

**This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.**

**Author(s):** Sievänen, Tero; Jokela, Tiina; Hyvärinen, Matti; Korhonen, Tia-Marje; Pylvänäinen, Kirsi; Mecklin, Jukka-Pekka; Karvanen, Juha; Sillanpää, Elina; Seppälä, Toni T.; Laakkonen, Eija K.

**Title:** Circulating microRNA signature predicts cancer incidence in Lynch syndrome : a pilot study

**Year:** 2024

**Version:** Published version

**Copyright:** © 2024 The Authors; Published by the American Association for Cancer Research

**Rights:** CC BY-NC-ND 4.0

**Rights url:** <https://creativecommons.org/licenses/by-nc-nd/4.0/>

**Please cite the original version:**

Sievänen, T., Jokela, T., Hyvärinen, M., Korhonen, T.-M., Pylvänäinen, K., Mecklin, J.-P., Karvanen, J., Sillanpää, E., Seppälä, T. T., & Laakkonen, E. K. (2024). Circulating microRNA signature predicts cancer incidence in Lynch syndrome : a pilot study. *Cancer Prevention Research, 17*(6), 243-254. <https://doi.org/10.1158/1940-6207.capr-23-0368>



# Circulating miRNA Signature Predicts Cancer Incidence in Lynch Syndrome—A Pilot Study

Tero Sievänen<sup>1</sup>, Tiina Jokela<sup>1</sup>, Matti Hyvärinen<sup>1</sup>, Tia-Marje Korhonen<sup>1</sup>, Kirsi Pylvänäinen<sup>2</sup>, Jukka-Pekka Mecklin<sup>2,3</sup>, Juha Karvanen<sup>4</sup>, Elina Sillanpää<sup>1,2</sup>, Toni T. Seppälä<sup>5,6,7,8</sup>, and Eija K. Laakkonen<sup>1</sup>

## ABSTRACT

Lynch syndrome (LS) is the most common autosomal dominant cancer syndrome and is characterized by high genetic cancer risk modified by lifestyle factors. This study explored whether a circulating miRNA (c-miR) signature predicts LS cancer incidence within a 4-year prospective surveillance period. To gain insight how lifestyle behavior could affect LS cancer risk, we investigated whether the cancer-predicting c-miR signature correlates with known risk-reducing factors such as physical activity, body mass index (BMI), dietary fiber, or NSAID usage. The study included 110 c-miR samples from LS carriers, 18 of whom were diagnosed with cancer during a 4-year prospective surveillance period. Lasso regression was utilized to find c-miRs associated with cancer risk. Individual risk sum derived from the chosen c-miRs was used to develop a model to predict LS cancer incidence. This model was validated using 5-fold cross-validation. Correlation and pathway analyses were applied to inspect biological functions of c-miRs. Pearson correlation was used to examine the associations of

c-miR risk sum and lifestyle factors. hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615 were identified as cancer predictors by Lasso, and their risk sum score associated with higher likelihood of cancer incidence (HR 2.72, 95% confidence interval: 1.64–4.52, C-index = 0.72). In cross-validation, the model indicated good concordance with the average C-index of 0.75 (0.6–1.0). Coregulated hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p targeted genes involved in cancer-associated biological pathways. The c-miR risk sum score correlated with BMI ( $r = 0.23$ ,  $P < 0.01$ ). In summary, BMI-associated c-miRs predict LS cancer incidence within 4 years, although further validation is required.

**Prevention Relevance:** The development of cancer risk prediction models is key to improving the survival of patients with LS. This pilot study describes a serum miRNA signature-based risk prediction model that predicts LS cancer incidence within 4 years, although further validation is required.

## Introduction

Lynch syndrome (LS) is the most common inherited cancer predisposition syndrome, with an estimated prevalence of

1:300 (1, 2). Distinct LS phenotypes are caused by germline mutations in DNA mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6*, and *PMS2* (2). The impaired MMR manifests as an increased risk of multiple cancers, and depending on the cancer type, the risk is modified by lifestyle factors such as physical activity, body weight, consumption of dietary resistant starch, and NSAID usage (2–8). LS cancer spectrum includes various cancer types, colorectal cancer and endometrial cancers being most common (6). As the cancer risk varies greatly among pathogenic MMR variant carriers (6), it is pivotal to innovate risk stratification biomarkers that could be used to identify LS carriers who may develop cancer in the near future.

Circulating miRNAs (c-miR) are short, noncoding RNA molecules that function as intercellular messengers by migrating throughout the body (9). They play a crucial role in cancer biology by regulating core cellular processes, such as proliferation and apoptosis, through the suppression of target gene translation (10). Multiple studies have reported c-miRs as potential biomarkers for various sporadic cancers (11–15) by demonstrating differential expression (DE) between the c-miR signatures of patients with cancer and healthy controls. In most of these prior studies, the analysis of c-miR signatures has been limited to patients who have already received a colorectal cancer diagnosis, making it challenging to ascertain their

<sup>1</sup>Gerontology Research Center and Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland. <sup>2</sup>The wellbeing services county of Central Finland, Jyväskylä, Finland. <sup>3</sup>Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland. <sup>4</sup>Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland. <sup>5</sup>Applied Tumor Genomics Research Program, University of Helsinki, Helsinki, Finland. <sup>6</sup>Department of Abdominal Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland. <sup>7</sup>Department of Gastroenterology and Alimentary Tract Surgery and TAYS Cancer Centre, Tampere University Hospital, Tampere, Finland. <sup>8</sup>Faculty of Medicine and Health Technology, University of Tampere, Tampere, Finland.

T.T. Seppälä and E.K. Laakkonen contributed equally to this article.

**Corresponding Authors:** Tero Sievänen, Gerontology Research Center and Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä 40700, Finland. E-mail: tero.o.sievenen@jyu.fi; and Eija K. Laakkonen, eija.k.laakkonen@jyu.fi

Cancer Prev Res 2024;17:1–12

doi: 10.1158/1940-6207.CAPR-23-0368

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

©2024 The Authors; Published by the American Association for Cancer Research

potential utility in risk stratification. Interestingly, a recent study by Raut and colleagues showed that altered c-miR expression could predict sporadic colorectal cancer incidence several years prior the diagnosis (16). However, it has remained unclear whether this observation extends to LS.

In LS, the risk of various cancers is significantly elevated by sedentary behavior and excess body weight, while physical activity, maintaining a healthy body weight, and the consumption of dietary resistant starch and NSAIDs have been shown to mitigate these risks (3, 7, 8). Although it is well acknowledged that adopting an optimal lifestyle can reduce cancer incidence, the underlying molecular mechanisms remain less elucidated. c-miRs, due to their capacity to modulate pathophysiologic responses to changing lifestyle behaviors (9) and their ability to exhibit DE profiles between sedentary and physically active individuals (17), offer potential insights into how lifestyle behaviors influence LS-associated cancer risk.

We were first to report that the c-miR signature of cancer-free LS carriers is associated with carcinogenesis by displaying aberrant expression compared with healthy population but similar expression when compared with patients with sporadic rectal cancer (18). To build on that, the primary aim of this study was to investigate whether c-miRs can be used in LS cancer risk prediction during a 4-year prospective surveillance period. Considering the modulatory role of c-miRs in lifestyle habits, our secondary aim was to explore whether any of the LS cancer predictive c-miRs are associated with physical activity, body weight, dietary fiber, or NSAID usage.

## Materials and Methods

The study flow chart and general outline is detailed in **Fig. 1**.

### Patients and sample collection

The clinical data of our study were derived from the nationwide Finnish Lynch Syndrome Research Registry (LSRFi, [www.lynchsyndrooma.fi](http://www.lynchsyndrooma.fi), accessed November 2022). Age, sex, MMR mutation status, family cancer history, and all cancer diagnoses with the cancer type and date of each diagnosis were confirmed from hospital medical records and national cancer registries upon recording in the LSRFi. To date, LSRFi includes 1,800 LS carriers from 400 families and contains clinicopathologic information on all cancers of the registered individuals. In the current study, we reviewed baseline medical records of Finnish cancer-free LS carriers whose c-miR expression profile was characterized ( $n = 110$ ). Ethnicity throughout the study population was White Caucasian.

