SCIENCE CHINA PRESS

SCIENCE CHINA Life Sciences
life.scichina.com link.springer.com
https://doi.org/10.1007/s11427-023-2439-7

RESEARCH PAPER

Prioritization of risk genes in colorectal cancer by integrative analysis of multi-omics data and gene networks

Ming Zhang[1,2†], Xiaoyang Wang[1,3†], Nan Yang[1,2†], Xu Zhu[4†], Zequn Lu[1], Yimin Cai[1], Bin Li[1], Ying Zhu[1], Xiangpan Li[5], Yongchang Wei[6], Shaokai Zhang[3*], Jianbo Tian[1,2*] & Xiaoping Miao[1,2,7,8*]

[1]Department of Epidemiology and Biostatistics, School of Public Health; Department of Gastrointestinal Oncology, Zhongnan Hospital of Wuhan University; Department of Radiation Oncology, Renmin Hospital of Wuhan University, TaiKang Center for Life and Medical Sciences, Wuhan University, Wuhan 430071, China;
[2]Research Center of Public Health, Renmin hospital of Wuhan University, Wuhan University, Wuhan 430060, China;
[3]Department of Cancer Epidemiology, The Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital, Henan Engineering Research Center of Cancer Prevention and Control, Henan International Joint Laboratory of Cancer Prevention, Zhengzhou 450008, China;
[4]Department of Gastrointestinal Surgery, Renmin Hospital of Wuhan University, Wuhan 430060, China;
[5]Department of Radiation Oncology, Renmin Hospital of Wuhan University, Wuhan 430060, China;
[6]Department of Gastrointestinal Oncology, Hubei Cancer Clinical Study Center, Zhongnan Hospital of Wuhan University, Wuhan 430062, China;
[7]Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430073, China;
[8]Jiangsu Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China

†Contributed equally to this work
*Corresponding authors (Shaokai Zhang, email: shaokaizhang@126.com; Jianbo Tian, email: tianjb@whu.edu.cn; Xiaoping Miao, email: xpmiao@whu.edu.cn)

Received 31 July 2023; Accepted 26 August 2023; Published online 21 September 2023

Genome-wide association studies (GWASs) have identified over 140 colorectal cancer (CRC)-associated loci; however, target genes at the majority of loci and underlying molecular mechanisms are poorly understood. Here, we utilized a Bayesian approach, integrative risk gene selector (iRIGS), to prioritize risk genes at CRC GWAS loci by integrating multi-omics data. As a result, a total of 105 high-confidence risk genes (HRGs) were identified, which exhibited strong gene dependencies for CRC and enrichment in the biological processes implicated in CRC. Among the 105 HRGs, *CEBPB*, located at the 20q13.13 locus, acted as a transcription factor playing critical roles in cancer. Our subsequent assays indicated the tumor promoter function of *CEBPB* that facilitated CRC cell proliferation by regulating multiple oncogenic pathways such as MAPK, PI3K-Akt, and Ras signaling. Next, by integrating a fine-mapping analysis and three independent case-control studies in Chinese populations consisting of 8,039 cases and 12,775 controls, we elucidated that rs1810503, a putative functional variant regulating *CEBPB*, was associated with CRC risk (OR=0.90, 95%CI=0.86–0.93, $P=1.07\times10^{-7}$). The association between rs1810503 and CRC risk was further validated in three additional multi-ancestry populations consisting of 24,254 cases and 58,741 controls. Mechanistically, the rs1810503 A to T allele change weakened the enhancer activity in an allele-specific manner to decrease *CEBPB* expression via long-range promoter-enhancer interactions, mediated by the transcription factor, REST, and thus decreased CRC risk. In summary, our study provides a genetic resource and a generalizable strategy for CRC etiology investigation, and highlights the biological implications of *CEBPB* in CRC tumorigenesis, shedding new light on the etiology of CRC.

susceptibility genes | gene screening models | multi-omics | GWAS | *CEBPB* | long-range promoter-enhancer interactions

INTRODUCTION

Colorectal cancer (CRC), the second major cause of cancer deaths, imposes a major health burden worldwide. In China, CRC is the third most common cancer diagnosed in adults and the fifth leading cause of death from cancer (Chen et al., 2016; Ju et al., 2023). Given the 12%–35% heritability of CRC (Jiao et al., 2014), human genetic approaches can promote understanding of biological mechanisms and contribute to the development of effective clinical treatments. To date, more than 140 common genetic variants have been found to be associated with CRC risk through genome-wide association studies (GWASs), providing genetic basis for dissecting CRC etiology and biology (Li et al., 2020). However, over 90% of GWAS risk variants are intergenic and cannot be directly mapped to specific genes (Consortium, 2012), impeding interpretation of biological implications of GWAS findings. One default approach to identifying target genes is to assign the most significant single-nucleotide polymorphism (SNP) to the nearest gene at each locus. This seems to ignore the fact that most SNPs affect the activity of regulatory elements (e.g., enhancers, silencers) (Maurano et al., 2012), which can influence gene expression over long genomic ranges. Indeed, studies based on expression quantitative trait loci (eQTL) analysis indicate that two thirds of the causal genes at GWAS loci are not the nearest genes (Brænne et al., 2015; Kapoor et al., 2021; Zhu et al., 2016).

