Machine Learning Reveals the Unique Biomarkers of Clonal Hematopoiesis in Patient With Early-Stage Colorectal Neoplasia: A Case Control Study

Yin-Chen Hsu
National Taiwan University College of Medicine
Sin-Ming Huang
National Taiwan University College of Medicine
Li-Chun Chang
National Taiwan University Hospital
Yan-Ming Chen
National Taiwan University College of Medicine
Wei-Tzu Chiu
National Taiwan University College of Medicine
Jing-Wei Lin
National Taiwan University College of Medicine
Chien-Chia Lin
National Taiwan University College of Medicine
Ching-Wen Chen
National Taiwan University College of Medicine
Han-Mo Chiu
National Taiwan University College of Medicine
SUNG-LIANG YU (slyu@ntu.edu.tw)
National Taiwan University https://orcid.org/0000-0003-4535-9036

Research

Keywords: Colorectal neoplasia, Clonal hematopoiesis, Early diagnosis, Machine learning

DOI: https://doi.org/10.21203/rs.3.rs-431856/v1

Abstract

BACKGROUND: Blood test has a better uptake for colorectal cancer screening than stool test and colonoscopy but suboptimal detection of early-stage colorectal neoplasia (CRN), including advanced adenoma and stage I cancer, limits its application. The present study aimed to evaluate whether clonal hematopoiesis (CH) from peripheral blood can be used as a biomarker for early-stage CRN screening and improve the detection of blood tests by machine-learning approach.

METHODS: The CH profile was evaluated in 63 early-stage CRNs and 32 controls by error-corrected sequencing and classified by machine-learning method. Diagnostic performance was measured by receiver operator characteristic analysis. Additional 20 early-stage CRNs and 10 controls were used to validate the machine-learning model. We simultaneously used mutational signature analysis to study predictors based on CH.

RESULTS: We identified 1,446 variants and clarified the uniqueness of variants from the peripheral bloods of early-stage CRNs. The machine learning model identified early-stage CRNs from controls and its AUC, sensitivity and specificity were 0.988, 94.2% and 99.3%, respectively. The CH-based CRN detection model was further verified. The accuracy, sensitivity, and specificity were 0.933 (p=0.00065), 95.0%, and 90.0%, respectively. Furthermore, the mutational signature analysis of those unique variants in CRNs revealed the influence of genetic architecture on DNA damages.

CONCLUSIONS: Our results reveal the potential of CH to a mark produced by the carcinogenesis in early-stage CRN. We developed a CH-based blood test with machine learning approach, which not only increase screening uptake but also improve the detection rate of early-stage CRN.

1. Background

Colorectal cancer (CRC) draws attentions because it is the third most common cancer and the second leading cause of cancer-related deaths worldwide[1]. In Taiwan, the incidence of CRC was 42.9 cases per 100,000 residents in 2017 and was the most common cancer than lung cancer and breast cancer. CRC has high mortality in advanced stages but it is preventable by removal of precancerous lesions (advanced adenoma, AA) or detection of CRC in early stage[2]. Among the various screening modalities, colonoscopy remains the gold standard and reduces the mortality rates by 67%[3,4]. However, it remains debated whether colonoscopy qualified as first-line screening modality because of its inconvenience, invasiveness, and cost, which limit the adherence of general population to the screening program. Currently, fecal immunochemical test (FIT) and fecal DNA test are approaches as non-invasive tests for screening CRC. Although FIT can help reduce the mortality of CRC, its low specificity and sensitivity have hampered its clinical application in the identification of early-stage colorectal neoplasia (CRN)[5]. Even fecal DNA tests can detect early-stage neoplasia better than FIT, the social acceptance, high cost, and limited screening data also limit its utility as a screening purpose[6,7]. Compared with stool-based

screening, blood-based screening has an overwhelming advantage, which highlights the necessity and potential advantages of developing a highly reliable blood-based test to identify early-stage CRN.

In the past decade, emerging data have demonstrated that clonal hematopoiesis (CH) in the peripheral blood may imprecate the microenvironment in disease[8-14]. CH is a term referred to the clonal expansions of mutated hematopoietic cells and is common in aging human[15-18]. The *DNMT3A*, *TET2*, *ASXL1* and *JAK2* are canonical CH genes and common mutated in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS)[19,20]. Recent genetic studies have shown that CH is common in individuals without hematological malignancies and is associated with the cardiovascular disease and stroke by promoting inflammation of blood vessel walls[17,21,22]. CH was also reported to be relevant in the immune response to tumors by preventing T-cell exhaustion through the *DNMT3A* deletion in CD8+ T cells, which led to enhanced response to PD-1 checkpoint inhibitor in mice[23].

In addition, recent studies have shown the application of "error-corrected sequencing" (ECS) in CH research. Its "error suppression and mutation call" is designed as a "unique molecular index" (UMI), which improves the detection sensitivity of achieving variant allele frequency (VAF) ≥ 0.0001[24]. Genetic analysis of human blood samples using ECS has proven that rare hematopoietic clones have found in almost everyone over the age of 50[25,26]. It is worth noting that those rare hematopoietic clones are below the detection limit of the standard WES.