LS carriers were enrolled in the study, and whole blood was collected at their regular colonoscopy surveillance appointments at Helsinki University Central Hospital in Helsinki and Central Finland Central Hospital in Jyväskylä, Finland. Non-LS control samples were acquired from Biobank of Eastern Finland, Kuopio, and a previously studied Estrogenic Regulation of Muscle Apoptosis cohort consisting of healthy 47–55 years old women. To separate serum, the whole blood samples were

allowed to clot for 30 minutes at room temperature, centrifuged at  $1,800 \times g$  for 10 minutes and aliquoted. Methods of sample collection, preanalytic preparation, c-miR extraction, library preparation, and sequencing have been described previously in detail (18).

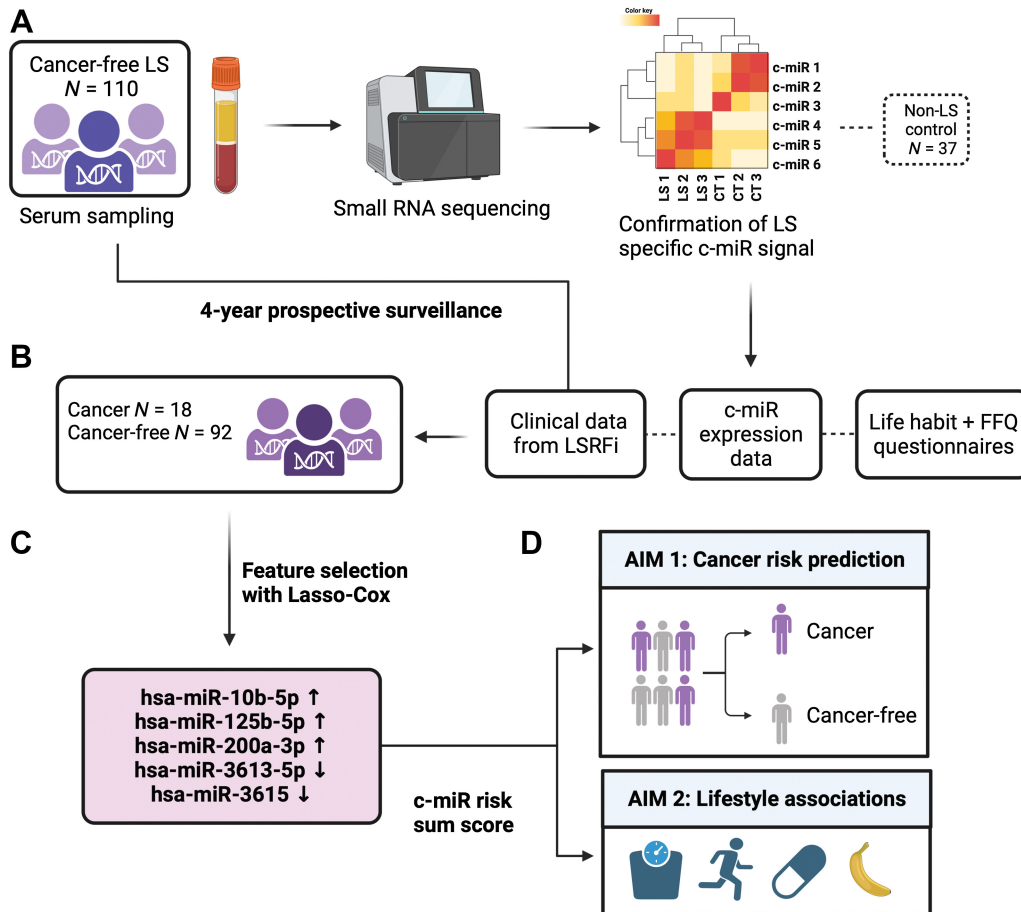
### Data collection and ethical issues

High-throughput c-miR expression data of cancer-free LS carriers ( $n = 86$ ) as well as of healthy non-carrier control samples ( $n = 37$ ) were generated as described earlier (18). Briefly, c-miRs were extracted using affinity column-based approach (miRNeasy Serum/Plasma advanced kit, Qiagen), ligated to sequencing adapters from both 5' and 3' end, reverse transcribed into cDNA using unique molecular identifier (UMI)-assigning primers, and purified with magnetic beads (Qiaseq miRNA Library preparation kit, Qiagen). Sequencing of the c-miR libraries were done with NextSeq 500 (Illumina) using NextSeq 500/550 High Output Kit v. 2.5 with 75 cycles aiming for depth of 5M reads per sample. Quality controls throughout the RNA isolation, library preparation, and sequencing protocols were conducted with qRT-PCR (Bio-Rad), TapeStation 4200 (Agilent) and Qubit fluorometer (Invitrogen), respectively. To increase the cohort size, we performed small RNA sequencing (RNA-seq) experiment on additional 24 LS carriers using the same analysis pipeline as described (18). Thus, the current study composed of 110 cancer-free LS carriers who are registered in the LSRFi and 37 healthy non-carrier control samples. Healthy non-carrier control samples were included only in the DE analysis to confirm previously reported LS-associated c-miR signature.

The corresponding lifestyle data of the cancer-free LS carriers in the current study were collected as described previously in detail (3). Briefly, questionnaires for anthropometric, socioeconomic, and lifestyle data collection were sent to adult Finnish LS carriers whose contact information was available in LSRFi in 2017 and 2020. Alongside with the lifestyle data collection, dietary habits data of the same persons were collected by a validated semiquantitative food frequency questionnaire (19). The average time period between the questionnaires' data collection and blood sampling was 2.0 (0.3–3.9) years. A written informed consent was obtained from all participants, and the Helsinki and Uusimaa Health Care District (HUS/155/2021) and Central Finland Health Care District Ethics Committee (KSSHP D# 1U/2018 and 1/2019 and KSSHP 3/2016) approved the study protocol. The study was conducted according to the guidelines of the Declaration of Helsinki.

### Missing data

There were no missing c-miR data. Missing lifestyle and dietary data [physical activity: 30.9%; body mass index (BMI): 4.5%; dietary fiber intake: 29.0% and NSAID usage: 29.0%; Supplementary Table S1] occurred due to incomplete questionnaire responses. Missing data were assumed to occur at random and multiple imputation with 50 iterations was used to create and analyze 50 multiply imputed datasets using mice"



**Figure 1.**

The study flow chart and general outline. **A**, Serum samples of 110 cancer-free LS carriers and 37 non-carrier controls were sequenced to confirm previously identified LS specific c-miR signature. **B**, LS clinical data were derived from LSRFi to assess the cancer status of cancer-free LS carriers after 4 years of prospective surveillance. Of the 110 cancer-free LS carriers, 18 had developed cancer during the surveillance period. Then, c-miR expression data were matched with the corresponding clinical and lifestyle data to investigate whether the c-miR signature can predict LS cancer risk during the surveillance. Lifestyle data were collected 2017 or 2020 with a questionnaire. Blood sample was taken at regular colonoscopy visit between 2018 and 2020. The average time period between lifestyle data collection and blood sample was 2.0 (0.3–3.9) years. **C**, Lasso-regularized Cox regression was used to select the most important predictor c-miRs from the entire cohort. Lasso-obtained c-miRs were used to compute c-miR risk sum score. Arrows indicate upregulation (↑) and downregulation (↓) of the c-miR in LS carriers when compared with healthy non-carriers. **D**, c-miR risk sum score was used in LS cancer risk prediction with 5-fold cross-validation and to inspect associations with lifestyle data. c-miR = circulating miRNA; Lasso = Least absolute shrinkage and selection method; LS = Lynch syndrome; LSRFi = Finnish Lynch Syndrome Research Registry. This figure was created with BioRender.com.

R-package (20) with default settings. All lifestyle variables as well as sex, age, pathogenic MMR-variant, cancer status, and c-miR expression were used for imputation of each lifestyle variable, and results were pooled using “pool” function in “mice”.

### DE analysis

DE analysis between cancer-free LS carriers and healthy non-carrier controls was performed with “DESeq2” R-package (ref. 21; RRID:SCR\_000154) using raw c-miR counts (Supplementary Materials and Methods S1). Sex and sequencing batch were added to the DE analysis design formula to account for their potential confounding effect. Normalization and variance stabilization transformations were done with DESeq2 by applying median of ratios method (21) and “rlog” function, respec-

tively. Low count c-miRs were filtered prior to DE analysis. Filtering was done with “filterByExpr” function in “edgeR” R-package (22) that excluded c-miRs with <1 count per million in 70% of samples. Benjamini–Hochberg procedure with FDR 0.05 was used to correct for multiple testing. Hierarchical clustering based on Euclidean distances and the “complete” method was applied to verify DE findings. “hclust” function in “stats” base R-package was used for the hierarchical clustering analysis.

