Transcriptome-wide Mendelian randomisation exploring dynamic CD4+ T cell gene expression in colorectal cancer development

Benedita Deslandes[1,2,a], Xueyan Wu[3,4], Matthew A. Lee[5], Lucy J. Goudswaard[1,2], Gareth W. Jones[6], Andrea Gsur[7], Annika Lindblom[8,9], Shuji Ogino[10,11,12], Veronika Vymetalkova[13], Alicja Wolk[14], Anna H. Wu[15], Jeroen R. Huyghe[16], Ulrike Peters[16,17], Amanda I. Phipps[16,17], Claire E. Thomas[16], Rish K. Pai[18], Robert C Grant[19], Daniel D. Buchanan[20,21,22], James Yarmolinsky[23], Marc J. Gunter[5,24], Jie Zheng[3,4*], Emma Hazelwood[1,2*], Emma E. Vincent[1,25*]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[3]Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China
[4]Shanghai National Clinical Research Center for Metabolic Diseases, Key Laboratory for Endocrine and Metabolic Diseases of the National Health Commission, Shanghai National Center for Translational Medicine, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China
[5]Nutrition and Metabolism Branch, International Agency for Research on Cancer, WHO, Lyon, France
[6]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK
[7]Center for Cancer Research, Medical University of Vienna, Vienna, Austria.
[8]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden.
[9]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden.
[10]Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

[11]Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
[12]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
[13]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic.
[14]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
[15]University of Southern California, Department of Population and Public Health Sciences, Los Angeles, California, USA.
[16]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, Washington, USA.
[17]Department of Epidemiology, University of Washington, Seattle, Washington, USA
[18]Department of Laboratory Medicine and Pathology, Mayo Clinic Arizona, Scottsdale, AZ, USA
[19]Division of Medical Oncology and Hematology, Princess Margaret Cancer Centre, Toronto, Canada.
[20]Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, VIC 3010, Australia
[21]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, VIC 3010, Australia
[22]Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Melbourne, VIC 3000, Australia
[23]Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, United Kingdom
[24]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, UK
[25]Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
*These authors contributed equally and are joint senior authors.
[a]Corresponding author (benedita.deslandes@bristol.ac.uk)

Abstract

Background

Recent research has identified a potential protective effect of higher numbers of circulating lymphocytes on colorectal cancer (CRC) development. However, the importance of different lymphocyte subtypes and activation states in CRC development and the biological pathways driving this relationship remain poorly understood and warrant further investigation. Specifically, CD4+ T cells – a highly dynamic lymphocyte subtype – undergo remodelling upon activation to induce the expression of genes critical for their effector function. Previous studies investigating their role in CRC risk have used bulk tissue, limiting our current understanding of the role of these cells to static, non-dynamic relationships only.

Methods

Here, we combined two genetic epidemiological methods – Mendelian randomisation (MR) and genetic colocalisation – to evaluate evidence for causal relationships of gene expression on CRC risk across multiple CD4+ T cell subtypes and activation stage. Genetic proxies were obtained from single-cell transcriptomic data, allowing us to investigate the causal effect of expression of 1,805 genes across five CD4+ T cell activation states on CRC risk (78,473 cases; 107,143 controls). We repeated analyses stratified by CRC anatomical subsites and sex, and performed a sensitivity analysis to evaluate whether the observed effect estimates were likely to be CD4+ T cell-specific.

Results

We identified six genes with evidence (FDR-P<0.05 in MR analyses and H4>0.8 in genetic colocalisation analyses) for a causal role of CD4+ T cell expression in CRC development – *FADS2, FHL3, HLA-DRB1, HLA-DRB5, RPL28,* and *TMEM258*. We observed differences in causal estimates of gene expression on CRC risk across different CD4+ T cell subtypes and activation timepoints, as well as CRC anatomical subsites and sex. However, our sensitivity analysis revealed that the genetic proxies used to instrument gene expression in CD4+ T cells also act as eQTLs in other tissues, highlighting the challenges of using genetic proxies to instrument tissue-specific expression changes.

Conclusions

Our study demonstrates the importance of capturing the dynamic nature of CD4+ T cells in understanding disease risk, and prioritises genes for further investigation in cancer prevention research.

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer-related deaths globally (1). While several lifestyle and environmental risk factors, such as obesity, alcohol consumption, and tobacco use, have been identified, the underlying biological pathways driving CRC development remain poorly understood (2). This gap in knowledge has limited the development of effective preventative or therapeutic interventions for CRC. A growing body of evidence suggests that circulating white blood cell (WBC) counts are linked to CRC risk, progression, severity, and mortality (3–7). WBC counts, commonly used in clinical practice as a marker of immune system and overall health, can be subdivided into basophils, eosinophils, lymphocytes, monocytes, and neutrophils (8). We have previously shown a potential protective effect of higher numbers of circulating lymphocytes on CRC risk (9). However, lymphocytes are a heterogeneous group of immune cells with distinct functional roles, and it remains unclear which specific subtypes mediate this protective effect. This highlights the need for further research into the underlying mechanisms.

Among lymphocytes, CD4+ T cells are well-known for their role in anti-tumour responses either through either direct action and/or by recruitment of other immune cells (10,11). These cells are strong candidates for contributing to the protective effect of lymphocyte count against CRC risk. Upon activation, CD4+ T cells undergo extensive gene expression remodelling to shape their effector function (12,13). This activation process occurs through distinct stages (Figure 1a), beginning in a resting state (0h), followed by a minimally active stage post-activation ("lowly active" [LA]), progressing through cell division (16h post-activation), post-division (40h post-activation), and culminating in the acquisition of effector functions (5d post-activation) (12). Each of these stages is characterized by unique functional and transcriptional profiles critical for immune surveillance and response.

CD4+ T cells also exhibit significant heterogeneity across their subtypes (12). Circulating CD4+ T cell subtypes can broadly be grouped into three functional clusters: i) naïve T cells (CD4 naïve) – cells that have not yet encountered an antigen; ii) memory T cells (CD4 memory) – essential for rapid and robust responses to previously encountered antigens; iii) regulatory T cells (nTreg) – pivotal for maintaining immune homeostasis by supressing excessive immune responses **(Figure 1b)** (12). Not all CD4+ T cells undergo all activation stages given some exist in a terminally differentiated state (refer to supplementary table 1 for more information). Despite their critical and distinct roles, most studies examining CD4+ T cells in CRC have relied on bulk tissue analyses, such as whole blood, which fail to capture the dynamic activation and gene expression changes of these cells. This limitation obscures the complexity of gene expression changes and hinders our ability to identify the biological mechanisms driving the protective effects of lymphocytes on CRC development.

Here, we aimed to identify key immune-related genetic drivers of CRC risk with potential to inform novel prevention or therapeutic strategies. To achieve this, we leveraged summary genetic data capturing associations between germline variants and single-cell transcriptomic data, allowing us to investigate gene expression across dynamic CD4+ T cell subtypes and activation timepoints. We performed transcriptome-wide Mendelian randomisation analyses (MR), which, under certain assumptions, can provide causal estimates (14), to identify genes that may play a role in CRC development based on single-cell transcriptomic data generated previously from CD4+ T cells (12). We then performed genetic colocalisation to evaluate possible misinference due to linkage disequilibrium and add robustness to our results. Additionally, we applied an additional sensitivity analysis to determine whether the genetic instruments used to proxy gene expression in CD4+ T cells in MR analyses are also associated with gene expression in other tissues, such as colorectal tissue. Our approach combining causal methodologies with single-cell transcriptomic data not only advances our understanding of the protective roles of CD4+ T cells in CRC risk, but also highlights potential therapeutic targets for CRC prevention.