In this study, we hypothesized that CH variants in peripheral blood may reflect the beguiling environment and serve as a detection landmark for early cancer screening. Therefore, we enrolled the coloscopy-informed CRN patients and risk-matched healthy controls for CH identification by ECS. A unique spectrum of CH variants was discovered and CH-based predictors were identified by machine-learning approach.

2. Methods

2.1. Patients

This study protocol was approved by the Institutional Review Board of the National Taiwan University Hospital (No.201712033RIN). Respective colorectal neoplasm patients and risk-matched controls were retrieved from Health Management Center at National Taiwan University Hospital. Eligible patients were at least 40 years of age and classified based on the histological examination and coloscopy assay. Early-stage CRN was defined as AA and stage I cancer.

2.2. Blood collection and DNA extraction

Whole blood was collected in K_2EDTA tube and processed from 2-6 hours after blood drawn. Peripheral blood DNA was extracted from PBMC using QIAamp Blood DNA Mini kit (Qiagen, Hilden Germany) according to the manufacturer's instructions. Additional fragmentation was performed from 100 ng peripheral blood DNA using KAPA Freg kit (KAPA Biosystems, Wilmington, MA) under the condition of 37°C incubation for 30 minutes.

2.3. Library preparation and next-generation sequencing

Sequencing libraries were prepared according to the manufacturer's instructions of AVENIO Expanded kit (Roche sequencing solution, Pleasanton, CA) with 30 ng of fragmented peripheral blood DNA. The library profile was analyzed with the Agilent 2200 Bioanalyzer (Agilent Technologies, Palo Alto, CA) and quantified using QuBit dsDNA HS Assay kit (ThermoFisher, Waltham, MA). For a single sequencing run, the 8 or 16-multiplexed library was created by pooling the libraries and sequenced on a lane of Illumina HiSeq 4000 flow cell or NovaSeq 6000 S1 flow cell with 2×150-bp paired-end reads. Raw sequencing was analyzed using *AVENIO* ctDNA analysis software (version 1.1.0) with the setting of "Default-Expanded-Panel-Workflow ".

2.4. Statistical analysis

Statistical analysis was performed in R (version 3.6.1). Two-sample T test was used for comparisons of the VAF distributions. Wilcoxon rank-sum test was performed to test the probability of difference in VAFs observed between the CRN patients and controls. Cosine correlation similarity was in using the "MutationalPatterns" package to measure the closeness of mutational signatures. Visualizations of variants in different groups were carried out in using the R package "ggplot2" and "reshapre2".

2.5. Model-training method

A semi-supervised learning framework with pre-processing filtering was used for the establishment of statistical model. The classification framework of model-training used the R package "caret" and "mlbench". Confidence intervals for AUC, sensitivity and specificity of training-model were generated by 70:30 resamplings and 200 bootstrap with the setting as 10-fold CV as a leave-one-out cross-validation framework.

2.6. Mutation signature analysis

The contribution of known mutation process to the mutations was assessed by "Mutational Signatures in Cancer" (MuSiCa) using the COSMIC signature set v2[27]. Mutations were pooled across individuals to evaluate mutation signatures present in CRN patients, risk-matched controls or model-training in CRN patients for given comparisons.

3. Results

3.1. Reshaping the mutational spectrum of CH in peripheral blood by error-corrected sequencing

As a step towards developing a non-invasive method for early-stage CRN screening, we characterized mutations including CH-related variants in peripheral blood mononuclear cells (PBMCs) using the ECS approach (Supplementary Fig. S1.). Initially, 63 early-staged CRN patients and 32 risk-matched healthy individuals were enrolled for the feature discovery (Table 1). Before further analysis of the variants, we implemented a quality control to ensure that the unique depth of each individual is greater than 3,000X

(Supplementary Fig. S2). By using ECS method in this study, 1,446 variants were identified in the testing cohort (Fig. 1). There were 1,171 variants identified in CRN patient group and 753 of 1,171 (64.3%) variants were unique and not shared with the risk-matched controls. On the other hand, 693 variants were identified in risk-matched control group and 275 of 693 (39.7%) variants were unique and not shared with CRN patient group. The average number of variants in CRN patients and risk-matched controls were 128.2 (92-167) and 127.2 (101-166), respectively (Supplementary Fig. S3). Somatic mutations with VAF greater than 0.02 are the traditional clonal hematopoiesis indeterminate potential (CHIP) definition, only 7 (11.11%) were found in 63 CRN patients, and 3 (9.38%) were found in 32 risk-matched controls (Supplementary Table1).

Not surprisingly, the rare hematopoietic clones (defined as VAF lower than 0.02, also refer to "low-allele-fraction variants" in this study) were detected in all samples (Fig. 2). The average proportion of rare hematopoietic clones was 39.05% (12.63%-53.33%) in CRN patients and 37.07% (8.6%-54.3%) in risk-matched controls. Notably, the mean proportion of variants with VAF lower than 0.001 were 15.88% (3.16%-27.5%) and 16.1% (3.31%-26.53%) in CRN patients and risk-matched controls, respectively. There was no statistical significance in the difference of the variant quantities between CRN patients and risk-matched controls.