### Covariates

#### c-miRs

c-miR expression data were derived from small RNA-seq experiments and measured as counts relative to sample library size where counts represent molecules in blood serum. DESeq2

normalized and variance stabilized c-miR counts were used for all analyses.

### Physical activity

Physical activity was assessed by a self-reported questionnaire. The questionnaire included four questions about the frequency, intensity, and duration of leisure time physical activity and commuting activity. On the basis of the responses, the metabolic equivalent task hours per day for leisure time physical activity was calculated.

### BMI

Body weight and height were measured by the clinician during the study subjects' regular colonoscopy appointment. If body weight and height information were missing, we used the last known self-measured weight and height measurement. BMI was calculated as weight in kilograms divided by the height squared in meters ( $\text{kg}/\text{m}^2$ ) according to World Health Organization guidelines.

### Dietary fiber

Dietary fiber including resistant starch amount was derived from self-reported food frequency questionnaire and assessed as grams per day.

### NSAID usage

Study subjects self-reported whether (yes/no) they used NSAIDs, such as acetylsalicylic acid, ibuprofen or ketoprofen products frequently.

### Construction and validation of the LS cancer risk prediction model

Least absolute shrinkage and selection method (Lasso; ref. 23) regularized Cox regression was used to find predictor c-miRs from the pool of identified LS-associated DE c-miRs using the entire study sample. Optimal value for the Lasso regularization parameter lambda was chosen with 10-fold cross-validation. The expression levels of the Lasso-obtained c-miRs were used to compute an individual risk sum score (linear predictor) for all the participants by using formula:

$$\text{Risk sum score} = \text{Expr}(\text{miR}_A) * \beta(\text{miR}_A) + \text{Expr}(\text{miR}_B) * \beta(\text{miR}_B) \dots,$$

where  $\text{Expr}(\text{miR})$  represents the normalized and variance stabilized c-miR expression and  $\beta(\text{miR})$  indicates the regression coefficient in Lasso-Cox regression model (16). By using univariate and multivariate Cox regression models, the c-miR risk sum score was then applied to predict the risk of cancer incidence. We used the entire study sample ( $n = 110$ ) for fitting the risk prediction model. The predictive performance of the risk prediction model was validated with 5-fold cross-validation and the model concordance evaluated with Harrel C-index (scale 0.5–1.0) where 0.5 indicates poor performance and 1.0 indicates excellent performance (ref. 24; Supplementary Materials and Methods S1).

The surveillance time used for risk prediction was determined from the timepoint of initial serum sampling (2018–2020) until the latest update of LSRFi (November 2022). The response variable in the risk prediction model was the age at the time of cancer diagnosis (event) or the age at the final update date of LSRFi (censoring). HR and 95% confidence intervals (CI) of the c-miR risk sum score were estimated for unadjusted model as well as for sex and MMR-variant adjusted model. Proportional hazards assumption was tested using Schoenfeld residuals (Supplementary Fig. S1). Regarding the risk prediction model development and validation, we followed Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) reporting checklist (25). We used “glmnet” R-package (26) for the cross-validation procedure as well as for Lasso-regularized Cox regression. “survival” R-package was used for Cox regression modeling (27).

### Pathway analysis

We identified potential targets genes of the Lasso-obtained c-miRs from miRTarBase (28) by using miRWalk online tool (29). We considered only the genes with experimental validation in MiRTarBase (28) to exclude low evidence targets. The obtained target gene list was applied to overrepresentation analysis with hypergeometric tests using Search Tool for Retrieval of Interacting Genes/Proteins (STRING; ref. 30) and Reactome (31) databases.

### Statistical analysis

All statistical analyses were performed in R-programming environment (v.4.2.2) using RStudio and in-house R-scripts. Levene test was used to inspect homoscedasticity. Study subject characteristics are presented as means and SDs for continuous variables and as number of study subjects and percentages for categorical variables. Regarding **Table 1**, Welch two-sample  $t$  test was used for continuous variables whereas  $\chi^2$  test was applied for categorical variables. Because of skewed nature of RNA-seq data, Spearman method was applied to inspect correlations between the Lasso-obtained c-miRs. Pearson method was applied to examine correlations between the multiple imputed lifestyle data and c-miR risk sum score.

### Data availability

The sequence data generated in this study are publicly available in Sequence Read Archive (SRA) at PRJNA1088397.

## Results

### Study subject characteristics

The study subjects' clinical characteristics are described in **Table 1**. Most had a pathogenic *MLH1* germline variant (67.3%) followed by *MSH2* (17.3%), *MSH6* (13.6), and *PMS2* (1.8), respectively. Of the 110 study subjects, 18 (13 males and 5 females) developed cancer during the prospective surveillance. The mean surveillance time for those who developed cancer was 1.3 years whereas for those who remained cancer-free it

**Table 1.** Study subject characteristics.

Parameter	Total cohort	Cancer during surveillance	Cancer-free after surveillance	P-value
N (%)	110	18 (16.4)	92 (83.6)	
Sex, N (%)				0.071
Male	57 (51.8)	13 (72.2)	44 (47.8)	
Female	53 (48.2)	5 (27.8)	48 (52.2)	
MMR status, N (%)				0.777
MLH1	74 (67.3)	14 (77.8)	60 (65.2)	
MSH2	19 (17.3)	2 (11.1)	17 (18.5)	
MSH6	15 (13.6)	2 (11.1)	13 (14.1)	
PMS2	2 (1.8)	—	2 (2.2)	
Physical activity, MET/hours/day (SD) <sup>a</sup>	4.4 (± 4.5)	7.6 (± 7.2)	3.7 (± 3.6)	0.094
BMI, kg/m <sup>2</sup> (SD) <sup>a</sup>	27.8 (± 5.8)	27.9 (± 4.4)	27.7 (± 6.1)	0.875
Dietary fiber intake, g/day <sup>a</sup> (SD)	23.4 (± 10.0)	21.1 (± 9.8)	23.9 (± 10.0)	0.379
NSAID usage, N (%) <sup>a</sup>				0.736
Yes	26 (33.3)	3 (25.0)	23 (34.8)	
No	52 (66.7)	9 (75.0)	43 (65.2)	
Age at the start of surveillance, years <sup>a</sup> (SD)	57.5 (± 11.8)	57.6 (± 14.3)	57.7 (± 11.4)	0.967
Age at the end of surveillance, years <sup>a</sup> (SD)	60.7 (± 12.0)	58.9 (± 14.4)	61.0 (± 11.5)	0.575
Surveillance time, years <sup>a</sup> (SD)	3.1 (± 1.1)	1.3 (± 1.1)	3.5 (0.6)	<0.001
Cancer history, N (%)				0.636
Yes	54 (49.1)	10 (55.6)	44 (47.8)	
No	56 (50.9)	8 (44.4)	48 (52.2)	
Cancer, N (%)	18 (16.4)	18 (16.4)	—	
CRC	9 (50.0)	9 (50.0)	—	
Other <sup>b</sup>	9 (50.0)	9 (50.0)	—	

Abbreviations: BMI: body mass index; CRC: colorectal cancer; MET: metabolic equivalent task; MMR: mismatch-repair gene; NSAID: non-steroidal anti-inflammatory drug; SD: standard deviation.

<sup>a</sup>Missing values, total cohort: Physical activity,  $n = 34$ ; BMI,  $n = 5$ ; dietary fiber intake,  $n = 32$ ; NSAID usage,  $n = 32$ . Missing values, cancer: Physical activity,  $n = 6$ ; dietary fiber intake,  $n = 6$ ; NSAID usage,  $n = 6$ . Missing values, cancer-free: Physical activity,  $n = 28$ ; BMI,  $n = 5$ ; dietary fiber intake,  $n = 26$ ; NSAID usage,  $n = 26$ .