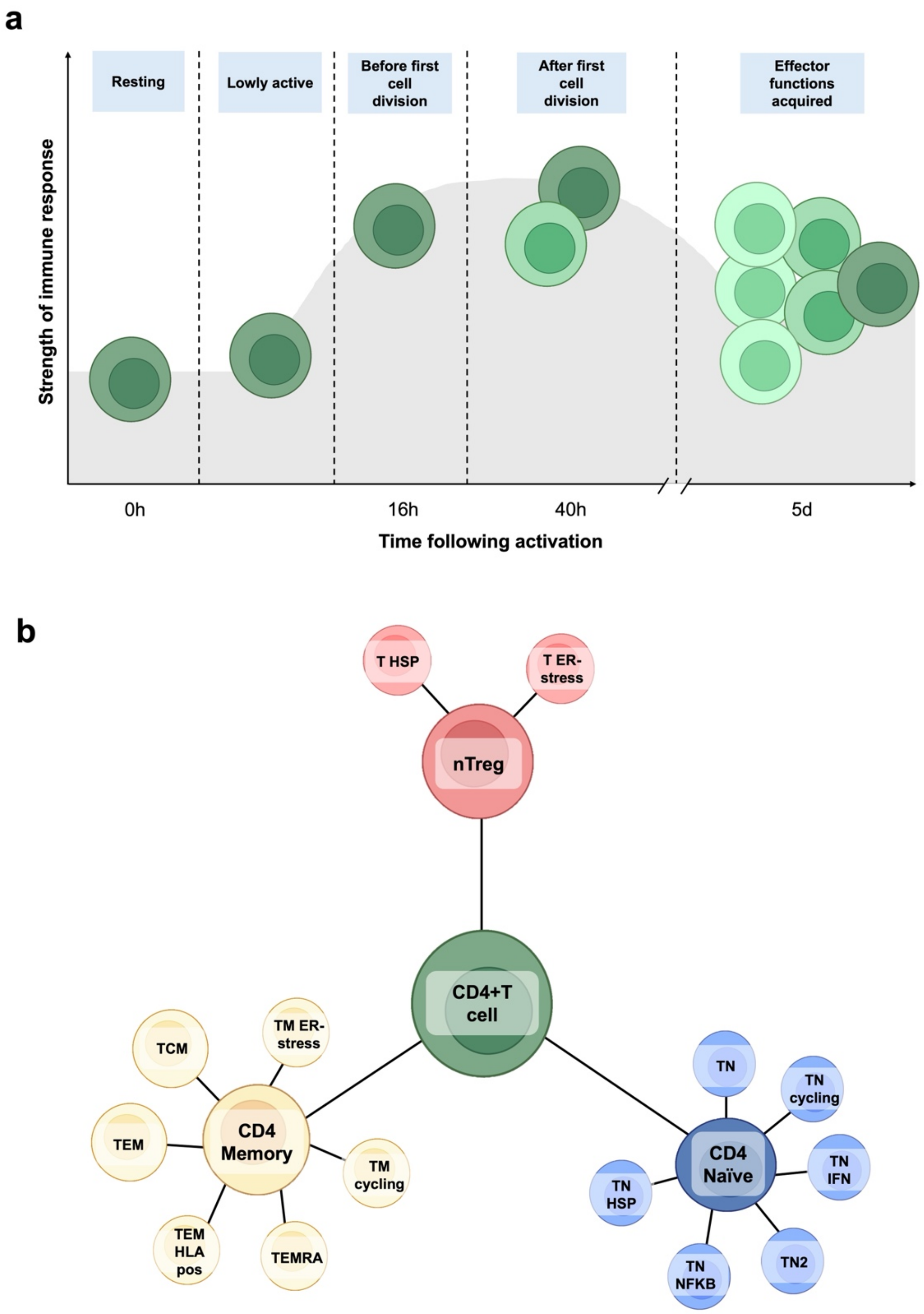

Figure 1. Overview of CD4+ T cell activation states and subtypes **a)** CD4+ T cells undergo complete remodelling of gene expression to shape their effector function upon activation. This activation occurs in distinct stages, progressing from a resting stage (0h), to minimally active

cells following activation ("lowly active" [LA]), before undergoing cell division (16h post-activation), after completion of the first cell division (40h post-activation), and after acquiring effector functions (5d post-activation). Each activation timepoint, or state, reflects a unique functional and transcriptional profile crucial for immune surveillance and response. Colour shadings reflect transcriptomic changes. **b)** Circulating CD4+ T cell subtypes can broadly be grouped into three functional clusters: i) naïve T cells (CD4 naïve) – cells that have not yet encountered an antigen; ii) memory T cells (CD4 memory) – essential for rapid and robust responses to previously encountered antigens; iii) regulatory T cells (nTreg) – pivotal for maintaining immune homeostasis by supressing excessive immune responses.

Methods

To identify key gene expression alterations across subtypes of CD4+ T cells with a role in CRC development, we conducted summary-level MR analyses. To evaluate the robustness of our results to misinference by linkage disequilibrium (LD), we then performed genetic colocalisation. For genes with robust evidence for an effect on CRC risk across both analyses, we performed a sensitivity analysis to evaluate the possibility of bias arising from horizontal pleiotropy through gene expression changes in other tissues. An overview of our methodology is represented in **Figure 2.**

Study populations

(i) CD4+ T cell gene expression genome-wide association study (GWAS)

Soskic *et al.* isolated peripheral mononuclear cells (PMBCs) from blood samples obtained from 119 healthy individuals (56% male; 44% female) of European ("British") ancestries (mean age 47 ± 15.61). CD4+ T cells were then further isolated from the PBMCs and activated with anti-CD3/anti-CD28 human T-Activator Dynabeads (Invitrogen) at a 1:2 beads-to-cells ratio. This resulted in a total of 655,349 CD4+ T cells across all individuals available for the single-cell transcriptomics (12).

We used summary-level expression quantitative trait loci (eQTL) results derived from single-cell transcriptomic data (12). These eQTLs were identified from the transcriptomes of CD4+ T cells sampled at five distinct timepoints following the *ex vivo* activation described above, representing a spectrum from resting to activated states **(Figure 1a)**. The timepoints included resting cells (0h), lowly active (LA), before undergoing cell division (16h post-activation), after the first cell division (40h post-activation), and after acquiring effector functions (5d post-activation) (12). Additionally, unsupervised clustering was performed by Soskic *et al.* using the transcriptomic data to group cells based on gene expression patterns and activation markers across the activation timepoints. This process identified 17 distinct CD4+ T cell subtypes which fall into three main groups – CD4 memory, CD4 naïve and nTreg **(Figure 1b)**. Combining the 17 cell types with different activation timepoints, Soskic *et al.* described a total of 46 distinct CD4+ T cell gene expression profiles (e.g. CD4_naive_0h represents CD4 naïve cells immediately following activation; note that all five activation timepoints are not available for every subtype, as they can become terminally activated before reaching the final timepoint). All CD4+ T cell subtypes included in our analyses are described in Supplementary table 1.