3.2. Machine learning constructs the CH-based CRN detection model for early-stage CRN identification

In this case-control study, those variants were scattered and it was difficult to deal with those statistical extremes through traditional analysis methods. Therefore, we implemented machine learning (ML) methods to discriminate early-stage CRN patients from risk-matched controls. Initially, the pre-processing filtering was performed using the 1,446 variants from the discovery cohort with following filters: (1) removed the shared variants present in CRN patient and risk-matched controls with a Wilcoxon rank sum test *p*-value >0.1, (2) removed the unique variants (only in CRN patients or controls) with less than two cases in the discovery cohort, (3) rescued the filtered unique variants which reported with the pathogenicity significantly. With the filter criteria, 108 resulting variants were used as the variables for model training (Fig. 3A). Next, we performed the model training with the ML package based on a leave-one-out cross-validation framework method. The result of model training was that the AUC was 0.988, the sensitivity was 94.2%, and the specificity was 99.3% (Fig. 3B). It is worth noting that in this cohort, the FIT test sensitivity of the AA group and the stage I cancer group were 48.9% and 72.2%, respectively.

Another validation cohort, including 20 early-staged CRN patients and 10 risk-matched controls, was used to validate the model training. The accuracy of the validation study was 0.933 (95% CI, 0.779-0.982; p =0.00065), the sensitivity was 95.0%, and the specificity was 90.0% (Table 2).

3.3. Mutational signature analysis reveals the influence of genetic architecture on DNA damages

To understand the mechanism and influence of identified variants, we performed the mutational signature analysis by COSMIC mutational signature v2, in which may reflect the activity of 30 specific mutational processes. Initially, we compared the contribution of mutational signatures generated from

753 unique variants unique in the early-stage CRN group to 693 variants observed in the risk-matched control group (Fig. 4A). As the result as shown in Fig. 4C, signature 1, 3, 12, 13, 20, 22, 28 were contributed in both CRN unique group and risk-matched control group. Contrastively, signature 15, 18, 21 were uniquely contributed in the CRN-unique group as the related-contribution score as 0.024, 0.007, 0.013, respectively. Next, we analyzed the contribution of mutational signatures for 86 CRN-unique variants of 108 variants used in ML approach (Fig. 4B). Signature 3, 10, 12, 15, 20, 21, 22, 30 were contributed from those 86 variants and signature 10, 15, 21 and 30 were uniquely contributed from those 86 variants compared to the risk-matched control group. The cosine similarity value between 86 CRN-unique ML model variants and 753 CRN-unique variants was 0.855. Relatively, the cosine similarity value between 86 CRN-unique ML model variants and 693 variants in control group was 0.558.

About the characteristics of these four signatures, signature 10 has been found in six cancer types including CRC. The proposed etiology of signature 10 is to alter the activity of the error-prone polymerase POLE. Samples exhibiting signature 10 have been termed "ultra-hypermutators". Signature 15, has been found in gastric cancer and small cell lung cancer, is associated with defective DNA mismatch repair (MMR) and with high numbers of small indel (shorter than 3 bp) at mono/polynucleotide repeats. The etiology of signature 21 remains unknown but it was found only in samples that also have signature 15 and signature 20. As a result, signature 21 is probably related to microsatellite instable (MSI) tumors. The etiology of signature 30 remains unknown. Furthermore, we found that signature 1 did not contribute to the 86 variant groups, which may indicate that age-related variants are common in both groups[28] , which is why the ML method did not select the variant related to signature 1.

4. Discussion

In the present study, we reshaped the CH landscape in early-stage CRN patients. It was not a surprise that high rate of low-allele-fraction variants were measured in almost all people when using the ECS techniques for CH profiling. However, as far as we know, there is limited knowledge about the dynamics of genetic diversity in the large clonal-expansion in the regenerating hematopoietic stem cell populations. Indeed, most studies identified those low-allele-fraction variants in white blood cells and served those variants as "background" or "contaminate" in cell-free DNA (cfDNA) pools. What sets us apart in this study is that we reshaped the genetic variation pattern of each individual and determined the uniqueness of the variation distribution from this case-control study. It is worth noting that these unique variants are scattered in the cohort, which increases the difficulty of using traditional analysis methods to deal with those statistical extremes. Therefore, we proved the potential of machine learning methods in improving statistical analysis and established an ML-based CRN detection model with a performance of 0.988 in AUC. When applying this method to panel design, it is essential that the risk of model overfitting should be reduced. Otherwise, overfitting of the model may lead to overly optimistic results. Therefore, an independent verification queue is essential for model verification. On the other hand, sufficient unique sequencing depth (coverage of more than 3,000X as shown in this study) is required to ensure the accuracy of detection.

In the pathogenicity of view, most of those low-allele-fraction variants are not canonical pathogenic variants but still selected by the machine-learning method for the model training. To test the causality, the mutational signature analysis was performed in this study. As a result, three unique mutational signatures with ultra-hypermutator, defective DNA mismatch repair and microsatellite-instability were identified by the ML method. Interestingly, the characteristics of these three mutational signatures are related to the effect of DNA damages. In metastatic colon cancer, the test of MSI/ dMMR and the process of POLE mutation signature are used either in guiding the therapeutic decision, predicting the survival outcome[29,30], or in study the colorectal tumorigenesis[31-33]. For example, CRC tumors with MSI have defective in DNA repair enzymes and infiltrated by a large number of lymphocytes[34]. In addition, metastatic CRCs with MSI were found to have a better response to immune checkpoint inhibitors[35]. We also found that the score of related-contribution in signature 15 (Defective DNA MMR) and 21 (Unknown, stomach cancer/MSI) were increased by ML approach from 0.024 and 0.013 to 0.148 and 0.102, respectively. This observation indicated that the ML approach could effectively enrich CRN-specific variants improving the accuracy of early cancer screening and characterizing the possible causality of CRN. There is not enough information from our current data to figure out the consequence of the mutation process, but we would like to demonstrate the potential of mutational signature analysis in studying the tumorigenesis.