Connecting risk loci with their likely casual genes is a tremendous process that requires the integration of GWAS data with multi-omics features to provide comprehensive supporting evidence. Transcriptomics data generated form large-scale projects such as The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx) project help to depict gene

expression profiles and promote further functional exploration. Epigenomics data derived from Functional Annotation of the Mammalian Genome 5 (FANTOM5) (Lizio et al., 2015) and the 3D-genome Interaction Viewer and database (3DIV) (Yang et al., 2018) provide significant chromatin contacts, thus facilitating the investigation of regulatory networks. Recent studies have integrated GWAS with multi-omics datasets to identify causal variants and risk genes implicated in complex human diseases including schizophrenia, Alzheimer's disease, and depression (Jin et al., 2021; Wang et al., 2019a). While similar efforts have been targeted at CRC, they have predominantly focused on transcriptome data and performed eQTL analysis or transcriptome-wide association studies (TWASs) to functionally interpret CRC GWAS findings (Yin et al., 2022; Yuan et al., 2021). However, target genes for CRC risk loci remain largely unclear, suggesting that increasingly diverse omics data beyond transcriptomics are required to be integrated to identify genuine risk genes for CRC.

To take advantage of the complementary information contained in multi-omics datasets, we applied integrative risk gene selector (iRIGS) (Wang et al., 2019a), a Bayesian framework that can represent and integrate different layers of multi-omics data, to probabilistically infer prioritized risk genes at CRC GWAS loci. We predicted a set of high-confidence risk genes (HRGs), most of which are not the nearest genes to the GWAS index SNPs. Subsequent functional analyses revealed the biological implications of these HRGs in CRC. It was observed that *CEBPB* was identified as the HRG at the 20q13.13 locus. Previous research reported that the expression of CEBPB significantly increased in CRC tumors compared with normal colon mucosa (Rask et al., 2000). Tang et al. (2021) found that CEBPB could activate *UBQLN4* and thus promote CRC cell proliferation, migration and invasion. Therefore, we investigated the role of *CEBPB* in CRC and verified its tumor promoter function. The molecular mechanisms of the functional variant rs1810503 regulating *CEBPB* were also examined. Collectively, our study showcases the power of integrated multi-omics analysis to provide fresh insights into functional interpretation of GWAS findings and advance the understanding of CRC etiology and potential therapeutics.

RESULTS

Integrating multi-omics data to identify HRGs for CRC

We performed iRIGS on 148 CRC-associated loci (Figure 1A; Table S1 in Supporting Information) (Chang et al., 2018; Cui et al., 2011; Lu et al., 2019; Peters et al., 2013; Schmit et al., 2014; Tanikawa et al., 2018; Zeng et al., 2016) by integrating genomic features, including differential expression, mutation frequency, distal regulatory element (DRE)-promoter links, and distance to GWAS index SNP, to prioritize risk genes for CRC. A total of 105 HRGs with the highest posterior probability (PP) for each locus were identified (Figure 1B). We also defined 1,041 local background genes (LBGs) with a PP less than the median PP of all the candidates. As expected, different genomic features consistently exhibited supportive evidence for HRGs (Figure S1 in Supporting Information).

HRGs exhibit strong gene dependencies for CRC cells

To assess the essentiality of HRGs for cancer cell fitness, we queried their gene dependencies in CRC cells using genome-wide loss-of-function screening data. Both CRISPR-based and RNAi-based dependency scores were significantly lower for HRGs compared with LBGs (Figure 2A and B). When comparing HRGs with the nearest genes to the corresponding index SNPs, we observed modestly decreased gene dependency scores, although the differences were not statistically significant (Figure 2A and B). These results suggest that HRGs exhibit relatively stronger gene dependencies for CRC cells compared with other candidates at GWAS loci, and moreover, iRIGS is capable of prioritizing genes that are essential for the survival of CRC cells.

HRGs are enriched in the biological processes implicated in CRC

To investigate the biological functions of HRGs in CRC, we first analyzed their enrichment in MSigDB Hallmark subsets, using all genes across the loci as references. HRGs were significantly enriched in 19 Hallmark subsets, including E2F targets, G2/M checkpoint, TGF-β signaling, Myc targets, and others (Figure 2C). The nearest genes were significantly enriched in 11 Hallmark subsets, including TGF-β signaling, Wnt/β-catenin signaling, TNFα signaling via NF-κB, and others (Figure 2C). Intriguingly, the *P* values of LBGs were statistically significant in six subsets, but the odds ratio (OR) values were all less than 1.000 (Figure 2C). Gene Ontology (GO) and pathway analyses from three databases (KEGG, Wiki, and Reactome) revealed similar trends for functional enrichment (Figure S2 in Supporting Information). Next, we compiled a CRC-related gene set containing 2,065 reported genes derived from the literature to explore the involvement of HRGs, the nearest genes, and LBGs in this gene set. As shown in Figure 2D, compared with all genes at the 148 loci, both HRGs and the nearest genes had a significantly higher proportion of CRC-related genes (41.9% *vs.* 11.7%, and 33.1% *vs.* 11.7%, respectively), whereas LBGs had a lower proportion (6.8% *vs.* 11.7%).