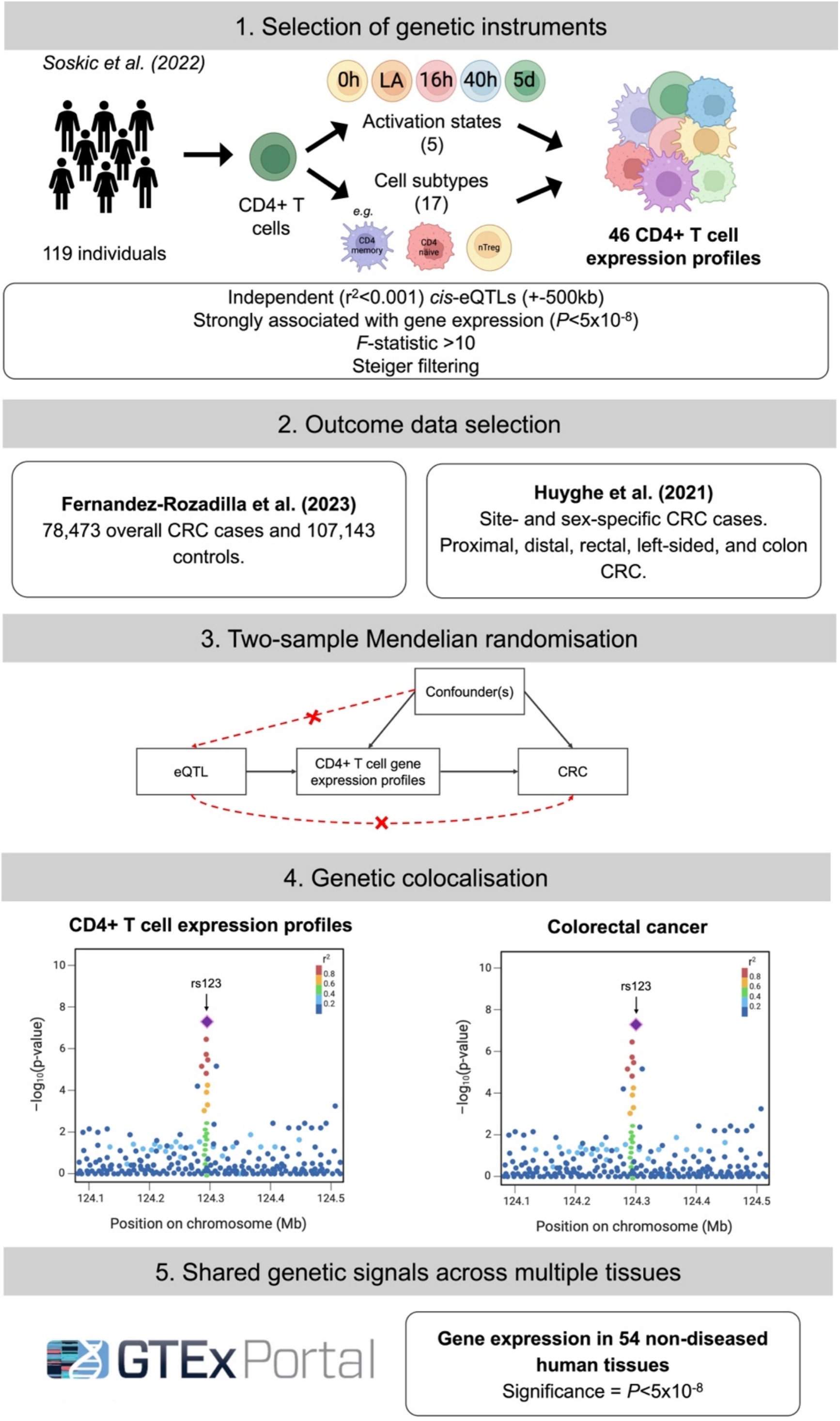
1. Selection of genetic instruments
Soskic et al. (2022)
119 individuals
CD4+ T cells
0h
LA
16h
40h
5d
Activation states (5)
Cell subtypes (17)
e.g.
CD4 memory
CD4 naïve
nTreg
46 CD4+ T cell expression profiles
Independent (r²<0.001) cis-eQTLs (+-500kb)
Strongly associated with gene expression (P<5x10⁻⁸)
F-statistic >10
Steiger filtering
2. Outcome data selection
Fernandez-Rozadilla et al. (2023)
78,473 overall CRC cases and 107,143 controls.
Huyghe et al. (2021)
Site- and sex-specific CRC cases.
Proximal, distal, rectal, left-sided, and colon CRC.
3. Two-sample Mendelian randomisation
Confounder(s)
eQTL
CD4+ T cell gene expression profiles
CRC
4. Genetic colocalisation
CD4+ T cell expression profiles
Colorectal cancer
rs123
r²
0.8
0.6
0.4
0.2
−log₁₀(p-value)
0
2
4
6
8
10
124.1
124.2
124.3
124.4
124.5
Position on chromosome (Mb)
5. Shared genetic signals across multiple tissues
GTEx Portal
Gene expression in 54 non-diseased human tissues
Significance = P<5x10⁻⁸

Figure 2. Flow chart describing study methodology. Note that genetic instruments were not available for all cell subtypes at every activation timepoint, meaning the total number of CD4+ T cell expression profiles investigated does not exactly equal the number of timepoints (5) multiplied by the number of cell subtypes (17). eQTL = expression quantitative trait loci; CRC = colorectal cancer; GTEx = Genotype-Tissue Expression.

(ii) CRC risk GWAS

We used summary-level data obtained from the largest available GWAS of CRC risk in European ancestries (15) (N cases = 78,473; N controls = 107,143) and the largest available

GWAS of sex-specific and site-specific CRC risk in European ancestries (16,17): proximal (14,416 cases, 43,099 controls), distal (12,879 cases, 43,099 controls), rectal (14,150 cases, 43,099 controls), left-sided (27,004 cases, 43,099 controls), colon (28,736 cases, 43,099 controls), female (24,594 cases, 23,936 controls), and male (28,271 cases, 22,351 controls). CRC classification was determined using ICD-10 codes, with the majority of cases being newly diagnosed. CRC subsites were categorized based on location: colon cancer includes the proximal colon (any primary tumour arising in the cecum, ascending colon, hepatic flexure, or transverse colon), the distal colon (any primary tumour arising in the splenic flexure, descending colon, or sigmoid colon), and cases with an unspecified site. Rectal cancer includes any primary tumour arising in the rectum or rectosigmoid junction.

MR analyses

MR uses genetic variants, typically single nucleotide polymorphisms (SNPs), which under specific assumptions can be used in an instrumental variable framework to obtain causal estimates. The three core assumptions are: (i) the genetic variant(s) must be associated with the exposure; (ii) there are no confounders of the association between the genetic variant(s) and the outcome; and (iii) the genetic variant(s) is/are only associated with the outcome via an association with the exposure (14). In addition to these core assumptions, additional assumptions, such as exposure and outcome data being obtained from non-overlapping populations from the same underlying population in summary-level MR, also exist (18); see Sanderson *et al.* (19) for a detailed overview of MR assumptions.

We identified *cis*-eQTLs for each gene as SNPs within the gene coding region (± 500kb) which had a P<5x10^{-8} and were independent of other associated SNPs within a 10kb window using a linkage disequilibrium (LD) r^2<0.001. We excluded weak instruments using an F-statistic<10 (20) and performed Steiger filtering (21) to exclude SNPs which may explain more variance in the exposure than the outcome, in order to avoid bias from potential reverse causation. As such, we identified 10,994 *cis*-SNPs associated with expression of 1,805 genes across the 46 CD4+ T cell gene expression profiles. For all MR analyses we used the Wald ratio to obtain causal estimates and the delta method to approximate standard errors (22). Benjamini-Hochberg correction (<0.05) was applied as a false discovery rate (FDR)-correction (23).

This manuscript was written following the STROBE-MR guidelines (24,25). A completed STROBE-MR checklist is included in the supplementary information.

Genetic colocalisation analyses

Genetic colocalisation uses GWAS summary statistics to distinguish between distinct causal variants underlying a shared causal signal at a specific locus for two (or more) traits (26). Genetic colocalisation evaluates the posterior probability of five mutually exclusive scenarios: H0: there are no variants in the given genomic region causal to either trait; H1: there is a causal variant in the given genomic region for the first but not second trait; H2: there is a causal variant in the given genomic region for the second but not the first trait; H3: there is a causal variant in the given genomic region for both traits, but this variant is different between the traits; and H4: the causal variant in the given genomic region is the same for both traits (26,27).

To evaluate the possibility of misinference due to LD in our MR analyses, we performed pair-wise conditional colocalisation (PWCoCo) for all genes meeting our predetermined threshold

(FDR-*P*<0.05) in MR analyses. Briefly, PWCoCo performs conditional and joint multi-SNP analysis (GCTA-COJO) to detect independent associations within a region. To achieve this, PWCoCo conditions each SNP on the sentinel SNP to identify conditionally independent SNPs for which colocalisation is then performed (28). Thus, in contrast with other genetic colocalisation methods, this approach retains the single causal variant assumption but allows for the testing of multiple causal variants within a genomic region. By combining our MR analyses with PWCoCo, we were therefore able to evaluate whether MR evidence was likely being driven by LD between distinct causal SNPs for gene expression and CRC risk; a possible violation of the third core MR assumption. Colocalisation was performed using PWCoCo for all SNPs within ±500kb of the gene coding region using prior probabilities ($p1=p2=1x10^{-5}$ and $p12=1x10^{-7}$) based on ~1,621 SNPs present within each window (suitable prior probabilities chosen based on an online calculator, see reference (29) and supplementary figure 1). We interpreted posterior probabilities as a scale of evidence for a shared causal variant and set a threshold of H4>0.8 as supporting evidence for colocalisation.

Shared genetic signals across multiple tissues

We evaluated whether evidence for a causal effect of gene expression on CRC risk from our MR and colocalisation analyses were specific to CD4+ T cells or whether the eQTLs used as genetic proxies may also instrument gene expression in other tissues. Where genetic instruments are associated with expression of the relevant gene in multiple tissues, this could suggest possible bias in our MR results from horizontal pleiotropy, a violation of the exclusion restriction assumption (i.e. the pathway from genetic instrument to CRC risk would not be through the presumed exposure of gene expression in CD4+ T cells specifically). This would be particularly pertinent if an eQTL was also associated with expression of the gene in the colon tissue itself, given the likely importance of local gene expression in disease development. To investigate this, we obtained summary statistics from the Genotype-Tissue Expression Program (GTEx) (30) for tissue-specific gene expression for all prioritised genes (i.e. those with FDR-*P*<0.05 in MR analyses and H4>0.8 in genetic colocalisation analyses). GTEx is a comprehensive public resource that maps how genetic variation influences gene expression across 54 non-diseased human tissues, based on samples from nearly 1,000 cadavers (sample size varies by tissue; see (30) for more information). We evaluated whether there was evidence for an association between the genetic instruments and expression of the instrumented gene in any of the available tissues in GTEx. We defined evidence of an association as genome-wide significance ($P<5x10^{-8}$). Where the genetic instrument used in MR analyses was not available in the GTEx dataset, we instead evaluated the SNP in highest LD with the eQTL which was available in GTEx data (minimum required r^2 = 0.8).