Additionally, the COSMIC signature 22 as "exposures to aristolochic acid " is observed in the control group, CRN-unique group and the model-training group but not found in the COSMIC CRC database. It makes sense because the exposures of aristolochic acid is a critical issue in Taiwan and almost one third Taiwanese were caught in the exposures of aristolochic acid[36]. Further studies with the comparison in different populations may provide the evidence to study the relationship between aristolochic acid exposure and higher incidence of CRC in Taiwan. We do not exclude the possibility that the prediction accuracy of ML-based CRN detection model might be affected by ethnicity and environmental agents. Whether this ML-based CRN detection is suitable for Caucasian population remains to further investigation.

5. Conclusion

We establish a peripheral blood-based non-invasive early-stage CRN detection by improved NGS sequencing technique and machine learning method for decoding the information from CH in CRN patients. Although the use of this CH-based blood test is still limited in clinical practice and further prospective studies with larger sample size would be needed to clarify the clinical effectiveness, the present study has demonstrated the potential of CH for early cancer diagnosis and helps to decode the aberration of CH for imprecating the microenvironment in disease.

Abbreviations

AA: advanced adenoma

AML: acute myeloid leukemia

AUC: area under the receiver operator characteristic curve

CH: clonal hematopoiesis

CHIP: clonal hematopoiesis of indeterminate potential

CRN: colorectal neoplasia

CRC: Colorectal cancer

ECS: error-corrected sequencing

FIT: fecal immunochemical test

MDS: myelodysplastic syndrome

ML: machine learning

MMR: defective DNA mismatch repair

MSI: microsatellite unstable

PBMCs: peripheral blood mononuclear cells

UMI: Unique Molecular Index

VAF: variant allele frequency

WES: whole exome sequencing

Declarations

Ethics approval and consent to participate:

This study protocol was approved by the Institutional Review Board of the National Taiwan University Hospital (No.201712033RIN).

Consent for publication:

All authors agree to the publication of our work titled "Machine learning reveals the unique biomarkers of clonal hematopoiesis in patient with early-stage colorectal neoplasia: a case control study" to *Journal of Experimental & Clinical Cancer Research*.

Availability of data and materials:

All error-corrected sequencing data from this study were deposited in the **Sequence Read Archive** (SRA) with the accession number of PRJNA678402. All other remaining data are available within the article and supplemental Files, or available from the authors upon request.

Competing interests:

The authors have declared that no competing interest exists.

Funding:

This work was supported by grants from Pharmacogenomics Laboratory of NCFB and Center of Precision Medicine' from The Featured Areas Research Center Program and by the Ministry of Science and Technology (MOST108-2319-B-002-001, MOST-109-2634-F-002-043).

Authors' contributions:

Y.-C.H., S.-M.H., L.-C.C., S.-L.Y. and H.-M.C. designed the study, analyzed the data and wrote the manuscript. Y.- C.H., S.-M.H., W.-T.C., J.-W.L., C.-C.L. and C.-W.C. acquired, processed patient specimens and performed experiments. Y.-C.H. and Y.-M.C. performed the analysis. L.-C.C. and H.-M.C. provided clinical samples and data. The manuscript was written by Y.-C.H, L.-C.C., S.-M.H., S.-L. Y. and H.-M.C. and was reviewed and edited by all authors.

Acknowledgements:

We would like to thank the technical support from the Pharmacogenomics Laboratory of National Core Facility for Biopharmaceuticals (NCFB), the NGS and Microarray Core Facility of NTU Centers of Genomic and Precision Medicine.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021.
2. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *N Engl J Med.* 2012;366(8):687-696.
3. Singh H, Nugent Z, Demers AA, Kliewer EV, Mahmud SM, Bernstein CN. The reduction in colorectal cancer mortality after colonoscopy varies by site of the cancer. *Gastroenterology.* 2010;139(4):1128-1137.
4. Kahi CJ, Imperiale TF, Juliar BE, Rex DK. Effect of screening colonoscopy on colorectal cancer incidence and mortality. *Clin Gastroenterol Hepatol.* 2009;7(7):770-775; quiz 711.
5. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med.* 2014;370(14):1287-1297.