To expand the understanding of HRGs contributing to cancer development, we next investigated whether HRGs preferentially harbored somatic copy-number alternation (SCNA) and mutation burdens. We performed GISTIC analysis using SCNA data from TCGA CRC tumor tissues and identified 11 significantly altered regions with amplifications or deletions (*G*-score>0.1), containing members of HRGs (Figure 2E). The corresponding results for the nearest genes and LBGs are shown in Figure S3A and B in Supporting Information. Furthermore, we integrated HRGs, the nearest genes, and LBGs that ranked in the top 20 among somatic mutated genes for further investigation. We observed that HRGs presented frequent SCNAs and mutations in 58.22% of CRC cases (Figure 2F), while the proportions were 54.04% for the nearest genes (Figure S3C in Supporting Information) and 31.64% for LBGs (Figure S3D in Supporting Information), respectively. These results suggest that SCNAs and mutations in HRGs may act as potential predicting markers and provide new insights into therapeutic targets. Considering that immune response can influence tumor development through immunosurveillance and proinflammatory factors, we assessed the associations between gene expression and the number of infiltrating immune cells. The majority of HRGs were closely associated with high infiltration of immune cells, especially macrophages and $CD4^+$ T cells (Figure 2G), whereas the nearest genes and LBGs were closely associated with high infiltration of T

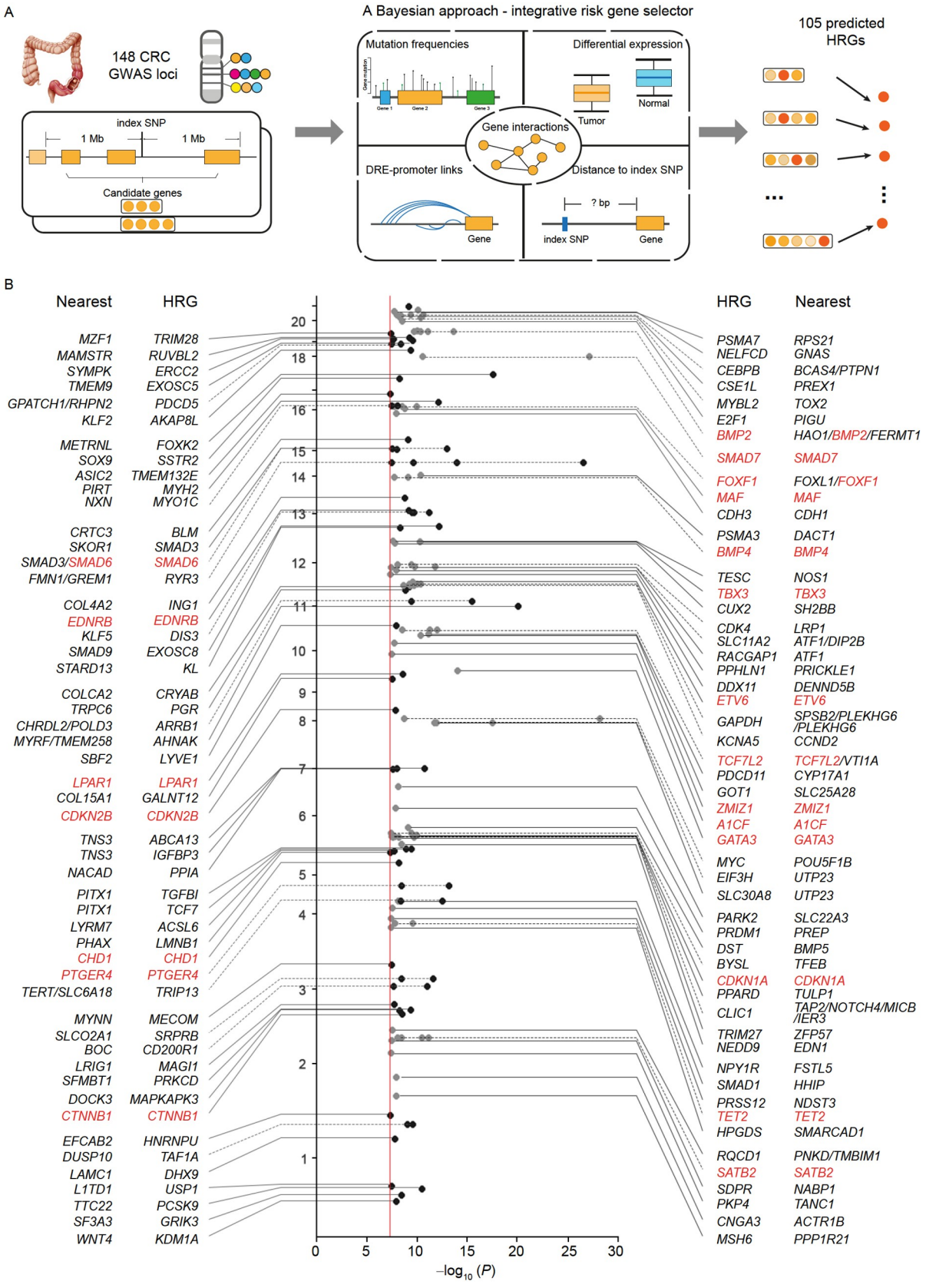

Figure 1. Identification of high-confidence risk genes for CRC via iRIGS. A, A schematic illustration of the iRIGS framework. B, A total of 105 HRGs were identified for CRC. The genes are sorted by chromosomes along the vertical axis with the location of chromosomes indicated. The horizontal axis shows the $-\log_{10}$(*P*-values) for index SNPs in the previous GWAS. The red line indicates the genome-wide significance *P* value threshold of 5×10^{-8}. The solid line represents a single HRG for a GWAS loci; the dotted line indicates that the HRGs are identical across multiple loci. nearest, the nearest genes to the GWAS index SNPs.