<sup>b</sup>Other cancers included bladder cancer ( $n = 1$ ), breast cancer ( $n = 1$ ), esophageal cancer ( $n = 1$ ), glioma ( $n = 1$ ), gastric cancer ( $n = 1$ ), prostate cancer ( $n = 3$ ), and pinocellular cancer ( $n = 1$ ).

was 3.5 years. Half of the diagnosed cancers were colorectal cancers and the other half consisted of several other cancer types (Supplementary Table S2). No loss to follow-up occurred.

### Confirmation of LS-associated c-miR signature

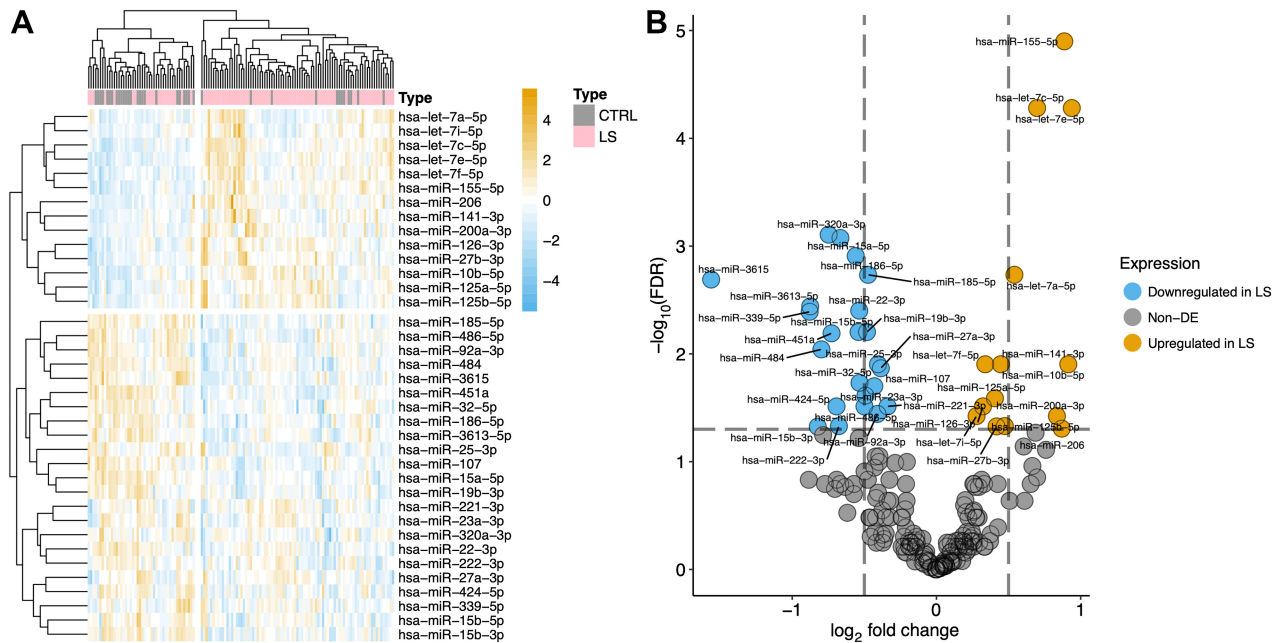
We have previously identified a c-miR signature that distinguished LS carriers from healthy non-carrier population. However, as the LS cohort used in the present study included 24 new cases, we reprocessed the data to seek for more LS-associated c-miRs and to verify our previous finding. DE analysis resulted in 37 DE c-miRs between cancer-free LS carriers and healthy non-carrier controls (Fig. 2A; Supplementary Table S3). We found 14 upregulated DE c-miRs and 23 downregulated DE c-miRs (Fig. 2B). These 37 DE c-miRs were confirmed as LS-associated and thus chosen for the downstream analyses.

### The expression levels of hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615 are associated with increased risk of cancer incidence

Several multi c-miR panels have been reported to have predictive or prognostic value in sporadic cancer risk assessment. Thus, we wanted to investigate whether the expression of

any of the LS-associated DE c-miRs showed potential in LS cancer risk prediction during the prospective surveillance. Out of the 37 DE c-miRs, Lasso selected hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615 as the best predictors that separated those LS carriers who developed cancer from those who remained cancer-free during the surveillance (Fig. 3A). The expression of all these c-miRs was higher in those LS carriers who developed cancer during the surveillance compared with those LS carriers who remained cancer-free, although only hsa-miR-3613-5p displayed statistical significance (Fig. 3B). Of them, only hsa-miR-10b-5p was independently associated with an increased cancer risk (HR 6.58, 95% CI: 1.43–30.21,  $\beta = 1.88$ ; Supplementary Table S4). The full model showed good concordance (C-index = 0.72; Supplementary Table S4).

Because efficient miR-based biological regulation relies on additive effects of multiple miRs (32), we wanted to investigate the pooled performance of the selected c-miRs on predicting LS cancer risk. We observed that c-miR risk sum score was significantly associated with increased risk of cancer incidence (HR 2.72, 95% CI: 1.64–4.52,  $\beta = 1.00$ , C-index = 0.72) also after adjusting for sex and MMR-variant (HR 2.71, 95% CI: 1.62–4.52,  $\beta = 1.00$ , C-index = 0.77; Fig. 3C). A 5-fold cross-validation of this risk prediction model resulted in average



**Figure 2.**

Confirmation of LS-associated c-miR signature. **A**, Heat map with hierarchical clustering of DE c-miRs ( $n = 37$ ) that separated cancer-free LS carriers and non-carrier controls. Orange color indicates c-miR upregulation in LS group whereas blue color indicates c-miR downregulation in LS group. The scale represents normalized and variance stabilized c-miR counts. LS samples are annotated with pink color and non-carrier controls with gray. **B**, Volcano plot of DE c-miRs that separated cancer-free LS carriers and non-carrier controls. Only the upregulated (orange) and downregulated (blue) c-miRs in LS group are annotated. Gray dots represent non-DE c-miRs. Y-axis indicates  $-\log_{10}$  FDR whereas X-axis indicate  $\log_2$  fold change of c-miR expression.

C-index of 0.75 (0.60–1.00; **Fig. 3D**) thus presenting good concordance (Supplementary Table S5). The mean c-miR risk sum score was higher in those LS carriers who developed cancer (mean = 44.0) compared with those who did not (mean = 43.1;  $P < 0.01$ ).

