (CXCL23A)	6.07E-03	Up	9.64E-03	Up	0	0	0
	LAMA5[4]	3.70E-03	Down	8.48E-03	Down	0	0	0
	POLD1[#,1,2,5]	1.10E-02	Down	5.14E-05	Down	0	0	0
	POLR1D[5]	1.50E-04	Down	2.94E-02	Down	0	0	0
	POLR2B[5]	2.36E-03	Up	4.34E-02	Up	0	0	0
	SERPINA3[^,1,2,3,4,5]	2.61E-02	Down	1.87E-02	Down	0	0	0
	TNFSF12–TNFSF13[3]	3.08E-04	Up	3.19E-02	Up	0	0	0
	UCKL1[5]	1.05E-03	Down	2.81E-03	Down	0	0	0
Controversial candidate	ACTN1[^,3,4,5]	3.60E-02	Down	1.90E-02	Down	36	273	172
	AHCY[3]	3.01E-02	Down	6.99E-03	Down	1	0	59
	AKR1B1[3]	5.97E-05	Up	8.21E-05	Up	3	0	0
	NT5E[5]	2.24E-02	Down	2.86E-04	Down	0	0	15

Notes: *Results from Fanayan et al.[18] ^Presentative *P*-value from hsa04060 cytokine–cytokine receptor interaction. #Presentative *P*-value from hsa03030 DNA replication. [1]hsa03420 nucleotide excision repair. [2]hsa03030 DNA replication. [3]hsa04060 cytokine–cytokine receptor interaction. [4]hsa01430 cell junctions. [5]hsa00240 pyrimidine metabolism.

In the controversial candidate group, AKR1B1 was the only upregulated gene. In contrast, ACTN1, AHCY, and NT5E were downregulated. Although ACTN1 was downregulated in more aggressive cell lines in the microarray data, there was reversed protein expression, and the PSC values of more aggressive cancer cell lines (LIM1899 and LIM2405) were 7.6-fold and 4.7-fold higher than for the LIM1215 cell line, respectively. In case of AHCY and NT5E downregulated direction, their corresponding proteins, however, expressed just in LIM2405 cell line. The corresponding AKR1B1 protein could only be found in LIM1215, although it belonged to the upregulated group.

Discussion

Meta-analysis is an approach based on combining available data sets to increase the statistical power and produce an estimation of the effect in a pooled analysis.[41] According to Hung et al.[42], the annual publication of meta-analyses has significantly increased since the 1990s. Although this approach was mainly applied in epidemiology and clinical medicine, several microarray-based gene expression meta-analysis studies have also been conducted.[14] In this paper, we integrated two *in vitro* microarray-based gene set expression data sets from GEO and the results of a previous proteomic analysis to suggest potential CRC candidates. Therefore, this approach takes into account both statistical processes and biological mechanisms. Among five significant gene sets, we found several significant genes in each gene set that could be novel candidates for further investigations.

ALDH3A2, ALDOA, LAMB2, MCM7, PARP4, and POLR2I had strong potential as prognostic candidates in our statistical analysis. However, except for ALDH3A2 and POLR2I, the roles of other genes and their corresponding proteins in cancer metastasis were well recorded in the literature, with or without direct evidence in CRC. The ALDOA expression level was significantly upregulated in various highly metastatic cancers, including CRC.[43,44] Downregulation of LAMB2, an extracellular matrix glycoprotein, has been reported to correlate with the advanced stages of ovarian and prostate cancer.[45] MCM7 was downregulated in our study, which is inconsistent with previous reports on different cancer types. Liu et al.[46] and Zhong et al.[47] showed that high expression of MCM7 was associated with shorter survival of non-small-cell lung carcinoma and esophageal squamous cell carcinoma, respectively. We applied PROgene,

a web application of gene expression-based survival analysis for multiple cancers, to evaluate the current evidence between MCM7 gene expression and survival.[48] The results were insignificant in 10 included data sets; the only significant result was poor overall survival in patients with MCM7 downregulation in the GSE16125 analysis (hazard ratio = 0.29, P-value < 0.05).[49] Although the roles of the PARP family in cancer biology were identified, understanding of the cancer-relevant roles of PARP4 is limited.[50] However, PARP4 might be a tumor suppressor in primary thyroid and breast cancer.[51] Microarray-based gene expression and proteomic analysis showed a trend of PARP4 in downregulation in more aggressive CRC cell lines, implying its potential as a prognostic candidate.