https://doi.org/10.1007/s11427-023-2439-7

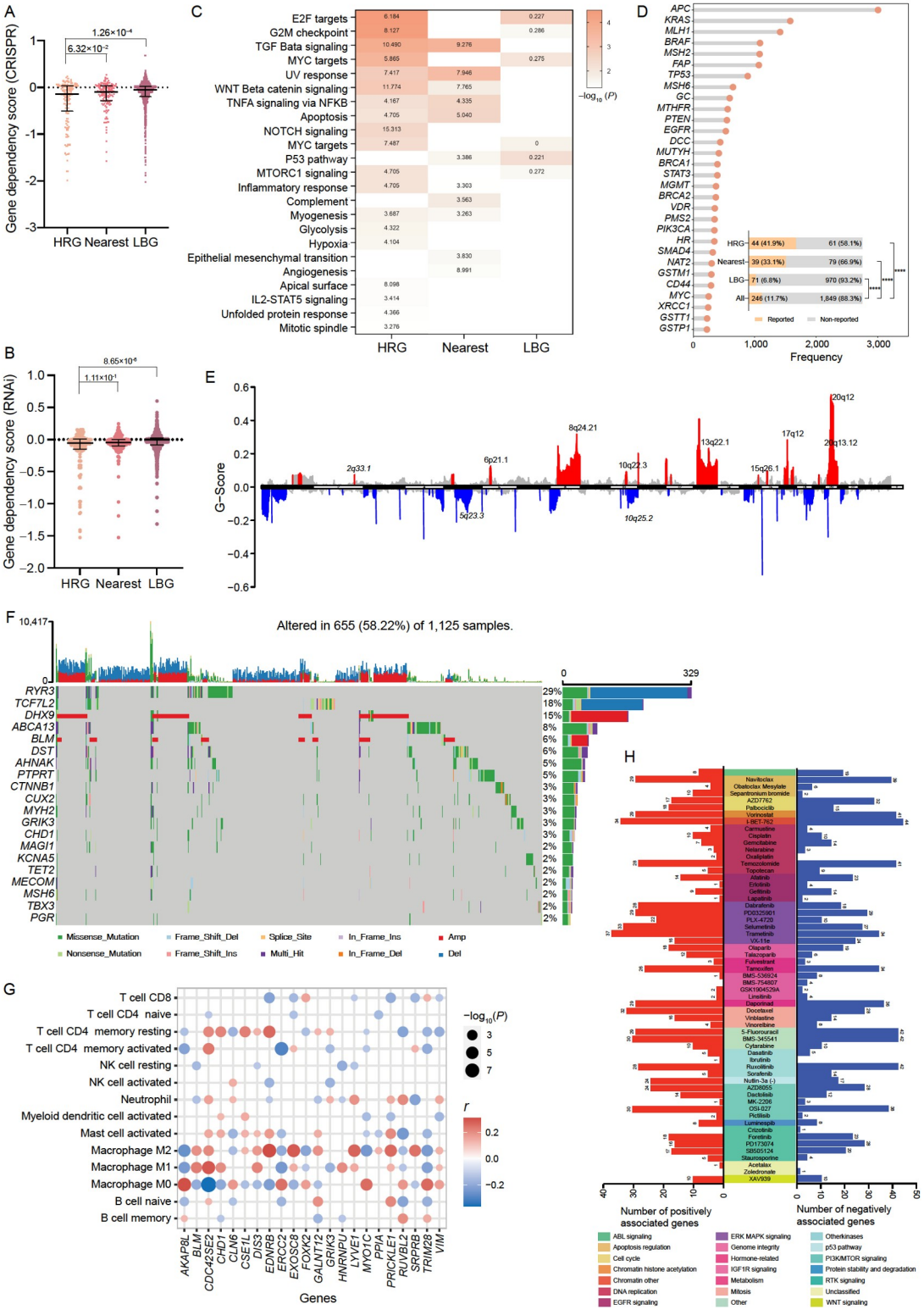

Figure 2. Functional characterization of predicted risk genes. A and B, HRGs exhibit strong gene dependencies for CRC cell lines. Both CRISPR-based (A) and RNAi-based (B) dependency scores are significantly lower for HRGs compared with LBGs. C, Enrichment analysis for HRGs, the nearest genes and LBGs in MSigDB Hallmark subsets. D, HRGs and the nearest genes have a significantly higher proportion of previously reported CRC genes, as compared with all genes at the 148 GWAS loci. E, GISTIC analysis shows that HRGs present frequent SCNAs (G-score>0.1) in TCGA CRC tumor tissues. The somatic copy-number alterations regions were indicated in the Figure. Red bar denotes amplification and blue bar denotes deletion. F, Somatic mutation and SCNA landscape of CRC tumor samples. The type of the top 20 altered HRGs is shown for every sample, and mutation subtypes as well as SCNA events are denoted by color. G, Bubble diagram depicting the correlations between the expression of HRGs and individual immune cell types in CRC tumor samples. Values displayed are the Spearman correlations of immune cell fractions (rows) with HRG expression (columns). Red indicates positive correlations (increasing proportions of indicated cell types with increasing gene expression), and blue indicates negative correlations, respectively. Size of the bubble indicates significance of the correlation. H, The identified HRG-drug pairs based on GDSC dataset. Blue bar denotes the negative association and red bar denotes the positive association, respectively.

