Identifying and Characterizing Genetic Variants Associated with Colorectal Cancer

Medha Kaul[1,2], Yao Yu, PhD[2], Ryan Bohlender, PhD[2], Chad Huff, PhD[2]

Johns Hopkins University Bloomberg School of Public Health[1], Baltimore, MD
Department of Epidemiology, The University of Texas MD Anderson Cancer Center[2], Houston, TX

THE UNIVERSITY OF TEXAS
MD Anderson
~~Cancer~~ Center
Making Cancer History®

Background

- Colorectal cancer (CRC) is the third most common cancer diagnosed in both men and women in the US[1].
- Several genes are known to affect CRC risk, but they only explain a small proportion of the disease heritability[2].

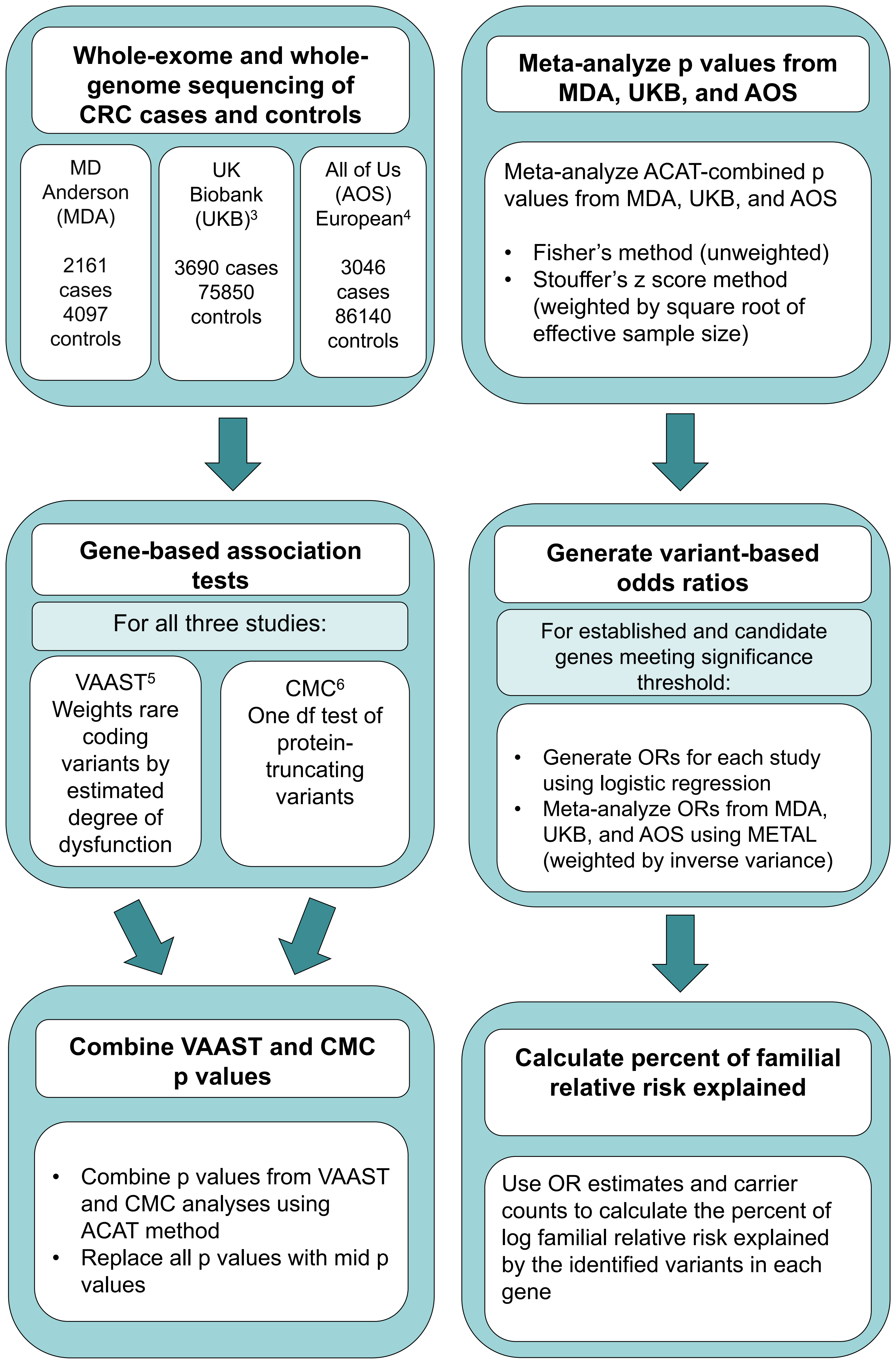
Hypothesis

The missing heritability of CRC is explained in part by undiscovered rare, intermediate-risk genetic variants.

Objectives

- Identify novel CRC susceptibility genes
- Produce risk estimates for variants in established CRC genes and novel candidate genes

Methods



Whole-exome and whole-genome sequencing of CRC cases and controls

MD Anderson (MDA)
2161 cases
4097 controls

UK Biobank (UKB)[3]
3690 cases
75850 controls

All of Us (AOS) European[4]
3046 cases
86140 controls

Gene-based association tests

For all three studies:

VAAST[5]
Weights rare coding variants by estimated degree of dysfunction

CMC[6]
One df test of protein-truncating variants

Combine VAAST and CMC p values

- Combine p values from VAAST and CMC analyses using ACAT method
- Replace all p values with mid p values

Meta-analyze p values from MDA, UKB, and AOS

Meta-analyze ACAT-combined p values from MDA, UKB, and AOS

- Fisher's method (unweighted)
- Stouffer's z score method (weighted by square root of effective sample size)

Generate variant-based odds ratios

For established and candidate genes meeting significance threshold:

- Generate ORs for each study using logistic regression
- Meta-analyze ORs from MDA, UKB, and AOS using METAL (weighted by inverse variance)

Calculate percent of familial relative risk explained

Use OR estimates and carrier counts to calculate the percent of log familial relative risk explained by the identified variants in each gene

Results

	MDA		UKB		AOS	
Gene	**VAAST pval**	**CMC pval**	**VAAST pval**	**CMC pval**	**VAAST pval**	**CMC pval**
MSH6	0.13	**0.012**	**$\leq 5.0 \times 10^{-7}$**	**$\leq 5.0 \times 10^{-7}$**	**$\leq 5.0 \times 10^{-5}$**	**2.3×10^{-3}**
MSH2	**3.5×10^{-5}**	**1.0×10^{-5}**	**2.0×10^{-4}**	**$\leq 5.0 \times 10^{-7}$**	**0.023**	**6.4×10^{-3}**
MLH1	**7.4×10^{-4}**	**1.8×10^{-4}**	**5.5×10^{-6}**	**$\leq 5.0 \times 10^{-7}$**	**5.9×10^{-3}**	0.69
APC	**3.7×10^{-3}**	**3.9×10^{-3}**	**0.024**	**5.6×10^{-3}**	**$\leq 5.0 \times 10^{-5}$**	**8.1×10^{-3}**
BRCA1	0.83	0.72	**2.5×10^{-6}**	**$\leq 5.0 \times 10^{-7}$**	**0.034**	0.13
CHEK2	0.053	0.34	**5.2×10^{-5}**	**6.5×10^{-6}**	0.54	0.39
BRCA2	0.85	0.74	**$\leq 5.0 \times 10^{-7}$**	**$\leq 5.0 \times 10^{-7}$**	0.38	0.14
ATM	**0.025**	0.066	0.087	**0.024**	**0.030**	0.21
PMS2	0.18	0.11	**0.013**	**0.019**	0.22	0.31
SDHA	**8.7×10^{-3}**	0.40	0.82	0.60	**0.037**	0.24
CDKN2A	0.99	0.99	**5.6×10^{-3}**	**0.023**	0.43	0.058
RAD51C	0.59	0.99	**6.3×10^{-4}**	**5.4×10^{-3}**	0.17	0.86

Table 1. Results of gene-based association analyses: VAAST and CMC p values from MDA, UKB, and AOS. Based on the meta-analysis results, the most significant established CRC genes and nominally significant a priori candidates with previous germline evidence for CRC were included in the table. Significant p values are bolded (genome-wide or nominal). P values with a ≤ sign are the smallest obtainable values given the number of permutations used in their respective analyses.