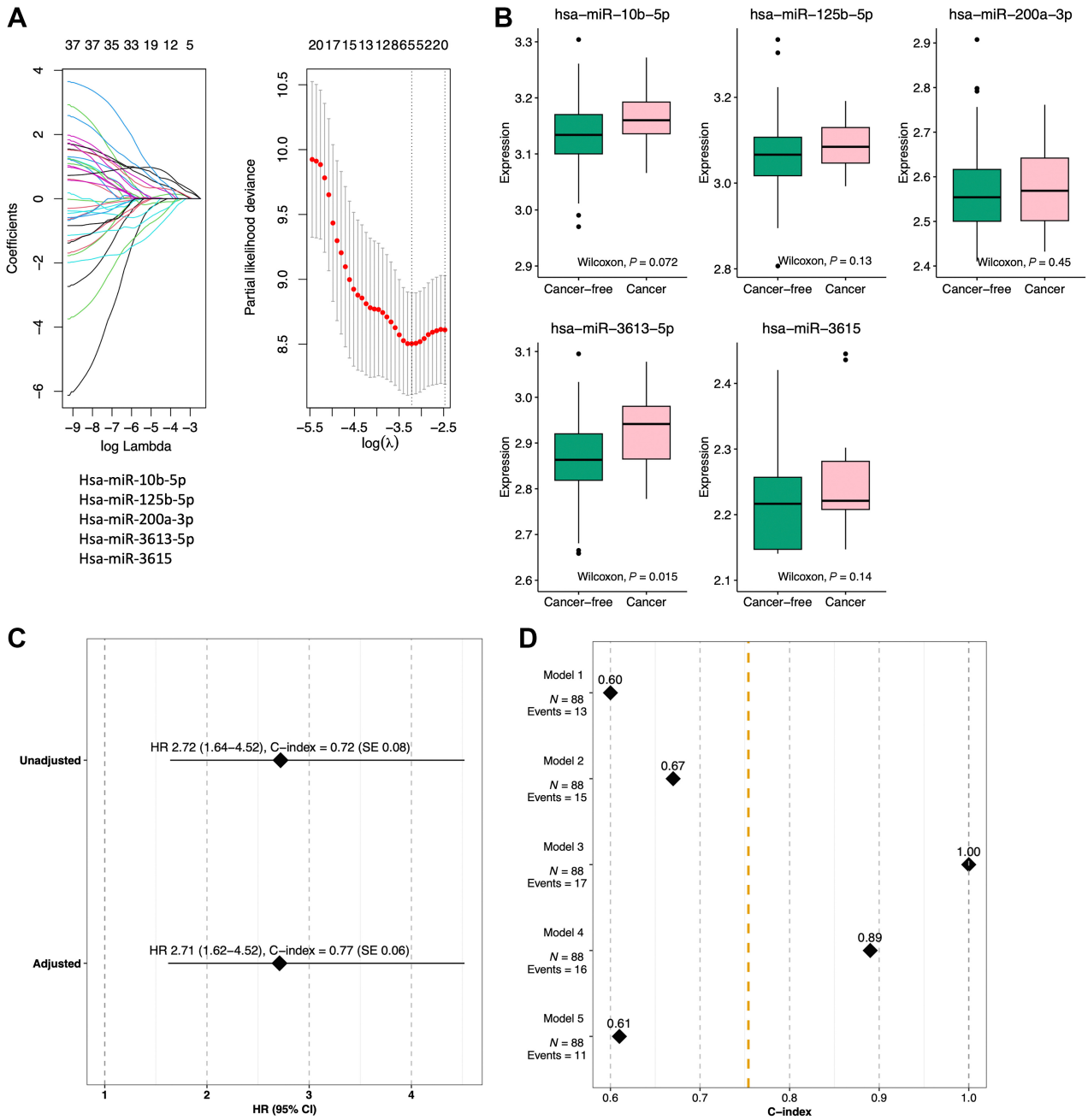
We also conducted two sensitivity analyses that included either *MLH1* carriers ( $N = 74$  of whom 14 developed cancer) or colorectal cancer cases ( $N = 101$  of whom 9 developed colorectal cancer; Supplementary Tables S6–S11). Lasso selected *hsa-let-7e-5p*, *hsa-miR-10b-5p*, and *hsa-miR-3613-5p* as the best predictors to separate those who developed cancer from those who did not in the *MLH1* subgroup. Regarding the colorectal cancer cases, *hsa-miR-10b-5p*, *hsa-miR-19b-3p*, *hsa-miR-200a-3p*, *hsa-miR-27b-3p*, and *hsa-miR-3615* were selected as the best predictors. Although a risk sum score in both sensitivity analyses was independently associated with increased cancer incidence after adjusting, an enhanced risk prediction performance was seen only in colorectal cancer-stratified model (C-index = 0.84) but not in *MLH1* model (C-index = 0.56) when compared with the unstratified model.

Taken together, risk prediction models composed of *hsa-miR-10b-5p*, *hsa-miR-125b-5p*, *hsa-miR-200a-3p*, *hsa-miR-3613-5p*, and *hsa-miR-3615* could classify between those LS carriers who developed cancer during the surveillance period and those who did not, also when stratified for *MLH1* or colorectal cancer. Higher prediagnostic expression levels of these c-miRs are associated with increased risk of cancer incidence.

### Pathway analysis links coregulated *hsa-miR-10b-5p*, *hsa-miR-125b-5p*, and *hsa-miR-200a-3p* to cell cycle regulation, programmed cell death, cellular senescence, and transcriptional regulation

The targeting of multiple genes within a specific pathway, as well as the additive effects of coregulated c-miR clusters, are key elements of effective c-miR regulation (32). First, we conducted a correlation analysis to inspect whether the Lasso-obtained c-miRs present possible coregulation. *Hsa-miR-10b-5p* correlated with *hsa-miR-200a-3p* ( $\rho = 0.28$ ,  $P < 0.01$ ) and with *hsa-miR-125b-5p* ( $\rho = 0.29$ ,  $P < 0.01$ ), *hsa-miR-200a-3p* correlated with *hsa-miR-125b* ( $\rho = 0.41$ ,  $P < 0.001$ ) whereas *hsa-miR-3613-5p* correlated only with *hsa-miR-3615* ( $\rho = 0.31$ ,  $P < 0.01$ ) thus displaying correlation and expression concordance (**Fig. 4A**). *hsa-miR-10b-5p*, *hsa-miR-125b-5p*, and *hsa-miR-200a-3p* were upregulated in LS whereas *hsa-miR-3613-5p* and *hsa-miR-3615* were downregulated when compared with the healthy non-carrier controls, respectively (**Fig. 2B**).

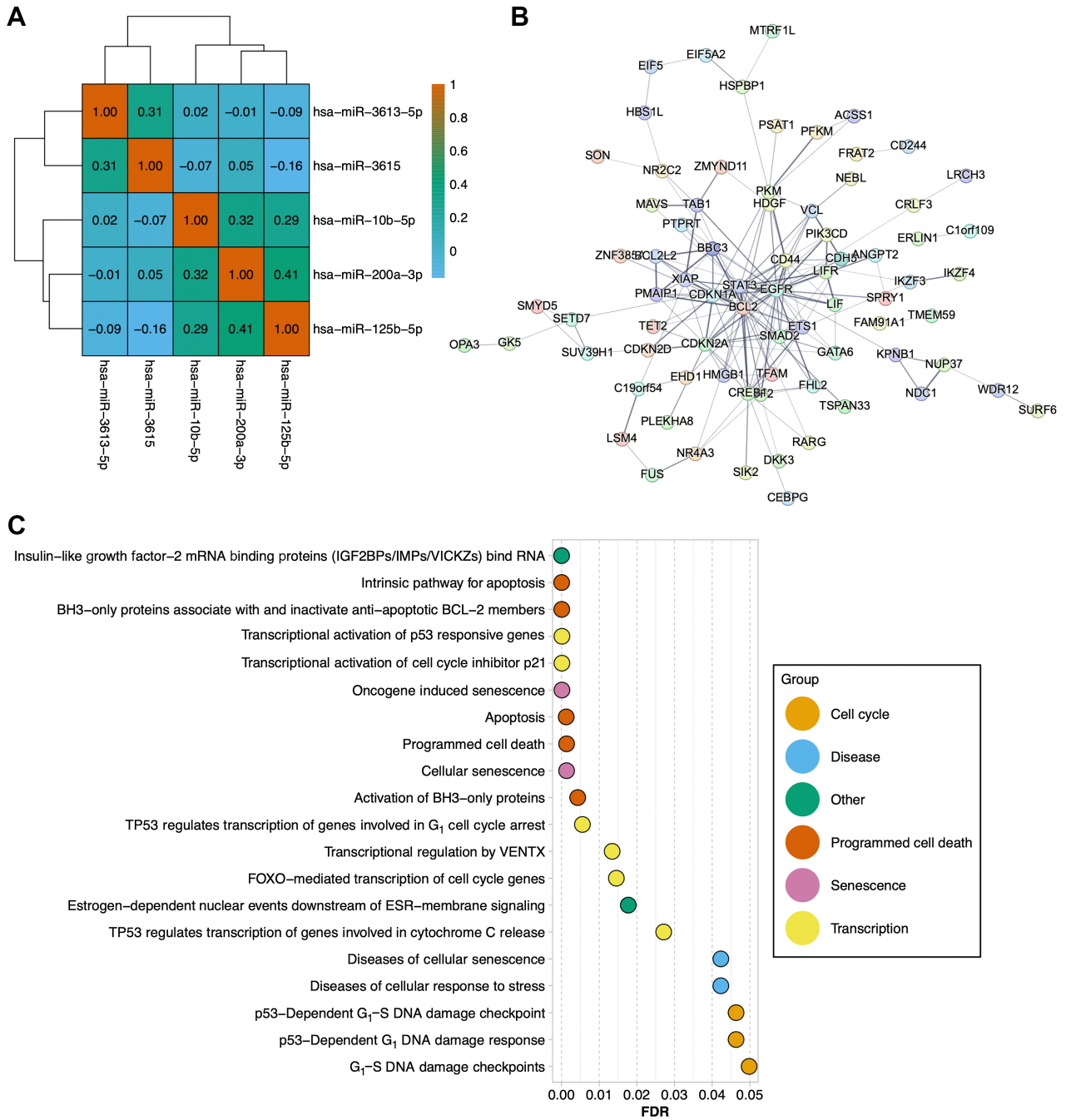
To gain insight on relevant biological processes of *hsa-miR-10b-5p*, *hsa-miR-125b-5p*, *hsa-miR-200a-3p*, *hsa-miR-3613-5p*, and *hsa-miR-3615*, we first predicted their putative target genes using miRWalk. We found 128 unique target genes for all the c-miRs expect for *hsa-miR-3613-5p* (Supplementary Table S12). The most important gene nodes are presented in **Fig. 4B**. These nodes had significant interactions among each other ( $P < 0.001$ ) which provided support for biological connection. Of them, *BCL2*, *EGFR*, *CDKN1A*, *CDKN2A*, *STAT3*, *SMAD2*, *CREB1*, *ETS1*, and *CD44* had the most