cells (Figure S3E and F in Supporting Information). The proportions of immune-related genes for HRGs, LBGs and the nearest genes are shown in Figure S4A–C in Supporting Information. Moreover, to understand the impact of HRGs on drug responses in cancer, we identified a total of 1,791 HRG-drug pairs based on the Genomics of Drug Sensitivity in Cancer (GDSC) drug dataset and further examined the pharmaceutical targets of these drugs (Figure 2H). Noticeably, the top drug category was drugs targeting chromatin other pathways (Figure 2H). For instance, I-BET-762, an inhibitor of BET, was revealed to be positively associated with 34 HRGs and negatively associated with 44 HRGs, respectively (Figure 2H). The drugs targeting signaling pathways linked to the nearest genes and LBGs are shown in Figure S3G and H in Supporting Information. Furthermore, we selected the top 10 drugs with the most associated genes and displayed associations between drug sensitivity and gene expression from distinct gene sets (HRGs, the nearest genes, and LBGs; Figure S4D–F in Supporting Information). Multiple HRGs such as *AHNAK*, *BLM*, *MYO1C*, and *TGFBI* exhibited a phenomenon that one gene was closely associated with several drugs (Figure S4D in Supporting Information), highlighting the clinical potential of HRGs in screening drug targets.

iRIGS prioritizes *CEBPB* to be the HRG at the 20q13.13 locus

Genes that are novel or non-canonical were nominated by iRIGS in the study, especially those located distally to the index SNPs. To further validate the predictions from iRIGS, we carried out an in-depth investigation on the rs1810502 locus at chromosome 20. At this locus, *PTPN1* is the nearest gene to the index SNP, while the top predicted gene is *CEBPB* (Table S2 in Supporting Information). We then assessed the biological function of *CEBPB* in CRC tumorigenesis. The mRNA levels of *CEBPB* were significantly higher in CRC tumor tissues compared with adjacent normal tissues in our 123 paired CRC samples (Figure 3A). Similar results were obtained from multiple databases including GEO (Figure 3B; Figure S5A in Supporting Information), TCGA (Figure 3C; Figure S5B in Supporting Information) and Oncomine (Figure 3D). The CRISPR-Cas9 screening data showed that *CEBPB* was essential for the survival of the CRC SW1463 cells (Figure 3E). We next examined the effect of *CEBPB* on CRC cell phenotypes. It was found that the overexpression of *CEBPB* in HCT116 and SW480 cells substantially increased CRC cell proliferation (Figure 3F; Figures S5C and S6A in Supporting Information), whereas the knockdown of *CEBPB* substantially reduced cell proliferation (Figure 3G; Figures S5D and S6B in Supporting Information). The colony formation ability of CRC cells was markedly stimulated by *CEBPB* overexpression (Figure 3H; Figures S5E and S6A in Supporting Information) but substantially attenuated by *CEBPB* knockdown (Figure 3I; Figures S5F and S6B in Supporting Information).

CEBPB exerts oncogenic effects on CRC by activating genes involved in MAPK, PI3K-Akt, and Ras signaling pathways

The protein product of this gene, CEBPB, is a specific CCAAT/enhancer-binding protein beta, functioning as a transcription factor (TF) playing regulatory roles in cancer. Thus, we used chromatin immunoprecipitation sequencing (ChIP-seq) data to explore its biological mechanism affecting CRC cell proliferation. We obtained 4,892 ChIP-seq binding peaks for CEBPB in HCT116 cells, and subsequently mapped these peaks to 1,880 genes. These CEBPB-binding genes presented high SCNA and mutation burdens, and were correlated with immune infiltration as well as drug response, which could be helpful in clinical applications (Figure S7 in Supporting Information). KEGG pathway enrichment analysis showed that CEBPB-binding genes were significantly enriched in signaling pathways regulating cancer cell proliferation such as MAPK, PI3K-Akt, Ras, and Rap1 signaling pathways (Figure 3J). We subsequently investigated the coexpression of *CEBPB* with the genes in the top 20 pathways in TCGA CRC samples, followed by the verification using an independent GEO dataset (GSE9348). A total of 33 genes were correlated with *CEBPB* expression in both two datasets, and exhibited coexpression relationships with each other (Figure 3K; Figures S8 and S9 in Supporting Information). Intriguingly, these genes were involved in MAPK, PI3K-Akt, and Ras signaling pathways, providing critical clues on the network of downstream targets regulated by CEBPB. We then validated the coexpression relationships between *CEBPB* and genes in these three pathways, using real-time qPCR in CRC cells with either *CEBPB* overexpression or knockdown. Ultimately, a total of five genes including *BCL2L1*, *EREG*, *MYC*, *PRKCA*, and *TAOK3* were significantly correlated with *CEBPB* expression (Figure S10 in Supporting Information). Moreover, previous studies have reported that CEBPB promotes CRC proliferation and metastasis by activating the transcription of a panel of genes (*UBQLN4*, *TRIM2*, *G-CSF*, *IL-6*, and others), thereby regulating downstream pathways including Wnt/β-catenin/c-Myc axis, p53 signaling, and myeloid-derived suppressor cell related immunosuppressive molecules (Fultang et al., 2020; Groth et al., 2019; Tang et al., 2021; Yang et al., 2017; Zhou et al., 2022; Zou et al., 2014) (Figure 3L). Altogether, these findings reveal that *CEBPB* pathologically activates multiple oncogenic pathways, thereby contributing to the proliferation and development of CRC.