Results

Mendelian randomisation analyses

Results are given as odds ratio (OR) of CRC risk per standard deviation (SD) higher expression of the gene in CD4+ T cells. Of 1,805 genes (across the 46 CD4+ T cell gene expression profiles), expression of 61 genes had evidence (FDR-*P*<0.05) for a potential causal effect on CRC risk **(Figure 3**; Supplementary table 2)**.** Among these, the activation state and cellular subtype that contributed the most were CD4+ T cells 40h post-activation **(Figure 4a)** and CD4 naïve respectively **(Figure 4b)**.

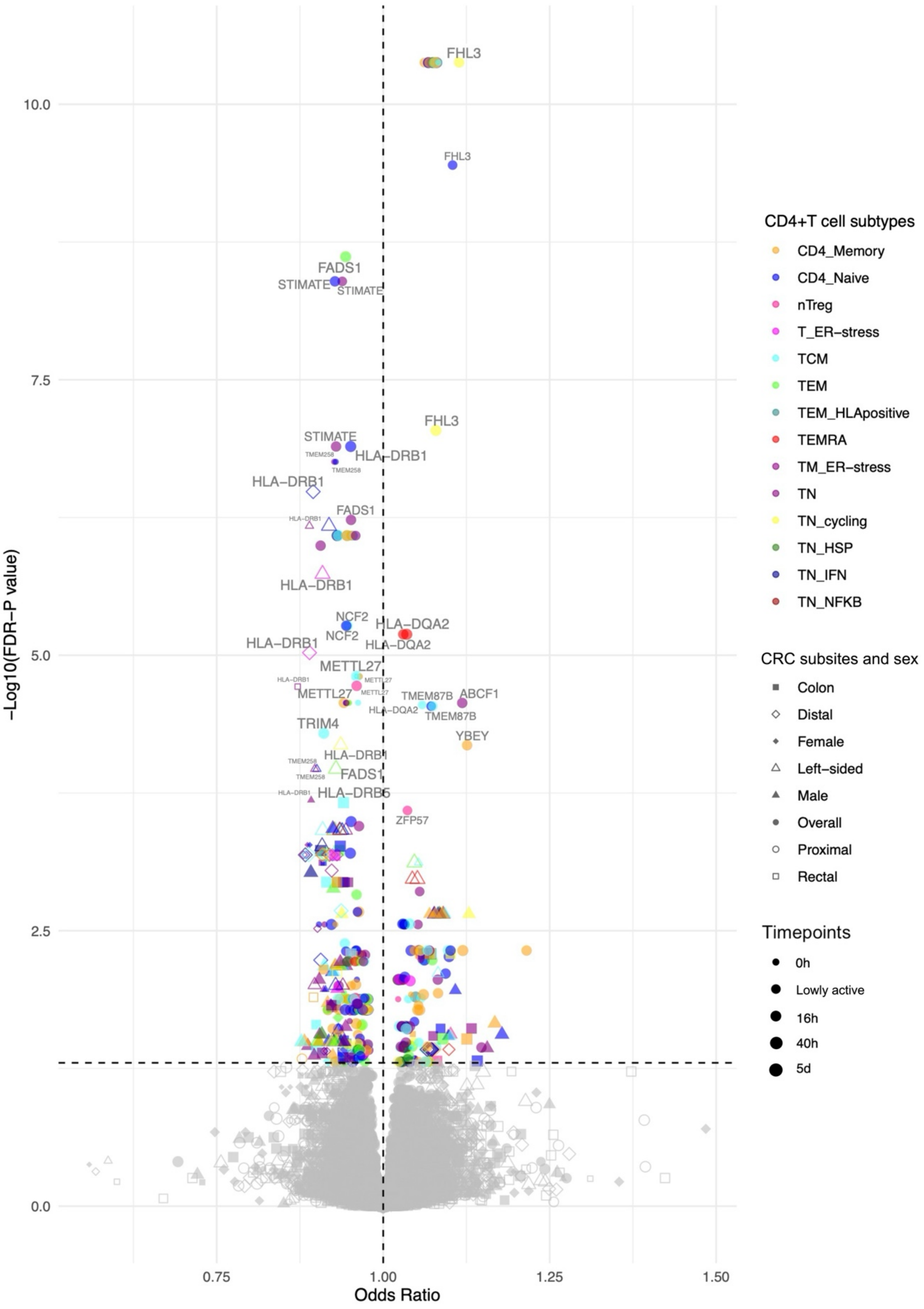

Figure 3. Volcano plot showing two-sample Mendelian randomisation results. Colours represent the different CD4+ T cell subtypes, shapes represent CRC subsites and sex, and point size represents activation state. Points labelled with gene names. Refer to supplementary table 1 for CD4+ T cell subtype definitions and abbreviations.

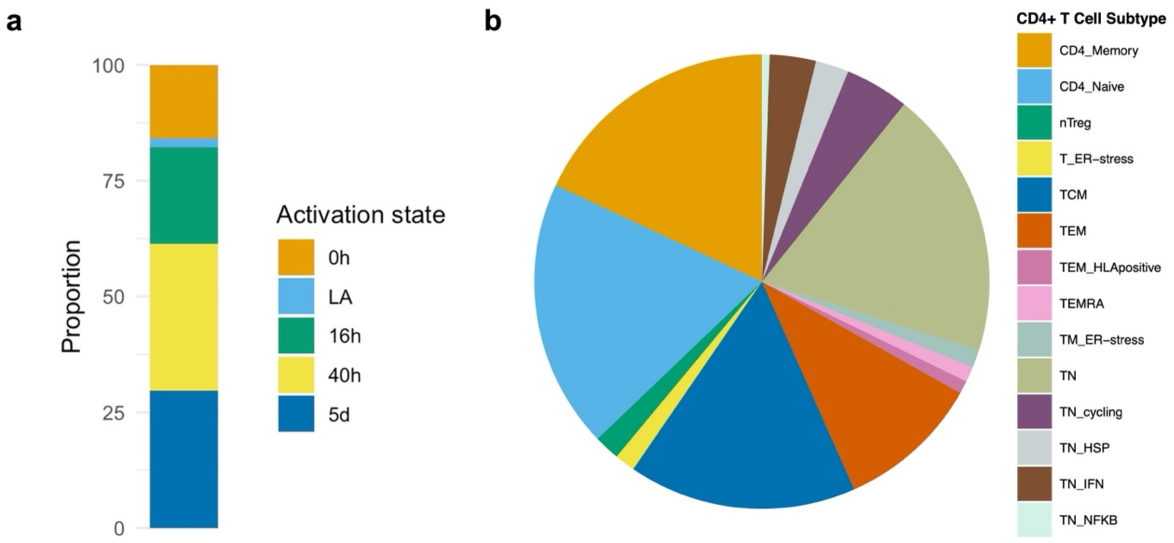

Figure 4. Proportion of CD4+T cell activation states and cell subtypes contributing to significant Mendelian randomisation results. **a)** Stacked bar plot representing proportion of CD4+ T cell activation states (0h, LA, 16h, 40h, 5d) contributing to main study. **b)** Pie chart representing proportion of CD4+ T cell subtypes contributing to main study results. Refer to supplementary table 1 for CD4+ T cell subtype definitions and abbreviations. LA = lowly active.

Genetic colocalisation analyses

Of the 61 genes with evidence for an association between their expression and CRC risk, six had evidence for a shared signal with CRC risk (H4>0.8): *FADS2, FHL3, HLA-DRB1, HLA-DRB5, RPL28, and TMEM258* (Table 1; full results in supplementary table 4).

Shared genetic signals across multiple tissues

For the six genes with robust evidence from MR (FDR-*P*<0.05) and genetic colocalisation (H4>0.8) analyses, we evaluated whether these observations could be explained through gene expression in tissues other than CD4+ T cells. Table 2 summarises the results from this sensitivity analysis.