Stouffer's Genome-Wide Rank	Gene	MDA ACAT pval	UKB ACAT pval	AOS ACAT pval	Fisher's Meta-Analysis pval	Stouffer's Meta-Analysis pval
1	*MSH6* *	**0.021**	**$\leq 5.0 \times 10^{-7}$**	**9.8×10^{-5}**	**4.3×10^{-10}**	**7.5×10^{-11}**
2	*MSH2* *	**1.6×10^{-5}**	**1.0×10^{-6}**	**0.010**	**7.5×10^{-11}**	**1.0×10^{-10}**
3	*MLH1* *	**2.9×10^{-4}**	**9.2×10^{-7}**	**0.012**	**1.2×10^{-9}**	**8.4×10^{-10}**
4	*APC* *	**3.8×10^{-3}**	**9.1×10^{-3}**	**9.9×10^{-5}**	**7.2×10^{-7}**	**3.1×10^{-7}**
6	*BRCA1*	0.79	**8.3×10^{-7}**	0.054	**5.9×10^{-6}**	**6.1×10^{-5}**
11	*CHEK2*	0.095	**1.2×10^{-5}**	0.46	**6.1×10^{-5}**	**2.9×10^{-4}**
12	*BRCA2*	0.81	**$\leq 5.0 \times 10^{-7}$**	0.22	**1.3×10^{-5}**	**3.7×10^{-4}**
38	*ATM*	**0.036**	**0.037**	0.052	**4.0×10^{-3}**	**1.6×10^{-3}**
202	*PMS2*	0.14	**0.015**	0.26	**0.020**	**0.011**
1230	*SDHA*	**0.017**	0.74	0.065	**0.028**	0.082
1642	*CDKN2A*	0.99	**9.0×10^{-3}**	0.11	**0.031**	0.11
2492	*RAD51C*	0.99	**1.1×10^{-3}**	0.55	**0.022**	0.17

Table 2. ACAT-combined p values from MDA, UKB, and AOS and meta-analysis p values. Based on the meta-analysis results, the most significant established CRC genes and nominally significant a priori candidates with previous germline evidence for CRC were included in the table. Significant p values are bolded (genome-wide or nominal). Genes with an asterisk reached genome-wide significance. P values with a ≤ sign are the smallest obtainable values given the number of permutations used in the VAAST and CMC analyses.

Gene	Control Carrier Frequency (%)	OR (95% CI)	Percent log FRR[1] explained (%)
MSH6			
Truncating	0.0820	**6.28 (4.23, 9.34)**	0.558
Pathogenic missense	0.0244	**4.01 (1.37, 11.7)**	0.0652
MSH2			
Truncating	0.0233	**16.4 (8.08, 33.4)**	0.688
Pathogenic missense	0.0199	2.90 (0.879, 9.59)	0.0236
MLH1			
Truncating	0.0299	**9.51 (5.13, 17.6)**	0.417
Pathogenic missense	0.0543	**4.75 (2.54, 8.91)**	0.212
APC			
Truncating	0.0310	**15.0 (8.44, 26.5)**	0.817
Pathogenic missense	0.240	1.12 (0.683, 1.84)	0.00138
BRCA1			
Truncating and pathogenic missense	0.297	1.38 (0.893, 2.14)	0.0163
CHEK2			
Truncating and pathogenic missense	0.810	1.23 (0.951, 1.58)	0.0157
BRCA2			
Truncating and pathogenic missense	0.437	**1.45 (1.06, 1.99)**	0.0333

Table 3. Meta-analyzed variant-based odds ratios and percent of log familial relative risk explained. Based on the meta-analysis results, the most significant established CRC genes and nominally significant a priori candidates with previous germline evidence were included in the table. Significant ORs are bolded.