The SNP rs1810503 is identified as a putative functional variant at the *CEBPB* locus

CEBPB is located in the 20q13.13 region; rs1810502 is the index SNP at the locus identified in European populations. However, causal variants in this region have not been systematically investigated. Since all SNPs in LD with rs1810502 ($r^2 \geqslant 0.2$) were far away from *CEBPB*, we speculated that the functional SNP modified the activity of an enhancer, regulating *CEBPB* expression through long-range interactions. Therefore, we first screened out 37 SNPs in this block that physically interacted with *CEBPB* in CRC cell lines (HT-29 and LoVo), a sigmoid colon tissue (SG1) or FANTOM dataset. Among them, 26 SNPs were significant eQTLs affecting *CEBPB* expression in colon tissues from GTEx database. We then performed functional annotation for the 26 SNPs by using multiple bioinformatics tools, including RegulomeDB, CADD and GWAVA. In all three scoring systems, rs1810503 ranked in the top three (Figure S11 in Supporting Information). Additionally, rs1810503 was marked by DNase hypersensitive peaks and active histone modifications (H3K27ac and H3K4me1) according to Cistrome DB's annotation, suggesting the enhancer activity of the fragment (Figure 4A). Finally, we validated the correlation between rs1810503 and *CEBPB* mRNA expression in 123 CRC