Among the 12 members belonging to the candidate group (all PSCs = 0) in our analysis, some candidates have direct or indirect evidence of their behavior in previous studies. For example, IL8 (CXCL8) is an excellent prognostic candidate. In a recent meta-analysis evaluating the clinicopathologic features and diagnostic accuracy of IL8 in CRC, the authors suggested its potential in detecting and predicting the prognosis of CRC.[52] Several indirect pieces of evidence showed that high IL23 (CXCL23A, another CXC chemokine) expression might be a predictor of invasiveness in esophageal squamous cell carcinoma and cutaneous melanomas.[53,54] Nevertheless, a preliminary study by Adamo et al.[55] showed no correlation between the IL23 concentration and CRC severity. Belonging to the serine protease inhibitor protein superfamily, SERPINA3 is a protease inhibitor involved in various inflammatory reactions and malignant tumors.[56] In a previous study, the SERPINA3 concentration was found to be significantly lower in CRC tissues than in normal tissues.[57] However, the authors found no significant difference in the SERPINA3 concentration among different CRC stages. Our gene expression data, on the other hand, showed a significant downregulation of gene expression in more aggressive cell lines. Interestingly, the SERPINA3 expression could not be detected in a proteomic evaluation of LIM1215, LIM1889, and LIM2405 cell lines. The opposite results demonstrate the need to further investigate SERPINA3 behavior, because it may be a potential diagnostic and prognostic candidate. ATP-binding cassette (ABC) transporters are involved in many important biological processes in humans.[58] Among our candidates, the ABCA3 gene was downregulated in more aggressive CRC cell lines. Decreased expression of ABCA3 might correlate with drug-resistant cell lines, as suggested for the A2780 ovarian cancer cell line.[59] However, the expression behaviors of ABCA3 require more validation.

Then again, there is lack of evidence for other candidates such as ACTN1, AHCY, AKR1B1, and NT5E, or the available evidence of these candidates showed reverse gene expression patterns compared to our meta-analysis. For example, NT5E was reported to be involved in CRC tumor invasion and metastasis,[60] and its corresponding protein showed increased expression in LIM2405 (more aggressive cell lines). However, microarray-based gene expression in our meta-analysis showed that NT5E gene expression was downregulated in the more aggressive cell lines. The disagreement may arise from the complicated biological processes of cancer, which are not fully understood. The current status requires further investigations regarding their potential as prognostic candidates.

Conclusion

Our study used a meta-analysis approach to better understand the gene expression behaviors of CRC cell lines with differing levels of aggressiveness. The gene set analysis, which considered the biological significance of gene expression, may provide a novel approach for identifying potential prognostic candidates for further validation.

Author Contributions

Conceived and designed the experiments: WJL, NPL, NTH, SJL, JHP, SWK. Analyzed the data: WJL, NPL. Wrote the first draft of the manuscript: NPL, WJL. Contributed to the writing of the manuscript: WJL, NPL, NTH, SJL, JHP, SWK. Agree with manuscript results and conclusions: WJL, NPL, NTH, SJL, JHP, SWK. Jointly developed the structure and arguments for the paper: WJL, NPL, NTH, SJL, JHP, SWK. Made critical revisions and approved final version: WJL, NPL, NTH, SJL, JHP, SWK. All authors reviewed and approved the final manuscript.

Supplementary Material

Supplementary File 1. The results from gene set analysis and five common significant gene sets.

Supplementary File 2. The results of leave-one-out cross-validation (LOOCV). T indicates that sample from aggressive cancer cell line or non-aggressive cell line was predicted to be an aggressive cancer cell line or non-aggressive cell line, respectively. F indicates that sample from aggressive cancer cell or non-aggressive cell line was predicted to be an non-aggressive cancer cell line aggressive cell line, respectively.