**Figure 3.**

hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615 are associated with increased risk of cancer incidence. **A**, Left panel presents a Lasso feature selection graph where every colored line indicate one of the 37 DE c-miRs found between cancer-free LS carriers and non-carrier controls. Regression coefficient is presented as a function of the tuning parameter lambda. Right panel shows the partial likelihood deviance as a function of lambda. The area between the dashed lines presents the optimal lambda value ( $\lambda = -2.5$ ) after 10-fold cross-validation. **B**, Boxplots present expression differences of the Lasso-obtained c-miRs between LS carriers who got cancer (pink) and cancer-free LS carriers (green). All of the Lasso-obtained c-miRs were upregulated in those who developed cancer during the surveillance. The expression values on the Y-axis are presented as normalized and variance stabilized c-miR counts. **C**, Unadjusted and sex and MMR-variant-adjusted LS cancer risk prediction models. HRs, 95% CIs, and model C-indices are shown. **D**, A total of 5-fold cross-validated LS cancer risk prediction model. Number of samples and events are shown for the training folds (80% of data) whereas C-indices are shown for the validation fold (20% of data). Orange color indicates the mean C-index (0.75) across all folds.





**Figure 4.**

Pathway analysis linked coregulated hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p to cell cycle regulation, programmed cell death, cellular senescence, and transcriptional regulation. **A**, Heat map of correlations among the Lasso-obtained c-miRs hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615. hsa-miR-10b-5p correlated with hsa-miR-125b-5p ( $P < 0.01$ ) and hsa-miR-200a-3p ( $P < 0.01$ ), hsa-miR-125b-5p correlated with hsa-miR-200a-3p ( $P < 0.01$ ) whereas hsa-miR-3613-5p correlated with hsa-miR-3615 ( $P < 0.01$ ).  $P < 0.05$  was considered significant. The scale represents the magnitude of correlation. Blue indicates low correlation and orange indicate high correlation. **B**, The most important gene nodes included *EGFR*, *CDKN1A*, *CDKN2A*, *STAT3*, *SMAD2*, *CREB1*, *ETS1*, and *CD44* and were observed to have significant interactions ( $P < 0.001$ ) with each other. Edge thickness indicates the strength of data support between the nodes. **C**, Pathway analysis of 86 experimentally verified target genes of hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p were significantly enriched ( $FDR < 0.05$ ) in several pathways linked to cell cycle regulation, programmed cell death, cellular senescence, and transcriptional regulation.

interactions. hsa-miR-10b-5p targeted tumor suppressor genes *CDKN1A*, *CDKN2A*, and *CREB1*, hsa-miR-125b-5p targeted oncogenes *BCL2* and *STAT3*, proto-oncogene *ETS1* as well as *CD44*, hsa-miR-200a-3p targeted oncogene *EGFR* and tumor suppressor gene *SMAD2*, that further supported possible coregulation of these c-miRs. The complete gene node map is presented in Supplementary Fig. S2.

Next, we conducted a pathway analysis on the experimentally confirmed c-miR target genes reported in MiRTarBase (28). A total of 86 out of 128 of the found target genes were significantly enriched in several pathways related to cell cycle regulation, programmed cell death, cellular senescence as well as transcriptional regulation (Fig. 4C). The observed pathways, such as those linked to DNA damage response and programmed cell death, are also in line with the acknowledged biology of cancers. These pathways along with cellular senescence pathways were targeted by coregulative and upregulated hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p (Supplementary Table S13). In summary, hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p showed potential coregulation by displaying reciprocal correlation and by targeting genes involved in several biological pathways relevant to cancers.

#### c-miR risk sum score correlates with BMI

c-miRs modulate multisystemic adaptations in the human body in response to lifestyle behavior. Therefore, we investigated whether the five c-miR risk sum score was associated with lifestyle factors that are reported to reduce LS cancer risk, or age which is a significant cancer risk factor in LS. Of the chosen lifestyle factors, only BMI showed significant correlation with the c-miR risk sum score (Table 2). Using the multiple imputed datasets did not show significant differences to a complete-case analysis (Supplementary Table S14). These findings indicate that the expression levels of hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a, hsa-miR-3613-5p, and hsa-miR-3615 might be affected by BMI thus suggesting potential link between lifestyle, c-miRs and LS cancer risk.

**Table 2.** Pearson correlations of c-miR risk sum score and physical activity, BMI, dietary fiber consumption, NSAID usage, and age.

	<i>r</i>	95% CI	<i>P</i> -value
Physical activity	0.03	[-0.19, 0.26]	0.76
BMI	0.23	[0.04-0.43]	<b>0.01</b>
Dietary fiber intake	0.04	[-0.18, 0.26]	0.71
NSAID usage	-0.03	[-0.25, 0.18]	0.75
Age	-0.14	[-0.33, -0.05]	0.14

Note: Lifestyle data were collected 2017 or 2020 with a questionnaire. Blood sample was taken at regular colonoscopy visit. The average time-period between lifestyle data collection and blood sample was 2.0 (0.3–3.9) years. *P*-value significant at 0.05 level.

Abbreviations: *r* = Pearson correlation coefficient; 95% CI = 95% confidence interval; BMI = body mass index; NSAID = non-steroidal anti-inflammatory drug usage.

## Discussion

Our pilot study was the first to assess whether a c-miR expression signature could be used in LS cancer risk prediction during a 4-year prospective surveillance period. We also investigated whether this signature associates with lifestyle factors and age. Using Lasso regression and bioinformatics approaches, we showed that a risk sum score composed of hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615 associates with an increased risk of LS cancer incidence. We also observed that this c-miR risk sum score correlates positively with BMI.

Identifying reliable biomarkers has the potential to aid in risk stratification of high-risk patients (33). Integrating these biomarkers with clinicopathologic factors could enhance the accuracy of patient selection criteria for risk-based screening programs. In the current study, Lasso-Cox model successfully separated LS carriers who developed cancer from those who did not by using a c-miR signature. Our finding suggests that c-miR expression can classify high-risk cases in LS population, also when stratified for *MLH1*-variant or colorectal cancer, but further validation is required. This observation is valuable because the variation of cancer risk is high among LS carriers (34), and the implementation of intense screening programs is not uniformly effective (35). Therefore, a more nuanced approach is needed to identify those patients who are most likely to benefit from the screenings.

Cross-validations of the risk prediction models showed that c-miR risk sum scores have risk prediction potential also in randomly generated subsets with varying surveillance time and number of events. This finding is supported by previous research. For example, hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p, that were upregulated in those who developed cancer within the LS cohort, are well-recognized sporadic colorectal cancer miRs with multiple roles and reported biomarker potential (13, 14, 36–38). hsa-miR-3613-5p has been established as a colorectal cancer miR (39) whereas hsa-miR-3615 has been previously reported to display down-regulation in microsatellite unstable colorectal tumors, which are hallmark tumors of LS, when compared with their microsatellite stable counterparts (40).

Furthermore, these five c-miRs displayed correlation as well as higher expression in those LS carriers who developed cancer compared those who did not, thus suggesting potential coregulation and biological connection. In support, we found that four out of the five c-miRs (hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, and hsa-miR-3615) have been experimentally shown to target several well-established oncogenes and tumor suppressor genes, including *BCL2*, *EGFR*, *CDKN1A*, *CDKN2A*, *CREB1*, *STAT3*, and *SMAD2* (41). Also, these genes formed interconnected nodes, which indicates similar role and biological connection among them and provide more support for the suggested coregulation of these c-miRs. All of these genes are part of cancer-relevant biological pathways, such as those in apoptosis, DNA damage, and cellular senescence (42).