The lead SNP and any SNPs in high LD for *HLA-DRB5* expression were not available in GTEx, meaning we were unable to include this gene in this sensitivity analysis. Overall, the lead SNPs for all five other genes with robust evidence for a role in CRC risk were associated with gene expression in other tissues in the GTEx dataset ($P<5\times10^{-8}$). The proportion of tissues in which eQTLs were associated with gene expression ranged from 52-73% (mean=64%), making it difficult to decipher the relevant tissue through which the identified genes act to influence CRC risk. For three of these genes (*FADS2, HLA-DRB1, RPL28*), the tissues where the eQTL associated with gene expression included transverse and sigmoid colon, highlighting a biologically plausible mechanism of horizontal pleiotropy in our analyses.

Table 1. Genetic colocalisation results.

		Gene					
		FADS2	*FHL3*	*HLA-DRB1*	*HLA-DRB5*	*RPL28*	*TMEM258*
	Overall		CD4 Memory (0h, 16h, 40h, 5d); TN cycling (40h); CD4 Naive (16h, 40h, 5d); TCM (0h, 16h, 5d); TN (16h, 40h, 5d); TN HSP (5d); TEM (40h)	TM ER-stress (40h); CD4 Naive (5d); TEM (0h)		TCM (0h); CD4 Naive (5d)	
Sex	Male	TN IFN (5d)	TCM (0h); CD4 Memory (0h, 16h); TN HSP (5d); CD4 Naive (40h); TN (40h); TEM (40h)	TM ER-stress (40h)			
	Female						CD4 Naive (0h); TN (0h)
CRC subsites	Colon				TCM (5d)		
	Distal			TM ER-stress (40h)			
	Left-sided			TM ER-stress (40h)			
	Proximal						
	Rectal						

Results of genetic colocalisation analysis using PWCoCo. CRC = colorectal cancer, PWCoCo = PairWise Conditional and Colocalisation; refer to supplementary table 1 for CD4+ T cell subtype definitions and abbreviations.

Table 2. Shared genetic signals across multiple tissues

	SNP		Genome-wide significant threshold* reached		
Gene	Lead	LD (R^2)	Sigmoid colon tissue	Transverse colon tissue	All tissues (%)
FADS2	rs61897793	-	Y	Y	61.7
FHL3	rs67631072	-	N	N	73.0
HLA-DRB1	rs3104393	rs9272025 (0.993)	Y	Y	69.6
HLA-DRB5	*NA*	*NA*	*NA*	*NA*	*NA*
RPL28	rs4806665	-	N	Y	61.1
TMEM258	rs174538	-	Y	N	52.0

Associations of eQTLs and gene expression in colon tissue with a comparison to all other tissues using GTEx data. GTEx = Genotype-Tissue Expression; LD = linkage disequilibrium;

SNP = single nucleotide polymorphism. *Significance defined as genome-wide threshold = <5x10^{-8}; *NA* indicates that both the lead SNP and a SNP in high LD was not available in the GTEx dataset with a suitable proxy.

Discussion

We used a causal framework employing MR and genetic colocalisation to investigate whether CD4+ T cell subtype- and activation timepoint-specific gene expression may have a causal role in CRC risk. We identified six genes (*FADS2, FHL3, HLA-DRB1, HLA-DRB5, RPL28,* and *TMEM258*) with robust evidence for a causal role of expression in CD4+ T cells in CRC risk. Notably, *TMEM258* has not been previously linked to CRC. Furthermore, several of these observations were specific to different CD4+ T cell activation states (e.g., prior to antigen presentation). However, the lead genetic variants associated with these six genes in CD4+ T cell subtypes at different activation points were also associated with expression in multiple other tissues, for some genes including colon tissues, suggesting the effect of gene expression on CRC risk might not be entirely specific to CD4+ T cells.

We observed evidence for sex-specific effects of gene expression in our analyses **(Figure 5 and supplementary table 3)**. For example, MR results suggested a protective effect of higher *TMEM258* expression on female-specific CRC risk in CD4 naïve cells (OR= 0.89, confidence interval (CI)= 0.85 to 0.93) and TN cells at rest (OR= 0.89, CI= 0.85 to 0.93) but not on male-specific CRC risk. *TMEM258* is involved in protein synthesis, folding and trafficking (31). Previous research has demonstrated that dysregulation of *TMEM258* expression can lead to endoplasmic reticulum (ER) stress, consequently triggering activation of the unfolded protein response (UPR) (31). UPR is known to be beneficial to CD4+ T cells as it supports differentiation, activation, cytokine production and autophagy (32). This may explain the potential mechanism by which increased *TMEM258* expression could reduce CRC risk. Additionally, we observed sex-specific effects for *FADS2,* as its expression was significantly protective against male-specific CRC development. *FADS2*, which encodes the enzyme fatty acid desaturase 2, plays a crucial role in the biosynthesis of polyunsaturated fatty acids (PUFAs), including omega-3 and omega-6 fatty acids (33). While omega-3 PUFAs have been shown to have protective associations against various cancers, including CRC, the relationship between omega-6 PUFAs and CRC risk remains unclear (34–38). We found evidence for a protective effect of *FADS2* expression on male CRC risk in naïve T cells producing interferon-gamma (IFN) upon activation (TN IFN) five days post-activation (OR=0.89, CI=0.85 to 0.93). Metabolic reconstruction, mediated by gene expression changes, is an important aspect of CD4+ T cell activity (39,40). Moreover, we observed null results for an effect estimate of *FADS2* expression in the same cell type 40 hours post-activation **(Figure 5)**, which supports an activation timepoint-specific effect of this gene and demonstrates the importance of considering the dynamic nature of CD4+ T cells.

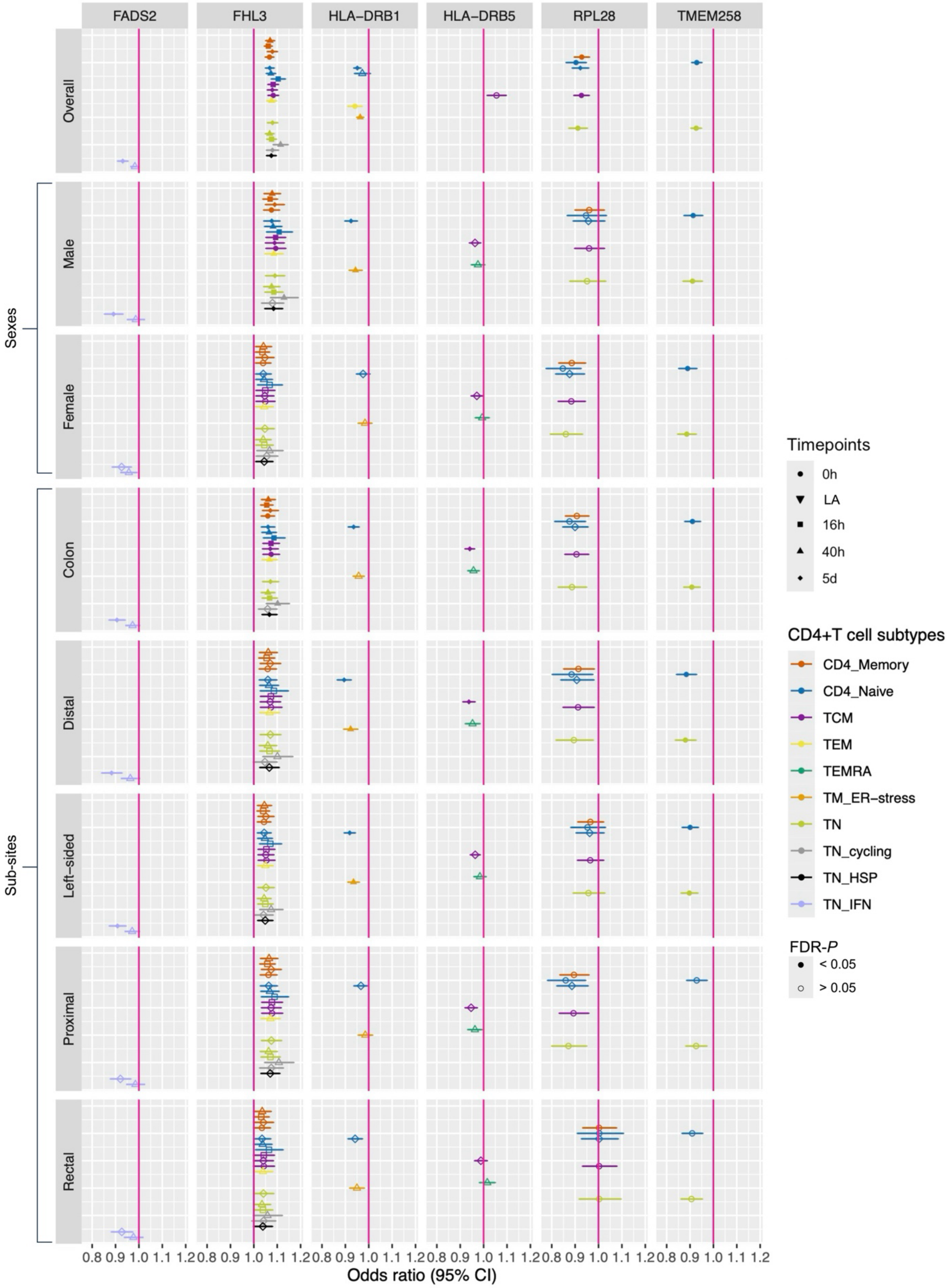

Figure 5. Mendelian randomisation results for genes that showed strong evidence of genetic colocalisation, separated by gene, sex and CRC subsites. Colours represent CD4+ T cell subtypes; shapes represent activation timepoint; and point fill represents whether FDR-*P* is < or > than 0.05. CRC = colorectal cancer; LA = lowly active.