[1]FRR= Familial Relative Risk

Gene	Control Carrier Frequency (%)	OR (95% CI)	Percent log FRR[1] explained (%)
ATM			
Truncating and pathogenic missense	0.417	**1.65 (1.19, 2.28)**	0.0638
PMS2			
Truncating and pathogenic missense	0.248	**1.74 (1.26, 2.40)**	0.0492
SDHA			
Truncating and pathogenic missense	0.131	**1.65 (1.07, 2.55)**	0.0203
CDKN2A			
Truncating and pathogenic missense	0.0521	**2.91 (1.13, 7.50)**	0.0622
RAD51C			
Truncating	0.107	1.47 (0.587, 3.68)	0.00886

Table 3 Continued

Discussion

- No novel genes reached genome-wide significance (2.5×10^{-6}).
- Of the a priori candidate genes with previous germline evidence for CRC, ***BRCA1/2*[7], *ATM*[7], *SDHA*[8], *CDKN2A*[7], and *RAD51C*[8] were nominally significant in the gene-based meta-analysis, supporting their potential association with CRC.**
- Few a priori candidates replicated in the gene-based analyses. If these are susceptibility genes, they may explain a modest proportion of FRR and likely require larger sample sizes for detection.
- Of the known CRC genes:
  - For *MSH2*, the effect size of pathogenic missense variants was attenuated relative to truncating variants, though the CIs' overlapped
  - *APC* pathogenic missense variants conferred non-significant risk, indicating potential false-positive classifications in ClinVar
  - *PMS2* truncating and pathogenic missense variants conferred only moderate risk despite it being a Lynch Syndrome gene[9] (OR= 1.74, 95% CI= 1.26, 2.40).
- While gene-based association results support *CHEK2* and candidates *BRCA1/2, ATM, SDHA*, and *RAD51C* as CRC susceptibility genes, their effect size estimates indicate that truncating and pathogenic missense variants confer at most a modest increase in risk (ORs ranging from 1.23-1.65)
- **Overall, the rare coding variants from the genes highlighted in this study explain approximately 3.1% of the log familial relative risk of colorectal cancer.**
- These findings will lead to more accurate CRC risk prediction models with clinical utility to aid early detection.

Responsible Conduct of Research

- Limited representation of diverse groups in MDA and UKB may exacerbate disparities in understanding genetic basis of CRC.
- These disparities will be explored through African, Asian, and Hispanic representation in AOS.

References

1) Colorectal cancer statistics: How common is colorectal cancer? American Cancer Society. Accessed August 9, 2023. https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html.
2) Schubert SA, Morreau H, de Miranda NFCC, van Wezel T. The missing heritability of familial colorectal cancer. Mutagenesis. 2020;35(3):221-231. doi:10.1093/mutage/gez027
3) Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12(3). doi:10.1371/journal.pmed.1001779
4) The "All of Us" Research Program. New England Journal of Medicine. 2019;381(7):668-676. doi:10.1056/nejmsr1809937
5) Yandell M, Huff C, Hu H, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011;21(9):1529-1542. doi:10.1101/gr.123158.111
6) Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am J Hum Genet. 2008;83(3):311-321. doi:10.1016/j.ajhg.2008.06.024
7) Pearlman R, Frankel WL, Swanson B, et al. Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. JAMA Oncol. 2017;3(4):464-471. doi:10.1001/jamaoncol.2016.5194
8) Xu T, Zhang Y, Zhang J, et al. Germline Profiling and Molecular Characterization of Early Onset Metastatic Colorectal Cancer. Front Oncol. 2020;10. doi:10.3389/fonc.2020.568911
9) Jang E, Chung DC. Hereditary colon cancer: Lynch syndrome. Gut Liver. 2010;4(2):151-160. doi:10.5009/gnl.2010.4.2.151

Acknowledgements: *This work was supported by NIH/NCI R25CA056452 (Medha Kaul, Shine Chang, Ph.D., Principal Investigator).* Thank you to my PI, Dr. Chad Huff, and my lab team members Yao and Ryan for all the guidance and support on this project.