Wikberg and colleagues observed that major changes of miR patterns occur mainly 3 years prior to sporadic colorectal cancer diagnosis by showing a temporal pattern of increase in miR-21-5p expression by using prediagnostic and postdiagnostic plasma samples (43). Raut and colleagues reported that a risk sum score of seven c-miRs was highly predictive for sporadic colorectal cancer risk in a prospective cohort with a follow-up time up to 14 years and median follow-up of 6.8 years (16). However, the c-miR signature we identified did not include any of the miRs observed by Raut and colleagues. In contrast to sporadic colorectal cancer that develops commonly in 10–15 years, the development of LS colorectal cancer is significantly accelerated, often taking only 1 to 3 years to progress to carcinoma with or without pre-existing adenoma (44), which may explain the discrepancies between our study and the study by Raut and colleagues. As LS carriers in our study were diagnosed with cancer in 1.3 years on average from the serum sampling, it is possible that the observed c-miR signature originates from tumors. However, it is also possible that the observed c-miR levels may reflect risk rather than tumor presence because our sample was not limited to colorectal cancers. Nonetheless, these studies as well as our bioinformatics analyses show promising results for using c-miRs in LS cancer risk prediction.

Interestingly, we found a positive correlation between the c-miR risk sum and BMI suggesting a potential link between excess body weight, c-miRs, and cancer risk. In support to our findings, hsa-miR-10b-5p, hsa-miR-125b-5p, and hsa-miR-200a-3p have been previously linked with increased levels of plasma total cholesterol, dysregulated lipid metabolism, and overweight/obesity in general (45–47). Mens and colleagues, reported that upregulation of hsa-miR-10b-5p and hsa-miR-125b-5p associate with increased total cholesterol (45). Conversely, Ortega and colleagues reported a positive correlation between decreased levels of hsa-miR-125b and BMI after surgery-induced weight loss in obese patients (46). Ruiz-Roso and colleagues showed upregulated miR-200a to regulate lipid metabolism-related genes in a mouse model (47), although we did not find hsa-miR-200a-3p to target those genes. Moreover, Dogan and colleagues reported 1,558 miR-target gene interactions in obesity, including miR-125b, that were also detected in multiple cancer types. They also showed that metabolism and growth signaling pathways are shared by obesity and obesity-related cancer (48). Of the pathways reported by Dogan and colleagues, p53-signaling pathway was also identified in our study as a key pathway targeted by the c-miRs of the risk sum score. In addition, cellular senescence and FOXO pathways emerged in our analysis. These pathways have been reported to associate with cancer metabolism and obesity via alteration of energy metabolism and adipose tissue (49, 50). However, it is important to note that c-miRs have multifaceted roles in metabolism, and their profiles change with disease progression. Without mechanistic studies, it is challenging to exclude the potential confounding effects of disease and genetics in our findings. As metabolomic abnormality is an acknowl-

edged cancer hallmark (42), these c-miRs could be promising targets to study when assessing the interactions of metabolic dysregulation and cancer.

A major strength of our study is that we were able to conduct an analysis using prediagnostic samples from a high-risk cohort under frequent surveillance. We also used robust methodology to interrogate c-miR signatures and their associations with LS cancer risk. All of the analyses were conducted carefully with state-of-the-art methods and tools. By utilizing Lasso regression, we were able to choose the most promising c-miRs and integrate them along with the surveillance time into well-established tool used for risk prediction, thus allowing comprehensive biomarker signature investigation. Missing values were handled with multiple imputation that is reported to have negligible bias when missingness occurs randomly (51). Finally, we followed TRIPOD checklist to enhance transparency in our risk prediction model development and validation as well as to improve reproducibility of these results.

As in many pilot experiments, the potential pitfall of our study is the small sample and effect size. Despite our best efforts to look for an external validation dataset, we unfortunately did not find a suitable candidate dataset nor had the opportunity to increase our sample size. For these reasons, we could not validate our predictor selection model. Because the majority of LS carriers are not most likely identified (44), and due to lack of resources, it is difficult to gather enough samples as well as it is costly to obtain enough small RNA-seq data for a more comprehensive investigation. An international collaboration study would be beneficial for such purposes. We also acknowledge that because the study population was comprised mainly of *MLH1* carriers, our results might have limited generalizability to other pathogenic MMR variant carriers. Finally, the average time period of 2.0 years between the lifestyle questionnaire data collection and blood sampling is also a potential limitation of this study.

To conclude, we report that a risk sum score composed of hsa-miR-10b-5p, hsa-miR-125b-5p, hsa-miR-200a-3p, hsa-miR-3613-5p, and hsa-miR-3615 has potential in LS cancer risk prediction, and thus may serve as a stratification biomarker signature for finding LS carriers at increased cancer risk in the future. However, more experiments with larger sample size are needed to confirm our findings. The molecular mechanisms underlying the associations of body weight, LS cancer risk and c-miRs remain to be elucidated in future studies.

## Authors' Disclosures

T. Jokela reports grants from European Commission Marie Skłodowska-Curie Individual Fellowships during the conduct of the study. T.T. Seppälä reports grants from Finnish Medical Foundation, Emil Aaltonen Foundation, Jane and Aatos Erkko Foundation, Relander Foundation, and Cancer Foundation Finland during the conduct of the study; personal fees from Amgen Finland, personal fees and other support from LS CancerDiag, grants from Academy of Finland, and other support from Healthfund Finland outside the submitted work. E.K. Laakkonen reports grants from Päivikki and Sakari Sohlberg Foundation during the conduct of the study. No disclosures were reported by the other authors.

## Authors' Contributions

**T. Sievänen:** Conceptualization, software, formal analysis, investigation, visualization, methodology, writing—original draft, writing—review and editing. **T. Jokela:** Software, investigation, writing—review and editing. **M. Hyvärinen:** Software, formal analysis, methodology, writing—review and editing. **T.-M. Korhonen:** Software, writing—review and editing. **K. Pylvänäinen:** Data curation, writing—review and editing. **J.-P. Mecklin:** Resources, data curation, funding acquisition, project administration, writing—review and editing. **J. Karvanen:** Conceptualization, software, methodology, writing—review and editing. **E. Sillanpää:** Conceptualization, supervision, writing—review and editing. **T.T. Seppälä:** Conceptualization, resources, data curation, supervision, funding acquisition, project administration, writing—review and editing. **E.K. Laakkonen:** Conceptualization, resources, data curation, supervision, funding acquisition, project administration, writing—review and editing.

## Acknowledgments

We would like to extend our gratitude to prof. Sarianna Sipilä (Gerontology Research Center and Faculty of Sport and Health Sciences,

University of Jyväskylä, Finland) for the lifestyle data acquisition. E.K. Laakkonen was supported by grants from Päivikki and Sakari Sohlberg Foundation. T. Jokela was supported by European Commission Marie Skłodowska-Curie Individual Fellowships (grant number: H2020-MSCA-IF-2020 #101026706). T.T. Seppälä was supported by funding from the Academy of Finland and iCAN Precision Medicine Flagship of Academy of Finland, and research grants by Jane and Aatos Erkkö Foundation, Finnish Medical Foundation, Sigrid Juselius Foundation, Emil Aaltonen Foundation, Cancer Foundation Finland, Relander Foundation, and state research funding from the Finnish Government, which is allocated as competed grants through employing institutions to researchers within their university hospital co-operation area (Tampere University Hospital/Helsinki University Hospital).

## Note

Supplementary data for this article are available at Cancer Prevention Research Online (<http://cancerprevres.aacrjournals.org/>).

Received September 6, 2023; revised January 3, 2024; accepted March 27, 2024; published first March 28, 2024.