In addition, we found evidence for a protective effect of two genes involved in the human leukocyte antigen (HLA) complex, *HLA-DRB1 and HLA-DRB5*, in CRC risk. We found evidence for a causal effect of higher *HLA-DRB1* expression on overall, male, distal and left-sided CRC risk in several subtypes of CD4+ T cells, particularly memory T cells experiencing ER stress (TM ER-stress) at 40h post-activation **(Figure 5)**. Cancer of the distal colon, which is part of left-sided CRC, is disproportionately diagnosed in males, which reflects these results. HLA-DRB1 is a class II major histocompatibility complex (MHC) protein involved in aiding the immune system distinguishing the body's own proteins from those made by foreign bodies (41). The HLA-DRB1 chain is a component of the MHC class II complex on the surface of antigen-presenting cells and is responsible for antigen presentation to CD4 T helper cells, thus supporting immune activation in response to peptides from pathogens (41). Our results suggest that the beneficial effect we observe of increased expression of this gene on CRC may be related to enhanced antigen presentation to CD4+ T cells, consequently boosting anti-tumour immunity. We found strong evidence for a protective effect of *HLA-DRB5* expression on colon-specific CRC risk in memory T cells (TCM) five days post-activation (OR= 0.94, CI= 0.92 to 0.96). Similar to *HLA-DRB1, HLA-DRB5* is an MHC class II protein (41). However, unlike *HLA-DRB1, HLA-DRB5* is only expressed in a subset of the population (42). This protein is involved in presenting peptides to T-helper cells, aiding in the initiation of an immune response (41). These findings reinforce the importance of immune vigilance/surveillance and its protective effect against CRC development (43). Higher *HLA-DRB5* expression may also enhance antigen presentation, resulting in stronger and more specific anti-tumour responses. Our results **(Figure 5)** show that higher *HLA-DRB5* expression in TCM cells 5 days post-activation has a protective effect risk of colon cancer, which aligns with this explanation. While we observed evidence for a protective effect of *HLA-DRB5* expression on colon-specific CRC risk in TCM cells, key players in long-term immune memory and response (44), we found little evidence for similar effect in effector memory cells re-expressing CD45RA (TEMRA). This highlights a cell type-specific relationship and underscores the importance of considering distinct subtypes of CD4+ T cells.

We observed a protective effect of higher *RPL28* expression on overall CRC risk in TCM cells at rest (OR = 0.92, CI = 0.89 to 0.96) and CD4 naïve cells 5 days post activation (OR = 0.89, CI = 0.82 to 0.96). The *RPL28* gene encodes for a ribosomal protein component of the large ribosome subunit (60S) and is involved in the assembly of ribosomes and the translation of messenger ribonucleic acid (mRNA) into proteins (45). As TCM cells play an important role in long-term immune memory and surveillance, *RPL28* may facilitate the production of necessary proteins that contribute to their maintenance and longevity. Similarly, CD4+ naïve T cells at 5 days post-activation require increased protein synthesis in order to support rapid proliferation and differentiation (12).

Finally, we identified strong evidence for a detrimental effect of higher *FHL3* expression on both overall and male-specific CRC risk across several subtypes of CD4+ T cells during rest and at three post-activation timepoints **(Figure 5)**. *FHL3* encodes the Four and a Half LIM Domains 3 (FHL3) protein, which is highly expressed in skeletal muscle. Although its specific function in this tissue is unclear (46), it is believed that the FHL family being localized to the nucleus, plays a critical role in transcription regulation (47). Differential expression of FHL3 could influence downstream gene expression. *FHL3* has been previously identified as a potential susceptibility gene for CRC, with abnormal expression patterns observed in several cancers. However, the direction of its association varies by cancer type (47–52). Given these findings, further research is necessary to elucidate the mechanisms underlying the detrimental effects suggested by our results.

Our study combined two complimentary methods which, taken together, provide evidence for a causal relationship that is robust to biases and confounding factors commonly associated with traditional epidemiological studies (14). Given the complexity and dynamic nature of CD4+ T cells, we aimed to identify genes with a role in CRC risk using single-cell data spanning a range of cell subtypes and activation points. However, several limitations to our analysis exist. First, we used data from European ancestries which, though we assume are homogenous and therefore satisfy our genetic instrument assumptions, means these results may not be generalizable to other populations. Second, sensitivity analyses revealed that genetic variants associated with gene expression in CD4+ T cells were also linked to gene expression in other tissues, suggesting that our findings may reflect broader tissue-level expression changes rather than being CD4+ T cell-specific. Third, two of the identified genes, *HLA-DRB1* and *HLA-DRB5*, are located within the MHC region on chromosome 6, a region known for high genetic variability and LD due to its complex genetic architecture. While biologically relevant, this variability may introduce biases to our study (53,54), consequently warranting cautious interpretation on findings related to these genes. Future methodologies may better resolve gene colocalisation within the MHC region (55). Fourth, though we investigated our outcome stratified by sex, it was not possible to do this for our exposures and it is unclear whether sex is an important factor in the genetic architecture of gene expression. Lastly, we acknowledge that, according to STROBE-MR guidelines, an ideal approach would include a replication dataset to validate our results, though this was not feasible given the novelty of the underlying data.

Conclusion

Our analysis identified six genes with robust evidence for a causal effect of expression in CD4+ T cell subtypes on CRC risk, including TMEM258, a gene not previously reported in relation to CRC development. This highlights its potential as a novel candidate for further research into CRC pathogenesis. Additionally, our findings revealed significant variability in causal estimates of CRC risk across different CD4+ T cell subtypes, activation time points, CRC anatomical subsites, and sex. These observations underscore the complex, context-dependent relationships between immune system dynamics and CRC risk.

List of abbreviations

CI	Confidence interval
CRC	Colorectal cancer
eQTL	Expression quantitative trait locus
ER	Endoplasmic reticulum
FADS	Fatty acid desaturase
FDR	False discovery rate
FHL3	Four and a Half LIM Domains 3
GCTA-COJO	Conditional and joint multi-SNP analysis
GECCO	Genetics and Epidemiology of Colorectal Cancer Consortium
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
HLA	Human leukocyte antigens
IFN	Interferon
LA	Lowly active
LD	Linkage disequilibrium
MHC	Major histocompatibility complex
MR	Mendelian randomisation
mRNA	Messenger ribonucleic acid
nTreg	Natural regulatory T cell
OR	Odds ratio
PUFAs	Polyunsaturated fatty acids
PWCoCo	PairWise Conditional and Colocalisation
RPL28	Ribosomal protein L28
rsID	Reference SNP cluster IDs
SD	Standard deviation
SNP	Single-nucleotide polymorphism
TMEM258	Transmembrane protein 258
UPR	Unfolded protein response
WBC	White blood cell

Declarations

Ethics approval and consent to participate

All GWAS obtained approval from the appropriate ethical committee(s) (12,15–17).

Availability of data and materials

Statistical analyses

The bulk of the analyses were performed using RStudio (version 2024.4.2.764) (56). MR analyses were performed using TwoSampleMR (version 0.5.7) (57). Data was manipulated using the following packages: arrow (version 16.1.0), biomaRt (version 2.58.0), data.table (version 1.14.10), dplyr (version 1.1.4), GenomicRanges (version 1.54.1), gwascat (version 2.34.0), gwasvcf (version 0.1.2), rtracklayer (version 1.62.0), stringr (version 1.5.1), tidyr (version 1.3.0) (58–68). Plots were created using ggrepel (version 0.9.5) and ggplot2 (version

3.4.4) (69,70). Allele frequencies were calculated using PLINK2.0 (71). Genetic colocalisation analyses were performed using Cmake (version 3.20.0) and PWCoCo (version 1.0) (28). Priors used in the colocalisation analyses were computed using link in reference (29).