6. Knudsen AB, Zauber AG, Rutter CM, et al. Estimation of Benefits, Burden, and Harms of Colorectal Cancer Screening Strategies: Modeling Study for the US Preventive Services Task Force. *Jama.* 2016;315(23):2595-2609.
7. Ladabaum U, Mannalithara A. Comparative Effectiveness and Cost Effectiveness of a Multitarget Stool DNA Test to Screen for Colorectal Neoplasia. *Gastroenterology.* 2016;151(3):427-439.e426.
8. Challen GA, Goodell MA. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood.* 2020;136(14):1590-1598.
9. Chen KZ, Kazi R, Porter CC, Qu CK. Germline mutations: many roles in leukemogenesis. *Curr Opin Hematol.* 2020;27(4):288-293.
10. SanMiguel JM, Young K, Trowbridge JJ. Hand in Hand: Intrinsic and Extrinsic Drivers of Aging and Clonal Hematopoiesis. *Exp Hematol.* 2020.
11. Severson EA, Riedlinger GM, Connelly CF, et al. Detection of clonal hematopoiesis of indeterminate potential in clinical sequencing of solid tumor specimens. *Blood.* 2018;131(22):2501-2505.
12. Steensma DP. Clinical consequences of clonal hematopoiesis of indeterminate potential. *Blood Adv.* 2018;2(22):3404-3410.
13. Swierczek SI, Agarwal N, Nussenzveig RH, et al. Hematopoiesis is not clonal in healthy elderly women. *Blood.* 2008;112(8):3186-3193.
14. Gao T, Ptashkin R, Bolton KL, et al. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. *Nat Commun.* 2021;12(1):338.
15. Martincorena I, Fowler JC, Wabik A, et al. Somatic mutant clones colonize the human esophagus with age. *Science.* 2018;362(6417):911-917.
16. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell.* 2013;153(6):1194-1217.
17. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med.* 2014;371(26):2488-2498.
18. Xie M, Lu C, Wang J, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med.* 2014;20(12):1472-1478.
19. Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med.* 2014;371(26):2477-2487.
20. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell.* 2018;173(7):1823.
21. Jaiswal S, Natarajan P, Silver AJ, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med.* 2017;377(2):111-121.
22. Coombs CC, Gillis NK, Tan X, et al. Identification of Clonal Hematopoiesis Mutations in Solid Tumor Patients Undergoing Unpaired Next-Generation Sequencing Assays. *Clin Cancer Res.* 2018;24(23):5918-5924.

23. Ghoneim HE, Fan Y, Moustaki A, et al. De Novo Epigenetic Programs Inhibit PD-1 Blockade-Mediated T Cell Rejuvenation. *Cell.* 2017;170(1):142-157.e119.
24. Wong WH, Bhatt S, Trinkaus K, et al. Engraftment of rare, pathogenic donor hematopoietic mutations in unrelated hematopoietic stem cell transplantation. *Sci Transl Med.* 2020;12(526).
25. Crowgey EL, Mahajan N, Wong WH, et al. Error-corrected sequencing strategies enable comprehensive detection of leukemic mutations relevant for diagnosis and minimal residual disease monitoring. *BMC Medical Genomics.* 2020;13(1):32.
26. Abbosh C, Swanton C, Birkbak NJ. Clonal haematopoiesis: a source of biological noise in cell-free DNA analyses. *Ann Oncol.* 2019;30(3):358-359.
27. Díaz-Gay M, Vila-Casadesús M, Franch-Expósito S, Hernández-Illán E, Lozano JJ, Castellví-Bel S. Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics.* 2018;19(1):224.
28. Bick AG, Weinstock JS, Nandakumar SK, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature.* 2020.
29. Sinicrope FA, Shi Q, Smyrk TC, et al. Molecular markers identify subtypes of stage III colon cancer associated with patient outcomes. *Gastroenterology.* 2015;148(1):88-99.
30. Phipps AI, Limburg PJ, Baron JA, et al. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology.* 2015;148(1):77-87.e72.
31. Temko D, Van Gool IC, Rayner E, et al. Somatic POLE exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response. *J Pathol.* 2018;245(3):283-296.
32. Castellsagué E, Li R, Aligue R, et al. Novel POLE pathogenic germline variant in a family with multiple primary tumors results in distinct mutational signatures. *Hum Mutat.* 2019;40(1):36-41.
33. Sun J, Wang C, Zhang Y, et al. Genomic signatures reveal DNA damage response deficiency in colorectal cancer brain metastases. *Nat Commun.* 2019;10(1):3190.
34. Dolcetti R, Viel A, Doglioni C, et al. High prevalence of activated intraepithelial cytotoxic T lymphocytes and increased neoplastic cell apoptosis in colorectal carcinomas with microsatellite instability. *Am J Pathol.* 1999;154(6):1805-1813.
35. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med.* 2015;372(26):2509-2520.
36. Ng AWT, Poon SL, Huang MN, et al. Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci Transl Med.* 2017;9(412).

Tables

Table 1

Demographic and clinical information

	Testing cohort		Validation cohort	
	CRN patients, n = 63	Risk-matched controls, n = 32	CRN patients, n = 20	Risk-matched controls, n = 10
Male, n (%)	30 (47.6)	21 (65.6)	7 (35.0)	5 (50.0)
Age, years (SD)	64.6 (9.6)	61.9 (6.4)	65.6 (9.9)	61.6 (5.6)
Location, n (%)				
Proximal	32 (50.8)	-	10 (50.0)	-
Distal	10 (15.9)	-	6 (30.0)	-
Rectum	21 (33.3)	-	4 (20.0)	-
Tumor size, mm (SD)	3.4 (1.7)	-	3.1 (1.3)	-
Histology, n (%)				
Advanced adenoma				-
Tubular	11 (17.5)	-	5 (25.0)	
Tubulovillous	31 (49.2)	-	11 (55.0)	-
Villous	3 (4.7)	-	0	-
Stage I cancer	18 (28.6)	-	4 (20.0)	-
FIT, n (%)				
Positive	35 (55.6)	0	10 (50.0)	0
Negative	28 (44.4)	32 (100)	10 (50.0)	10 (100)
FIT: fecal immunochemical test				