REFERENCES

1. Haggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg*. 2009;22(4):191–7.
2. Labianca R, Beretta GD, Kildani B, et al. Colon cancer. *Crit Rev Oncol Hematol*. 2010;74(2):106–33.
3. Boyle P, Langman JS. ABC of colorectal cancer: epidemiology. *BMJ*. 2000; 321(7264):805–8.
4. Bishehsari F, Mahdavinia M, Vacca M, Malekzadeh R, Mariani-Costantini R. Epidemiological transition of colorectal cancer in developing countries: environmental factors, molecular pathways, and opportunities for prevention. *World J Gastroenterol*. 2014;20(20):6055–72.
5. Labianca R, Merelli B. Screening and diagnosis for colorectal cancer: present and future. *Tumori*. 2010;96(6):889–901.
6. De Rosa M, Pace U, Rega D, et al. Genetics, diagnosis and management of colorectal cancer (Review). *Oncol Rep*. 2015;34(3):1087–96.
7. Hundt S, Haug U, Brenner H. Blood markers for early detection of colorectal cancer: a systematic review. *Cancer Epidemiol Biomarkers Prev*. 2007;16(10):1935–53.
8. Jain KK. Cancer biomarkers: current issues and future directions. *Curr Opin Mol Ther*. 2007;9(6):563–71.
9. Reimers MS, Zeestraten EC, Kuppen PJ, Liefers GJ, van de Velde CJ. Biomarkers in precision therapy in colorectal cancer. *Gastroenterol Rep*. 2013;1(3): 166–83.