https://doi.org/10.1007/s11427-023-2439-7

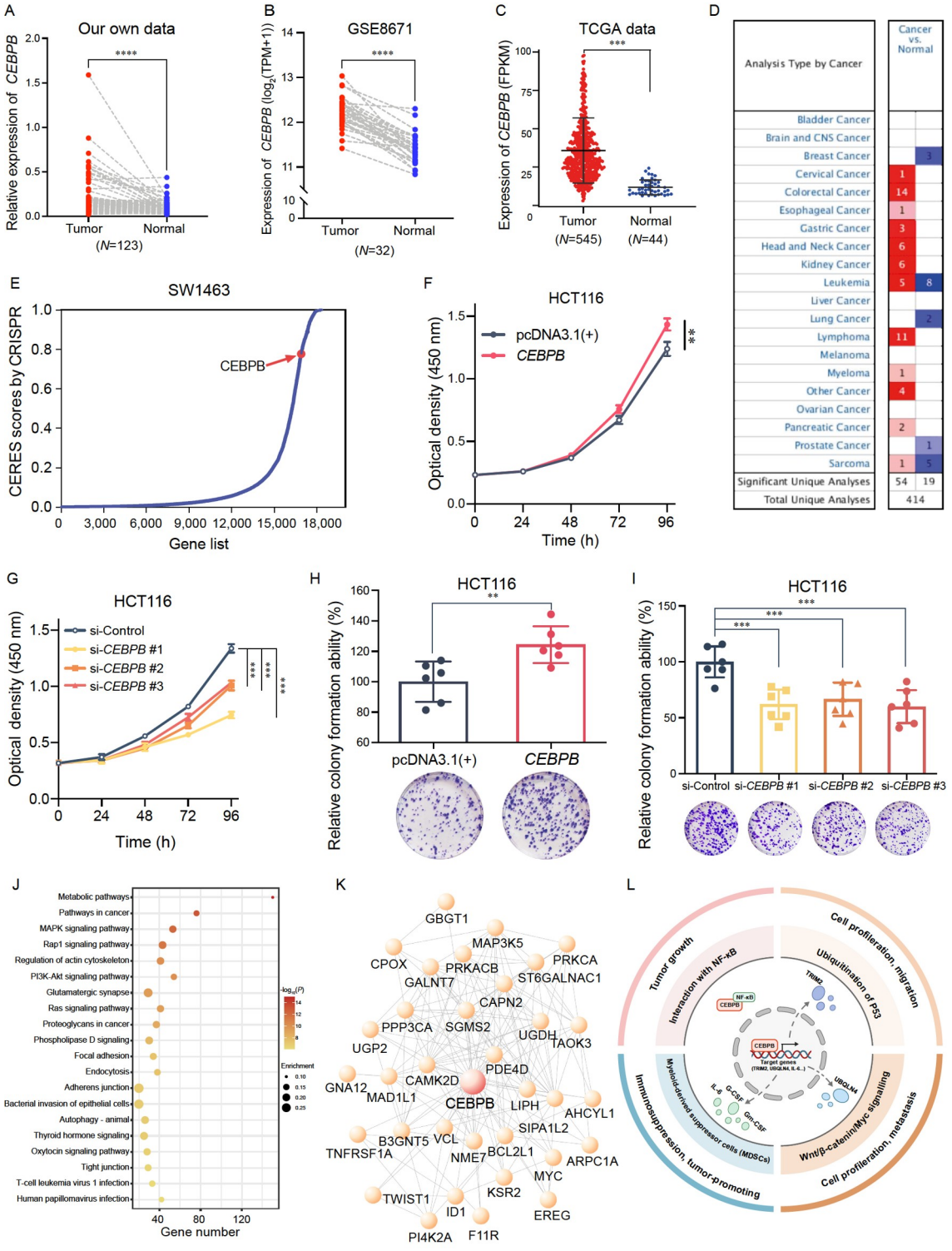

Figure 3. *CEBPB* located at 20q13.13 acts as an oncogene in CRC by promoting cell proliferation via MAPK, PI3K-Akt, and Ras pathways. A–C, *CEBPB* is significantly overexpressed in tumor tissues compared with that in normal tissues from our CRC patients (A), GEO (B) and TCGA (C) datasets. ***, $P<0.001$; ****, $P<0.0001$ (two-sided Student's *t*-test). D, *CEBPB* expression levels in multiple types of tumor tissues from the Oncomine database. E, *CEBPB* is essential for cell growth with high CERES scores in SW1463 cells from the genome-wide CRISPR/Cas9-based loss-of-function screening data. F and G, The effect of *CEBPB* overexpression (F) or knockdown (G) on CRC cell proliferation in HCT116 cells. Results are shown as mean±SEM from three experiments, each with three replicates. **, $P<0.01$; ***, $P<0.001$ (two-sided Student's *t*-test). H and I, The effect of *CEBPB* overexpression (H) or knockdown (I) on colony formation ability in HCT116 cells. The results present colony formation ability relative to control cells (set to 100%). Data are presented as the mean±SD from three independent experiments, each with two replicates. **, $P<0.01$; ***, $P<0.001$ (two-sided Student's *t*-test). J, The top 20 significant pathways from KEGG enrichment analysis for CEBPB-binding genes, as performed by KOBAS v3.0. The majority of these genes are implicated in oncogenic signaling including MAPK, PI3K-Akt, and Ras signaling pathways, among others. K, The coexpression network of *CEBPB* and CEBPB-binding genes in the top 20 KEGG pathways from TCGA and GEO (GSE9348) datasets, constructed by Cytoscape. L, The mechanisms of CEBPB contributing to CRC tumorigenesis. CEBPB interacts with NF-κB, or activates the transcription of multiple genes such as *TRIM2*, *UBQLN4*, *IL-6*, and thus regulates a panel of downstream oncogenic pathways to promote tumor growth and metastasis.