Table 2
ML-based prediction in validation cohort

Predictor	Reference	
	CRN	Controls
CRN	20	1
Controls	1	9
Accuracy (95% CI): 0.9333 (0.7793, 0.9918)		
P-value [Acc > NIR]: 0.00065		
Kappa: 0.85		
Sensitivity: 95.0 %		
Specificity: 90.0 %		

Figures

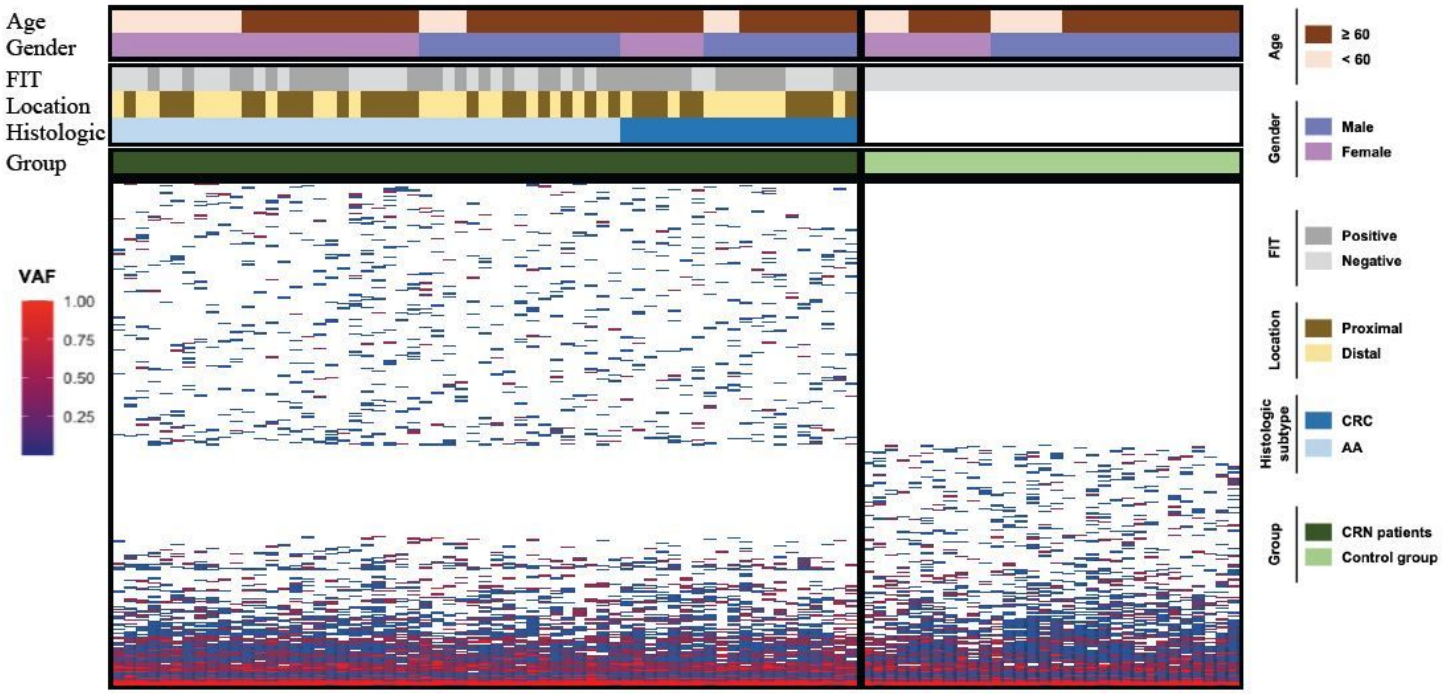

Figure 1

The mutational spectrum of CH in CRN patients and risk-matched controls The level of VAF is addressed from dark blue to red as low-allele fraction to high-allele fraction variants. The clinicopathological features are noted for each individual. FIT: fecal immunochemical test; AA: advanced adenoma.