10. García-Bilbao A, Armañanzas R, Ispizua Z, et al. Identification of a biomarker panel for colorectal cancer diagnosis. *BMC Cancer*. 2012;12(1):1–13.
11. Makrilia N, Kollias A, Manolopoulos L, Syrigos K. Cell adhesion molecules: role and clinical significance in cancer. *Cancer Invest*. 2009;27(10):1023–37.
12. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.
13. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat Methods Med Res*. 2016;25(1):472–87.
14. Lee WJ, Kim SC, Yoon JH, et al. Meta-analysis of tumor stem-like breast cancer cells using gene set and network analysis. *PLoS One*. 2016;11(2):e0148818.
15. Wu S, Xu Y, Feng Z, Yang X, Wang X, Gao X. Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. *BMC Bioinformatics*. 2012;13(1):1–12.
16. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*. 2002;8(8):816–24.
17. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*. 2009;10(1):1–17.
18. Fanayan S, Smith JT, Lee LY, et al. Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. *J Proteome Res*. 2013;12(4):1732–42.
19. Pradet-Balade B, Boulme F, Beug H, Mullner EW, Garcia-Sanz JA. Translation control: bridging the gap between genomics and proteomics? *Trends Biochem Sci*. 2001;26(4):225–9.
20. Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M, Quattrone A. Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis*. 2006;27(7):1323–33.
21. Carter DA, Dick AD, Mayer EJ. CD133+ adult human retinal cells remain undifferentiated in leukaemia inhibitory factor (LIF). *BMC Ophthalmol*. 2009;9:1.
22. Sagiv E, Starr A, Rozovski U, et al. Targeting CD24 for treatment of colorectal and pancreatic cancer by monoclonal antibodies or small interfering RNA. *Cancer Res*. 2008;68(8):2803–12.
23. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
24. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2008;4(1):44–57.
26. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*. 2002;99(10):6567–72.
27. Chu THaRTaBNaG. Pam: prediction analysis for microarrays. 2014.
28. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004;20(1):93–9.
29. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*. 2012;12(5):323–34.
30. Hwang WL, Yang MH, Tsai ML, et al. SNAIL regulates interleukin-8 expression, stem cell–like activity, and tumorigenicity of human colorectal carcinoma cells. *Gastroenterology*. 2011;141(1):279.e–291.e.
31. Fan F, Samuel S, Evans KW, et al. Overexpression of Snail induces epithelial–mesenchymal transition and a cancer stem cell–like phenotype in human colorectal cancer cells. *Cancer Med*. 2012;1(1):5–16.
32. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol*. 2014;15(7):465–81.
33. Leibeling D, Laspe P, Emmert S. Nucleotide excision repair and cancer. *J Mol Histol*. 2006;37(5–7):225–38.
34. Berndt SI, Platz EA, Fallin MD, Thuita LW, Hoffman SC, Helzlsouer KJ. Genetic variation in the nucleotide excision repair pathway and colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2006;15(11):2263–9.
35. Boyer AS, Walter D, Sorensen CS. DNA replication and cancer: from dysfunctional replication origin activities to therapeutic opportunities. *Semin Cancer Biol*. 2016;3(7–38):16–25.
36. Korkaya H, Liu S, Wicha MS. Regulation of cancer stem cells by cytokine networks: attacking cancer's inflammatory roots. *Clin Cancer Res*. 2011;17(19):6125–9.
37. Weinstein RS, Pauli BU. Cell junctions and the biological behaviour of cancer. *Ciba Found Symp*. 1987;125:240–60.
38. Martin TA, Mason MD, Jiang WG. Tight junctions in cancer metastasis. *Front Biosci (Landmark Ed)*. 2011;16:898–936.
39. Knights AJ, Funnell AP, Crossley M, Pearson RC. Holding tight: cell junctions and cancer spread. *Trends Cancer Res*. 2012;8:61–9.
40. Cheng Y, Xie G, Chen T, et al. Distinct urinary metabolic profile of human colorectal cancer. *J Proteome Res*. 2012;11(2):1354–63.
41. Haidich AB. Meta-analysis in medical research. *Hippokratia*. 2010;14(suppl 1):29–37.
42. Hung BT, Long NP, Hung LP, et al. Research trends in evidence-based medicine: a joinpoint regression analysis of more than 50 years of publication data. *PLoS One*. 2015;10(4):e0121054.
43. Peng Y, Li X, Wu M, et al. New prognosis biomarkers identified by dynamic proteomic analysis of colorectal cancer. *Mol Biosyst*. 2012;8(11):3077–88.
44. Du S, Guan Z, Hao L, et al. Fructose-bisphosphate aldolase a is a potential metastasis-associated marker of lung squamous cell carcinoma and promotes lung cell tumorigenesis and migration. *PLoS One*. 2014;9(1):e85804.
45. Kim YS, Hwan JD, Bae S, Bae DH, Shick WA. Identification of differentially expressed genes using an annealing control primer system in stage III serous ovarian carcinoma. *BMC Cancer*. 2010;10:576.
46. Zhong X, Chen X, Guan X, et al. Overexpression of G9a and MCM7 in oesophageal squamous cell carcinoma is associated with poor prognosis. *Histopathology*. 2015;66(2):192–200.
47. Liu Y-Z, Jiang Y-Y, Hao JJ, et al. Prognostic significance of MCM7 expression in the bronchial brushings of patients with non-small cell lung cancer (NSCLC). *Lung Cancer*. 2012;77(1):176–82.
48. Goswami CP, Nakshatri H. PROGgene: gene expression based survival analysis web application for multiple cancers. *J Clin Bioinforma*. 2013;3(1):1–9.
49. Reid JF, Gariboldi M, Sokolova V, et al. Integrative approach for prioritizing cancer genes in sporadic colon cancer. *Genes Chromosomes Cancer*. 2009;48(11):953–62.
50. Vyas S, Chang P. New PARP targets for cancer therapy. *Nat Rev Cancer*. 2014;14(7):502–9.
51. Ikeda Y, Kiyotani K, Yew PY, et al. Germline PARP4 mutations in patients with primary thyroid and breast cancers. *Endocr Relat Cancer*. 2016;23(3):171–9.
52. Xia W, Chen W, Zhang Z, et al. Prognostic value, clinicopathologic features and diagnostic accuracy of interleukin-8 in colorectal cancer: a meta-analysis. *PLoS One*. 2015;10(4):e0123484.
53. Chen D, Li W, Liu S, et al. Interleukin-23 promotes the epithelial-mesenchymal transition of oesophageal carcinoma cells via the Wnt/beta-catenin pathway. *Sci Rep*. 2015;5:8604.
54. Li W, Zhou Y, Su Y, et al. IL-23 promotes invasion of esophageal squamous cell carcinoma cells by activating DLL4/Notch1 signaling pathway. *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi*. 2015;31(6):812–5,820.
55. Adamo V, Franchina T, Minciullo PL, et al. Role of interleukin-23 circulating levels increase in resected colorectal cancer before and after chemotherapy: preliminary data and future perspectives. *J Cell Physiol*. 2011;226(11):3032–4.
56. Lopez-Otin C, Matrisian LM. Emerging roles of proteases in tumour suppression. *Nat Rev Cancer*. 2007;7(10):800–8.
57. Dimberg J, Strom K, Lofgren S, Zar N, Hugander A, Matussek A. Expression of the serine protease inhibitor serpinA3 in human colorectal adenocarcinomas. *Oncol Lett*. 2011;2(3):413–8.
58. Vasiliou V, Vasiliou K, Nebert DW. Human ATP-binding cassette (ABC) transporter family. *Hum Genomics*. 2009;3(3):281–90.
59. Januchowski R, Zawierucha P, Ruciński M, et al. Drug transporter expression profiling in chemoresistant variants of the A2780 ovarian cancer cell line. *Biomed Pharmacother*. 2014;68(4):447–53.
60. Pagnotta SM, Laudanna C, Pancione M, et al. Ensemble of gene signatures identifies novel biomarkers in colorectal cancer activated through PPARgamma and TNFalpha signaling. *PLoS One*. 2013;8(8):e72638.