## References

- Win AK, Jenkins MA, Dowty JG, Antoniou AC, Lee A, Giles GG, et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 2017;26:404–12.
- Peltomäki P, Nyström M, Mecklin J-P, Seppälä TT. Lynch syndrome genetics and clinical implications. *Gastroenterology* 2023;5:783–99.
- Sievänen T, Törmäkangas T, Laakkonen EK, Mecklin JP, Pylvänäinen K, Seppälä TT, et al. Body weight, physical activity, and risk of cancer in lynch syndrome. *Cancers* 2021;13:1849.
- Jamizadeh N, Walton Bernstedt S, Haxhijaj A, Andreasson A, Björk J, Forsberg A, et al. Endoscopic surveillance of Lynch syndrome at a highly specialized center in Sweden: an observational study of interval colorectal cancer and individual risk factors. *Front Oncol* 2023;13:1127707.
- Dashti SG, Win AK, Hardikar SS, Glombicki SE, Mallenahalli S, Thirumurthi S, et al. Physical activity and the risk of colorectal cancer in Lynch syndrome. *Int J Cancer* 2018;143:2250–60.
- Dominguez-Valentin M, Sampson JR, Seppälä TT, ten Broeke SW, Plazzer JP, Nakken S, et al. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the prospective lynch syndrome database. *Genet Med* 2020;22:15–25.
- Burn J, Sheth H, Elliott F, Reed L, Macrae F, Mecklin J-P, et al. Cancer prevention with aspirin in hereditary colorectal cancer (Lynch syndrome), 10-year follow-up and registry-based 20-year data in the CAPP2 study: a double-blind, randomised, placebo-controlled trial. *Lancet* 2020;395:1855–63.
- Mathers JC, Elliott F, Macrae F, Mecklin J-P, Möslin G, McRonald FE, et al. Cancer prevention with resistant starch in lynch syndrome patients in the CAPP2-randomized placebo controlled trial: planned 10-year follow-up. *Cancer Prev Res* 2022;15:623–34.
- Mori MA, Ludwig RG, Garcia-Martin R, Brandão BB, Kahn CR. Extracellular miRNAs: from biomarkers to mediators of physiology and disease. *Cell Metab* 2019;30:656–73.
- Goodall GJ, Wickramasinghe VO. RNA in cancer. *Nat Rev Cancer* 2021;21:22–36.
- Jung G, Hernández-Illán E, Moreira L, Balaguer F, Goel A. Epigenetics of colorectal cancer: biomarker and therapeutic potential. *Nat Rev Gastroenterol Hepatol* 2020;17:111–30.
- Shah R, Patel T, Freedman JE. Circulating extracellular vesicles in human disease. *N Engl J Med* 2018;379:958–66.
- Yang G, Zhang Y, Yang J. A five-microRNA signature as prognostic biomarker in colorectal cancer by bioinformatics analysis. *Front Oncol* 2019;9:1207.
- Zhang J, Song W, Chen Z, Wei J, Liao Y, Lei J, et al. Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *Lancet Oncol* 2013;14:1295–306.
- Adam RS, Poel D, Ferreira Moreno L, Spronck JMA, de Back TR, Torang A, et al. Development of a miRNA-based classifier for detection of colorectal cancer molecular subtypes. *Mol Oncol* 2022;16:2693–709.
- Raut JR, Schöttker B, Hollecsek B, Guo F, Bhardwaj M, Miah K, et al. A microRNA panel compared to environmental and polygenic scores for colorectal cancer risk prediction. *Nat Commun* 2021;12:4811.
- Bye A, Røsjø H, Aspenes ST, Condorelli G, Omland T, Wisløff U. Circulating MicroRNAs and aerobic fitness - the HUNT-study. *PLoS One* 2013;8:e57496.
- Sievänen T, Korhonen T-M, Jokela T, Ahtiainen M, Lahtinen L, Kuopio T, et al. Systemic circulating microRNA landscape in Lynch syndrome. *Int J Cancer* 2023;152:932–44.
- Kaartinen NE, Tapanainen H, Valsta LM, Similä ME, Reinivuo H, Korhonen T, et al. Relative validity of a FFQ in measuring carbohydrate fractions, dietary glycaemic index and load: exploring the effects of subject characteristics. *Br J Nutr* 2012;107:1367–75.
- van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1–67.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;26:139–40.
- Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med* 1997;16:385–95.
- Harrell FEJ, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–6.
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015;13:1–10.

26. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
27. Therneau TM, Grambsch PM. The Cox model BT - modeling survival data: extending the Cox model. In: Therneau TM, Grambsch PM, editors. New York (NY): Springer New York; 2000. p. 39–77.
28. Huang HY, Lin YCD, Li J, Huang KY, Shrestha S, Hong HC, et al. MiRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res* 2020;48:D148–54.
29. Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: an online resource for prediction of microRNA binding sites. *PLoS One* 2018;13:e0206239.
30. Snel B, Lehmann G, Bork P, Huynen MA. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 2000;28:3442–4.
31. Wu G, Haw R. Functional interaction network construction and analysis for disease discovery. *Methods Mol Biol* 2017;1558:235–53.
32. Bracken CP, Scott HS, Goodall GJ. A network-biology perspective of microRNA function and dysfunction in cancer. *Nat Rev Genet* 2016;17:719–32.
33. Sur D, Advani S, Braithwaite D. MicroRNA panels as diagnostic biomarkers for colorectal cancer: a systematic review and meta-analysis. *Front Med* 2022;9:915226.
34. International Mismatch Repair Consortium. Variation in the risk of colorectal cancer in families with Lynch syndrome: a retrospective cohort study. *Lancet Oncol* 2021;22:1014–22.
35. Möller P, Seppälä T, Bernstein I, Holinski-Feder E, Sala P, Evans DG, et al. Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut* 2017;66:464–72.
36. Sheedy P, Medarova Z. The fundamental role of miR-10b in metastatic cancer. *Am J Cancer Res* 2018;8:1674–88.
37. Pichler M, Ress AL, Winter E, Stiegelbauer V, Karbiener M, Schwarzenbacher D, et al. MiR-200a regulates epithelial to mesenchymal transition-related gene expression and determines prognosis in colorectal cancer patients. *Br J Cancer* 2014;110:1614–21.
38. Yamada A, Horimatsu T, Okugawa Y, Nishida N, Honjo H, Ida H, et al. Serum MIR-21, MIR-29a, and MIR-125b are promising biomarkers for the early detection of colorectal neoplasia. *Clin Cancer Res* 2015;21:4234–42.
39. Xiang F, Xu X. CirRNA F-circEA-2a suppresses the role of miR-3613-3p in colorectal cancer by direct sponging and predicts poor survival. *Cancer Manag Res* 2022;14:1825–33.
40. Slattery ML, Lee FY, Pellatt AJ, Mullany LE, Stevens JR, Samowitz WS, et al. Infrequently expressed miRNAs in colorectal cancer tissue and tumor molecular phenotype. *Mod Pathol* 2017;30:1152–69.
41. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2019;47:D941–7.
42. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov* 2022;12:31–46.
43. Wikberg ML, Myte R, Palmqvist R, van Guelpen B, Ljuslinder I. Plasma miRNA can detect colorectal cancer, but how early? *Cancer Med* 2018;7:1697–705.
44. Seppälä TT, Burkhart RA, Katona BW. Hereditary colorectal, gastric, and pancreatic cancer: comprehensive review. *BJS Open* 2023;7:zrad023.
45. Mens MMJ, Maas SCE, Klap J, Weverling GJ, Klatser P, Brakenhoff JJP, et al. Multi-omics analysis reveals MicroRNAs associated with cardiometabolic traits. *Front Genet* 2020;11:110.
46. Ortega FJ, Mercader JM, Catalán V, Moreno-Navarrete JM, Pueyo N, Sabater M, et al. Targeting the circulating microRNA signature of obesity. *Clin Chem* 2013;59:781–92.
47. Ruiz-Roso MB, Gil-Zamorano J, López de las Hazas MC, Tomé-Carneiro J, Crespo MC, Latasa MJ, et al. Intestinal lipid metabolism genes regulated by miRNAs. *Front Genet* 2020;11:707.
48. Dogan H, Shu J, Hakguder Z, Xu Z, Cui J. Elucidation of molecular links between obesity and cancer through microRNA regulation. *BMC Med Genet* 2020;13:161.
49. Smith U, Li Q, Rydén M, Spalding KL. Cellular senescence and its role in white adipose tissue. *Int J Obes* 2021;45:934–43.
50. Yadav RK, Chauhan AS, Zhuang L, Gan B. FoxO transcription factors in cancer metabolism. *Semin Cancer Biol* 2018;50:65–76.
51. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920–31.