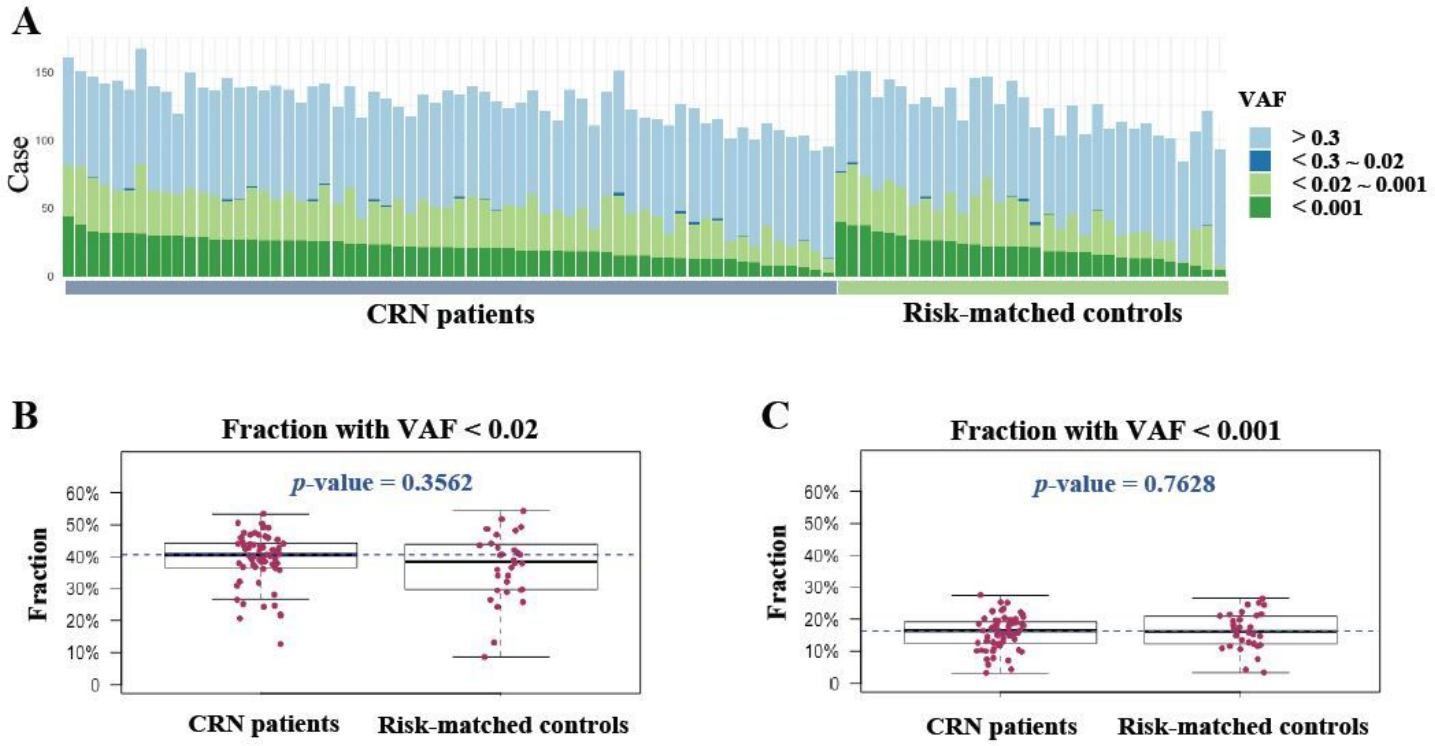

Figure 2

Genetic profiling of variants in CRN patients and risk-matched controls (A) Distribution of VAF in each individual was addressed in CRN and risk-matched control groups. The VAFs of PBMCs were determined by error corrected sequencing with greater than 3,000X unique-depth. (B) The prevalence of VAF < 0.02 in CRN patients and risk-matched controls. (C) The prevalence of VAF < 0.001 in CRN patients and risk-matched controls. P-values were calculated using two-sample T test with Welch's correction in (B) and (C).

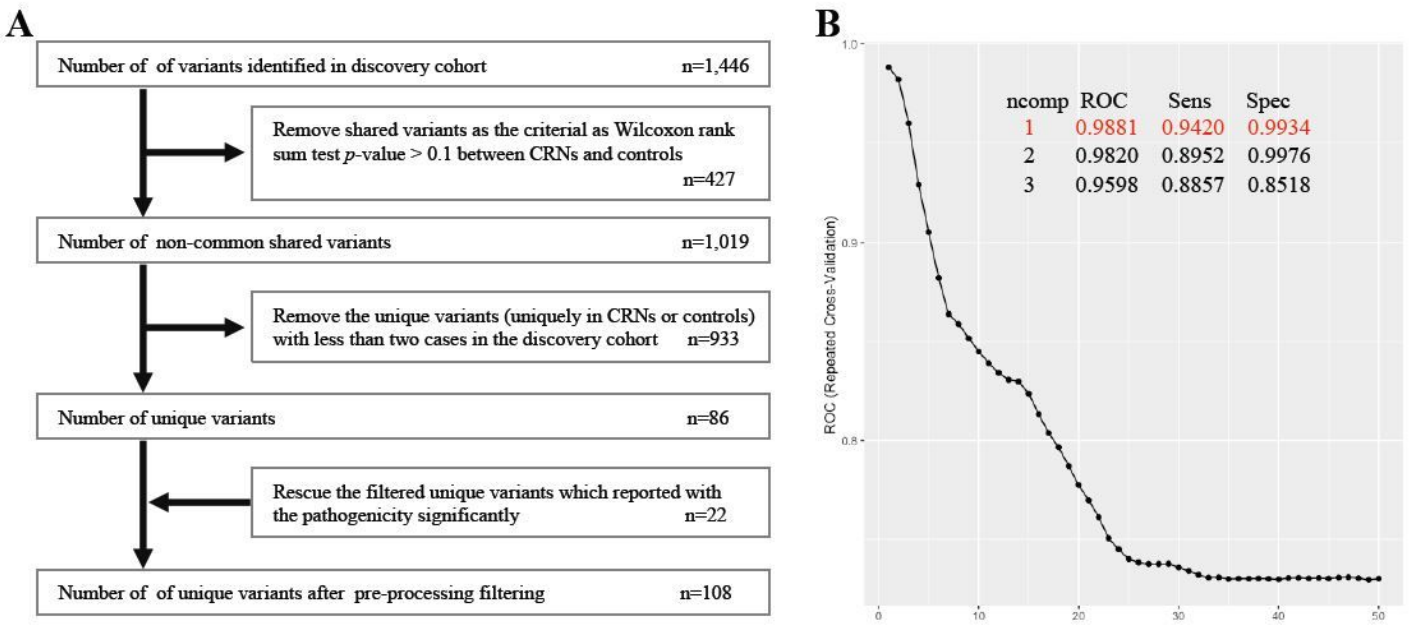

Figure 3

The flowchart and performance of model-training. (A) The diagram of pre-processing filtering. (B) The analysis method of machine-learning was based on a leave-one-out cross-validation framework. The resampling result across tuning parameters was listed.