Data and code availability

All data can be found in the manuscript, in the supplementary information, or in the links provided in the references. The GWAS of overall CRC in European ancestries can be accessed using the GWAS catalogue (https://www.ebi.ac.uk/gwas/) accession no. GCST90129505. The data where the genetic instruments were extracted for all MR analyses are available on this link (https://trynkalab.sanger.ac.uk). All code used to carry out analyses has been made publicly available on GitHub (https://github.com/bennydeslandes/CD4-T_cell_CRC). Further information on the TwoSampleMR package and PWCoCo can be found on https://github.com/MRCIEU/TwoSampleMR/, and https://github.com/jwr-git/pwcoco, respectively.

Competing interests

None to declare.

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Funding

BD is supported by a Wellcome Trust studentship (218495/Z/19/Z) at the University of Bristol. EH is supported by a Cancer Research UK Population Research Committee Studentship (C18281/A30905). BD, EH, and EV are supported by the CRUK Integrative Cancer Epidemiology Programme (C18281/A29019), and are part of the Medical Research Council Integrative Epidemiology Unit at the University of Bristol which is supported by the Medical Research Council (MC_UU_00032/03) and the University of Bristol. LJG is supported by a Cancer Research UK 25 (C18281/A29019) programme grant (the Integrative Cancer Epidemiology Programme). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Czech Republic CCS: This work was supported by the Grant Agency of the Ministry of Health of the Czech Republic (grants AZV NU22J-03-00033), and the project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102) - Funded by the European Union – Next Generation EU.

The Colon Cancer Family Registry (CCFR, www.coloncfr.org) is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). Support for case ascertainment was provided in part from the Surveillance, Epidemiology, and End Results (SEER) Program and the following U.S. state cancer registries: AZ, CO, MN, NC, NH; and by the Victoria Cancer Registry (Australia) and Ontario Cancer Registry (Canada). Additional funding for CCFR GWAS analysis was as follows: The CCFR Set-1 (Illumina 1M/1M-Duo) and Set-2 (Illumina Omni1-Quad) scans were supported by NIH awards U01 CA122839 and R01 CA143247 (to GC). The CCFR Set-3 (Affymetrix

Axiom CORECT Set array) was supported by NIH award U19 CA148107 and R01 CA81488 (to SBG). The CCFR Set-4 (Illumina OncoArray 600K SNP array) was supported by NIH award U19 CA148107 (to SBG) and by the Center for Inherited Disease Research (CIDR), which is funded by the NIH to the Johns Hopkins University, contract number HHSN268201200008I.

The Colon CFR graciously thanks the generous contributions of their study participants, dedication of study staff, and the financial support from the U.S. National Cancer Institute, without which this important registry would not exist. The content of this manuscript is solely the responsibility of the authors and does not necessarily reflect the views or policies of the NIH or any of the collaborating centers in the CCFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government, any cancer registry, or the CCFR.

Authors' contributions

BD: Data analysis – conducting MR and genetic colocalisation analysis, Writing – original draft, Writing – review and edit. XW: Data analysis – selecting genetic instruments, Writing – review and edit. MAL: Visualization – creating figures, Methodology – genetic colocalisation analyses, Writing – review and edit. LJG: Writing – review and edit. GWJ: Writing – review and edit. AG – data collection. AL – data collection. SO – data collection. VV – data collection. AW – data collection. AHW – data collection. JRH – data collection. UP – data collection. AIP – data collection. CET – data collection. JY: Methodology – genetic colocalisation analyses, Writing – review and edit. MJG: Writing – review and edit. JZ: Writing – review & editing, Writing – original draft, Methodology - conducting MR and genetic colocalisation analysis, Supervision, Project administration, Methodology and conceptualization. EH: Writing – review & editing, Writing – original draft, Methodology - conducting MR and genetic colocalisation analysis, Supervision, Project administration, Methodology and conceptualization. EEV: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology and conceptualization.

Acknowledgements

This work used the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bristol.ac.uk/acrc/. The authors would like to acknowledge the University of Bristol's High Performance Computing (HPC) team, without whom this work would not have been possible.

References

1. World Health Organization. Colorectal cancer [Internet]. 2023 [cited 2023 Oct 18]. Available from: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer
2. World Cancer Research Fund, American Institute for Cancer Research. Diet, Nutrition, Physical Activity and Colorectal Cancer: a Global Perspective. Continuous Update Project Expert Report 2018 [Internet]. 2018 [cited 2024 Oct 3]. Available from: dietandcancerreport.org
3. Wu J, Ge X, Zhu W, Zhi Q, Xu M, Duan W, et al. Values of applying white blood cell counts in the prognostic evaluation of resectable colorectal cancer. Mol Med Rep. 2019 Jan 10;
4. Prizment AE, Vierkant RA, Smyrk TC, Tillmans LS, Lee JJ, Sriramarao P, et al. Tumor eosinophil infiltration and improved survival of colorectal cancer patients: Iowa Women's Health Study. Modern Pathology. 2016 May;29(5):516–27.

5. Rosman Y, Hornik-Lurie T, Meir-Shafrir K, Lachover-Roth I, Cohen-Engler A, Munitz A, et al. Changes in peripheral blood eosinophils may predict colorectal cancer – A retrospective study. World Allergy Organization Journal. 2022 Oct;15(10):100696.
6. Liu Q, Luo D, Cai S, Li Q, Li X. Circulating basophil count as a prognostic marker of tumor aggressiveness and survival outcomes in colorectal cancer. Clin Transl Med. 2020 Jan 10;9(1).
7. Idos GE, Kwok J, Bonthala N, Kysh L, Gruber SB, Qu C. The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. Sci Rep. 2020 Feb 25;10(1):3360.
8. Nicholson LB. The immune system. Essays Biochem. 2016 Oct 31;60(3):275–301.
9. Constantinescu AE, Bull CJ, Jones N, Mitchell R, Burrows K, Dimou N, et al. Circulating white blood cell traits and colorectal cancer risk: A Mendelian randomisation study. Int J Cancer. 2023;
10. Tay RE, Richardson EK, Toh HC. Revisiting the role of CD4+ T cells in cancer immunotherapy—new insights into old paradigms. Cancer Gene Ther. 2021 Feb 27;28(1–2):5–17.
11. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. 2018 Oct 1;32(19–20):1267–84.
12. Soskic B, Cano-Gamez E, Smyth DJ, Ambridge K, Ke Z, Matte JC, et al. Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. Nat Genet. 2022 Jun 1;54(6):817–26.
13. Zhu J, Yamane H, Paul WE. Differentiation of Effector CD4 T Cell Populations. Annu Rev Immunol. 2010 Mar 1;28(1):445–89.
14. Davey Smith G, Ebrahim S. 'Mendelian randomisation': can genetic epidemiology contribute to understanding environmental determinants of disease?*. Int J Epidemiol. 2003 Feb;32(1):1–22.
15. Fernandez-Rozadilla C, Timofeeva M, Chen Z, Law P, Thomas M, Schmit S, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. Nat Genet. 2023 Jan 20;55(1):89–99.
16. Huyghe JR, Harrison TA, Bien SA, Hampel H, Figueiredo JC, Schmit SL, et al. Genetic architectures of proximal and distal colorectal cancer are partly distinct. Gut. 2021 Jul;70(7):1325–34.
17. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet. 2019 Jan 3;51(1):76–87.
18. Lawlor DA. Commentary: Two-sample Mendelian randomisation: opportunities and challenges. Int J Epidemiol. 2016 Jun;45(3):908–15.
19. Sanderson E, Glymour MM, Holmes M V., Kang H, Morrison J, Munafò MR, et al. Mendelian randomisation. Nature Reviews Methods Primers. 2022 Feb 10;2(1):6.
20. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomisation studies. Int J Epidemiol. 2011 Jun 1;40(3):755–64.
21. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. 2017 Nov 17;13(11):e1007081.
22. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomisation studies. Hum Mol Genet. 2018 Aug 1;27(R2):R195–208.
23. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple. Vol. 57, Source: Journal of the Royal Statistical Society. Series B (Methodological). 1995.
24. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomisation. JAMA. 2021 Oct 26;326(16):1614.
25. Skrivankova VW, Richmond RC, Woolf BAR, Davies NM, Swanson SA, VanderWeele TJ, et al. Strengthening the reporting of observational studies in epidemiology using

mendelian randomisation (STROBE-MR): explanation and elaboration. BMJ. 2021 Oct 26;n2233.

26. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014 May 15;10(5):e1004383.
27. Rasooly D, Peloso GM, Giambartolomei C. Bayesian Genetic Colocalisation Test of Two Traits Using *coloc*. Curr Protoc. 2022 Dec 14;2(12).
28. Robinson JW, Hemani G, Babaei MS, Huang Y, Baird DA, Tsai EA, et al. An efficient and robust tool for colocalisation: Pair-wise Conditional and Colocalisation (PWCoCo). bioRxiv [Internet]. 2022 Jan 1;2022.08.08.503158. Available from: http://biorxiv.org/content/early/2022/08/08/2022.08.08.503158.abstract
29. Chris Wallace. Prior explorer for coloc [Internet]. [cited 2024 Dec 12]. Available from: https://chr1swallace.shinyapps.io/coloc-priors/
30. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun 29;45(6):580–5.
31. Graham DB, Lefkovith A, Deelen P, de Klein N, Varma M, Boroughs A, et al. TMEM258 Is a Component of the Oligosaccharyltransferase Complex Controlling ER Stress and Intestinal Inflammation. Cell Rep. 2016 Dec;17(11):2955–65.
32. Li A, Song NJ, Riesenberg BP, Li Z. The Emerging Roles of Endoplasmic Reticulum Stress in Balancing Immunity and Tolerance in Health and Diseases: Mechanisms and Opportunities. Front Immunol. 2020 Feb 11;10.
33. Park HG, Park WJ, Kothapalli KSD, Brenna JT. The fatty acid desaturase 2 (*FADS2*) gene product catalyzes Δ4 desaturation to yield n-3 docosahexaenoic acid and n-6 docosapentaenoic acid in human cells. The FASEB Journal. 2015 Sep;29(9):3911–9.
34. Aglago EK, Huybrechts I, Murphy N, Casagrande C, Nicolas G, Pischon T, et al. Consumption of Fish and Long-chain n-3 Polyunsaturated Fatty Acids Is Associated With Reduced Risk of Colorectal Cancer in a Large European Cohort. Clinical Gastroenterology and Hepatology. 2020 Mar;18(3):654-666.e6.
35. Kim M, Park K. Dietary Fat Intake and Risk of Colorectal Cancer: A Systematic Review and Meta-Analysis of Prospective Studies. Nutrients. 2018 Dec 12;10(12):1963.
36. Wang J, Zhang Y, Zhao L. Omega-3 PUFA intake and the risk of digestive system cancers. Medicine. 2020 May;99(19):e20119.
37. Bull C, Hazelwood E, Bell JA, Tan V, Constantinescu AE, Borges C, et al. Identifying metabolic features of colorectal cancer liability using Mendelian randomisation. Elife. 2023 Dec 21;12.
38. Haycock PC, Borges MC, Burrows K, Lemaitre RN, Burgess S, Khankari NK, et al. The association between genetically elevated polyunsaturated fatty acids and risk of cancer. EBioMedicine. 2023 May;91:104510.
39. Phan LM, Yeung SCJ, Lee MH. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. Cancer Biol Med. 2014 Mar;11(1):1–19.
40. Dumitru C, Kabat AM, Maloy KJ. Metabolic Adaptations of CD4+ T Cells in Inflammatory Disease. Front Immunol. 2018 Mar 15;9.
41. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet. 2009 Jan 9;54(1):15–39.
42. Houtman M, Hesselberg E, Rönnblom L, Klareskog L, Malmström V, Padyukov L. Haplotype-Specific Expression Analysis of MHC Class II Genes in Healthy Individuals and Rheumatoid Arthritis Patients. Front Immunol. 2021 Aug 17;12.
43. Dunne MR, Phelan JJ, Michielsen AJ, Maguire AA, Dunne C, Martin P, et al. Characterising the prognostic potential of HLA-DR during colorectal cancer development. Cancer Immunology, Immunotherapy. 2020 Aug 18;69(8):1577–88.
44. Barnaba V. T Cell Memory in Infection, Cancer, and Autoimmunity. Front Immunol. 2022 Jan 3;12.

45. Labriet A, Lévesque É, Cecchin E, De Mattia E, Villeneuve L, Rouleau M, et al. Germline variability and tumor expression level of ribosomal protein gene RPL28 are associated with survival of metastatic colorectal cancer patients. Sci Rep. 2019 Sep 10;9(1):13008.
46. Coghill ID, Brown S, Cottle DL, McGrath MJ, Robinson PA, Nandurkar HH, et al. FHL3 Is an Actin-binding Protein That Regulates α-Actinin-mediated Actin Bundling. Journal of Biological Chemistry. 2003 Jun;278(26):24139–52.
47. Huang Z, Yu C, Yu L, Shu H, Zhu X. The Roles of FHL3 in Cancer. Front Oncol. 2022 May 24;12.
48. Cao G, Li P, He X, Jin M, Li M, Chen S, et al. FHL3 Contributes to EMT and Chemotherapy Resistance Through Up-Regulation of Slug and Activation of TGFβ/Smad-Independent Pathways in Gastric Cancer. Front Oncol. 2021 Jun 4;11.
49. Li P, Cao G, Zhang Y, Shi J, Cai K, Zhen L, et al. FHL3 promotes pancreatic cancer invasion and metastasis through preventing the ubiquitination degradation of EMT associated transcription factors. Aging. 2020 Jan 13;12(1):53–69.
50. Ye SB, Cheng YK, Li PS, Zhang L, Zhang LH, Huang Y, et al. High-throughput proteomics profiling-derived signature associated with chemotherapy response and survival for stage II/III colorectal cancer. NPJ Precis Oncol. 2023 May 31;7(1):50.
51. Yuan Y, Bao J, Chen Z, Villanueva AD, Wen W, Wang F, et al. Multi-omics analysis to identify susceptibility genes for colorectal cancer. Hum Mol Genet. 2021 Apr 27;30(5):321–30.
52. Andersson-Rolf A, Zilbauer M, Koo BK, Clevers H. Stem Cells in Repair of Gastrointestinal Epithelia. Physiology. 2017 Jul;32(4):278–89.
53. Blomhoff A, Olsson M, Johansson S, Akselsen HE, Pociot F, Nerup J, et al. Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype specific manner assessed mainly by DRB1*03 and DRB1*04 haplotypes. Genes Immun. 2006 Mar 1;7(2):130–40.
54. Davey Smith G, Hemani G. Mendelian randomisation: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014 Sep 15;23(R1):R89–98.
55. Butler-Laporte G, Lu T, Morris S, Zhang W, Band G, Hamilton F, et al. An accurate genetic colocalisation method for the HLA locus. 2024.
56. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2023 [cited 2025 Jan 17]. Available from: https://www.R-project.org/
57. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018 May 30;7.
58. Richardson N, Cook I, Crane N, Dunnington D, Françoise R, Keane J, et al. arrow: Integration to "Apache." https://CRAN.R-project.org/package=arrow; 2024.
59. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009 Aug 23;4(8):1184–91.
60. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439–40.
61. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T. data.table: Extension of `data.frame`. https://CRAN.R-project.org/package=data.table; 2024.
62. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr; 2023.
63. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013 Aug 8;9(8):e1003118.
64. Hemani G. gwasvcf: Tools for Dealing with GWAS Summary Data in VCF Format. https://github.com/mrcieu/gwasvcf; 2024.

65. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009 Jul 15;25(14):1841–2.
66. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations. https://CRAN.R-project.org/package=stringr; 2023.
67. Wickham H, Vaughan D, Girlich M. tidyr: Tidy Messy Data. https://CRAN.R-project.org/package=tidyr; 2024.
68. Carey V. gwascat: representing and modeling data in the EMBL-EBI GWAS catalog. 2024.
69. Slowikowski K. ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2.” https://CRAN.R-project.org/package=ggrepel; 2024.
70. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
71. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Dec 25;4(1):7.

Supporting information

Supplementary table 1. CD4+ T cell subtypes description.

Supplementary table 2. Results of the two-sample MR analysis on the causal role of CD4+ T cell expression profiles on CRC development.

Supplementary table 3. Genes in the two-sample MR results (supplementary table 2) which pass an FDR-*P*<0.05.

Supplementary table 4. Full genetic colocalisation results for CD4+ T expression profiles and CRC development.

Supplementary table 5. Number of SNPs lost during rsID matching.

Supplementary table 6. Number of SNPs lost during allele frequency matching.

Supplementary figure 1. Priors set for the genetic colocalisation analysis.

Supplementary note. STROBE-MR checklist.