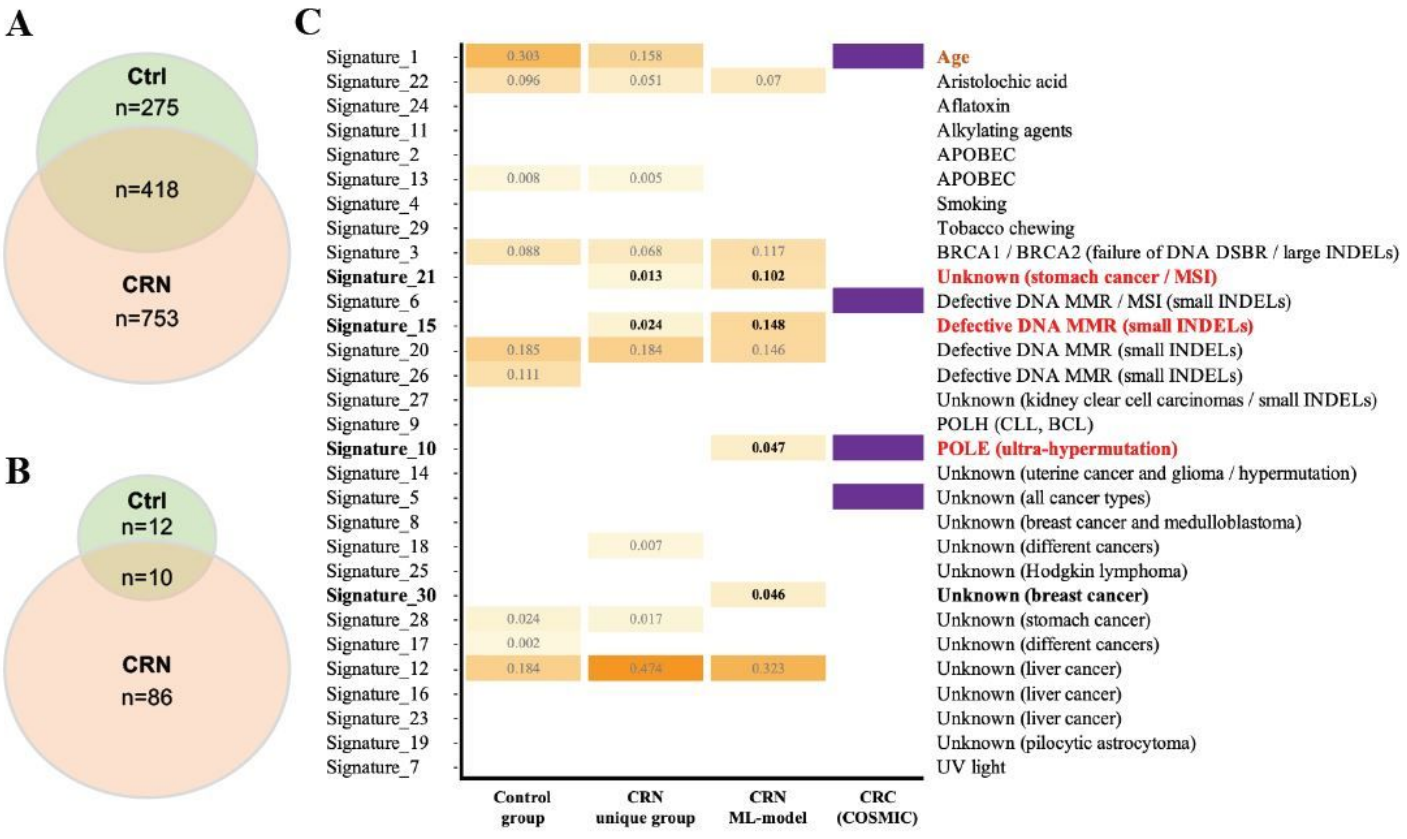

Figure 4

Characterization of the genetic landscape with mutation signature analysis. The COSMIC mutation signature v.2 was used for the characterization. (A) There are 753 unique variants (named CRN-unique group) identified in the CRN group and 693 variants (named control group) identified in the risk-matched control group. Among 693 variants, 418 variants are commonly found in both CRN and control groups. CRN-unique and control groups were subjected to COSMIC mutation signature v.2 analysis. (B) 86 variants resulted from pre-processing filtration for machine-learning approach in CRN-unique group was named as "CRN ML-model" group. c The COSMIC mutation signature v.2 analysis was shown with the level of contribution in Control group, CRN-unique group and CRN ML-model group. The purple patterns in CRC group indicated the specific mutational signatures in CRC released of COSMIC in 2015.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixHsuYC.docx