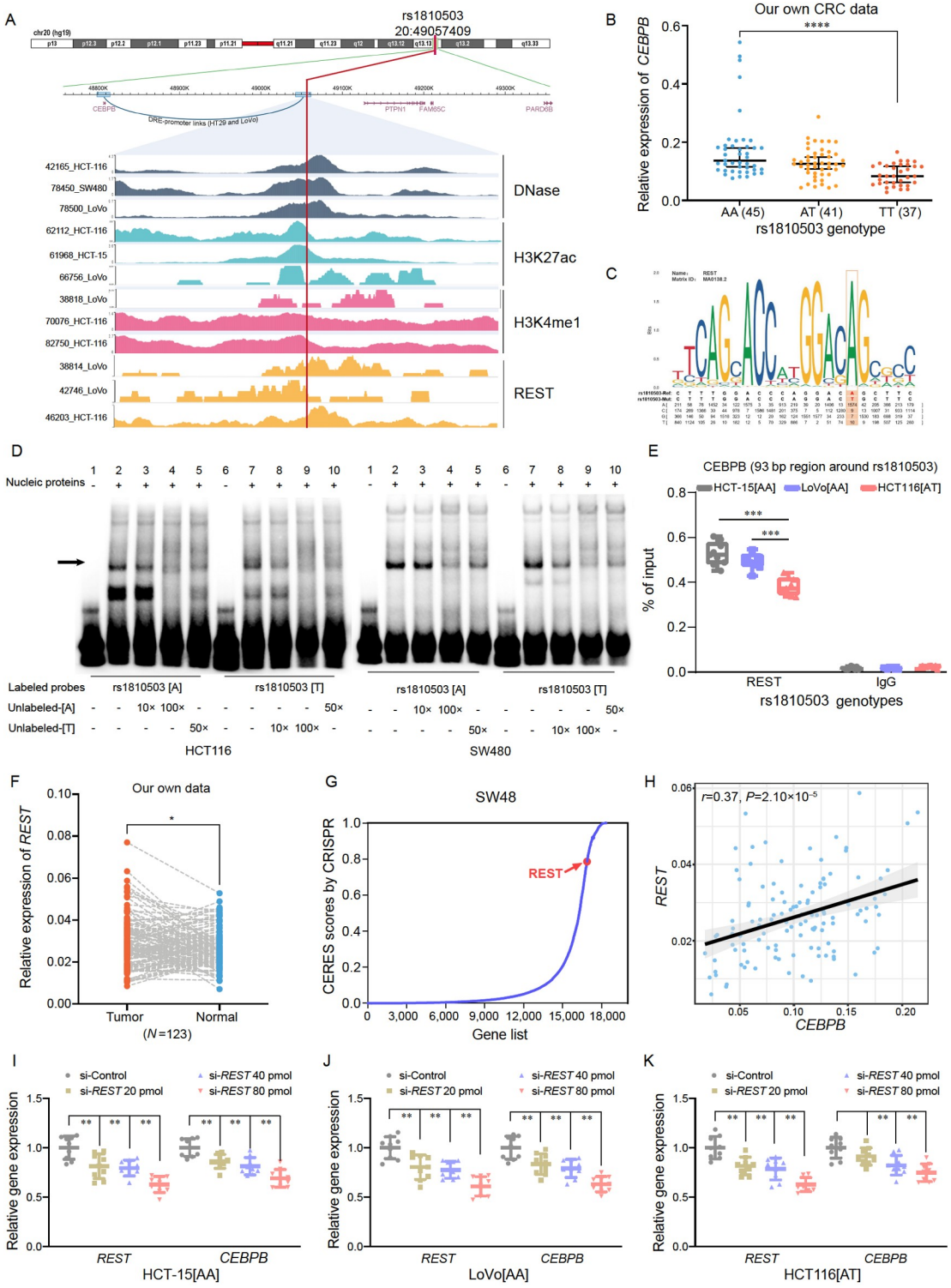

Figure 4. The allele-specific affinity of REST with the rs1810503 sequences. A, Epigenetic annotation for the region surrounding rs1810503 in CRC cell lines. Data including DNase peaks, transcription factor (REST) peaks, and multiple histones (H3K4me1 and H3K27ac) modification peaks were obtained from Cistrome DB. B, eQTL analysis of *CEBPB* expression with the rs1810503 genotypes in 123 CRC tumor tissues. C, The variant rs1810503 resides within the REST-binding motif according to JASPAR. D, EMSAs with biotin-labeled probes containing rs1810503 in HCT116 and SW480 cells. Arrows indicate allele-specific bands of probes with nuclear proteins in the cells. Additionally, 10×, 50×, and 100× represent a 10-fold, 50-fold, and 100-fold excess of the unlabeled probe compared with the labeled probe, respectively. "+", added; "-", not added. E, The allele-specific binding of REST to the region surrounding rs1810503 as measured by ChIP-qPCR assays in HCT-15[AA], LoVo[AA], and HCT116[AT] cell lines. Data are shown from three repeated experiments, each with three replicates. ***, $P<0.001$ (two-sided Student's *t*-test). F, *REST* is significantly overexpressed in tumor tissues compared with adjacent normal tissues from our 123 CRC patients. Data are shown as mean±SD. *, $P<0.05$ (two-sided paired Student's *t*-test). G, *REST* is essential for CRC cell growth. *REST* shows a high CERES score in SW48 cells from the genome-wide CRISPR-Cas9 screening data. H, The expression of *CEBPB* is positively correlated with *REST* expression in our 123 CRC tumor tissues. The *P* value and *r* value were calculated by Pearson's correlation analysis. I–K, The knockdown of *REST* attenuates *CEBPB* expression in a dose-dependent manner in HCT-15 (I) and LoVo cells (J), but not in HCT116 cells transfected with a low-dose siRNA (20 pmol, K). **, $P<0.01$ (two-sided Student's *t*-test).

https://doi.org/10.1007/s11427-023-2439-7

tumor tissues, and the result showed that carriers with the T allele of rs1810503 had lower *CEBPB* expression than those with the A allele (Figure 4B). The similar effect of the SNP on *CEBPB* expression in colon tissues from GTEx data is displayed in Figure S12 in Supporting Information.

The rs1810503 A to T change is strongly associated with a decreased risk of CRC

To further verify the association between rs1810503 and CRC risk, we conducted three independent case-control studies in Chinese populations consisting of 8,039 cases and 12,775 controls. Descriptive characteristics of the study subjects are detailed in Table S3 in Supporting Information. As expected, the significant association of rs1810503 with susceptibility to CRC was observed under all the models (dominant, recessive, and additive) in all the three studies after adjusting for gender, age, smoking status and drinking status (Table 1). Moreover, we performed a combined analysis and observed that the rs1810503 [T] allele was strongly associated with a decreased risk of CRC with an odds ratio (OR) of 0.90 (95% confidence interval (CI) =0.86–0.93, $P=1.07\times10^{-7}$) under the additive model. Similar results were also found under the dominant model (OR=0.86, 95%CI=0.81–0.92, $P=1.37\times10^{-5}$) and the recessive model (OR=0.87, 95%CI=0.81×0.93, $P=1.73\times10^{-5}$). We further validated the effect of rs1810503 on CRC risk in multi-ancestry populations (mostly Europeans) derived from Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), UK Biobank and Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), consisting of a total of 24,254 CRC cases and 58,741 controls. Demographic characteristics of the subjects are shown in Table S4 in Supporting Information. The regional plot around the rs1810503 locus generated by LocusZoom based on GWAS data from UK Biobank is shown in Figure S13 in Supporting Information. Consistently, statistically significant associations were observed under the additive, dominant and recessive models in all the three populations. The subsequent combined analysis showed that the T allele of rs1810503 was strongly associated with a decreased risk of CRC, having an OR of 0.83 (95%CI=0.79–0.87, $P=7.58\times10^{-14}$). The results under the

Table 1. The association between rs1810503 and colorectal cancer risk in Chinese and multi-ancestry populations[a)]

Study stage	Genotypes	Chinese populations (*N*=20,814)			Database	Genotypes	Multi-ancestry populations (*N*=82,995)		
		Cases/Control	OR (95%CI)	*P*			Cases/Control	OR (95%CI)	*P*
Study I (*N*=1,857)					PLCO (*N*=8,533)				
	AA	214/159	1.00 (Ref)			AA	444/2,236	1.00 (Ref)	
	AT	490/505	0.69 (0.54–0.88)	3.16×10^{-3}		AT	594/3,707	0.81 (0.71–0.93)	2.59×10^{-3}
	TT	219/270	0.80 (0.70–0.92)	2.09×10^{-3}		TT	181/1,371	0.79 (0.73–0.88)	3.55×10^{-6}
	Dominant		0.67 (0.53–0.85)	8.82×10^{-4}		Dominant		0.78 (0.68–0.88)	1.14×10^{-4}
	Recessive		0.76 (0.61–0.94)	1.04×10^{-2}		Recessive		0.75 (0.63–0.89)	1.00×10^{-3}
	Additive		0.77 (0.67–0.89)	2.72×10^{-4}		Additive		0.82 (0.75–0.89)	9.33×10^{-6}
Study II (*N*=7,488)					UK Biobank *N*=36,722)				
	AA	588/859	1.00 (Ref)			AA	2,121/11,575	1.00 (Ref)	
	AT	1,483/2,420	0.87 (0.78–0.99)	4.06×10^{-2}		AT	2,406/15,044	0.88 (0.82–0.94)	9.16×10^{-5}
	TT	752/1,386	0.89 (0.83–0.96)	1.56×10^{-3}		TT	719/4,857	0.81 (0.73–0.88)	5.99×10^{-6}
	Dominant		0.85 (0.76–0.96)	7.94×10^{-3}		Dominant		0.86 (0.81–0.91)	1.65×10^{-6}
	Recessive		0.84 (0.76–0.94)	2.05×10^{-3}		Recessive		0.86 (0.79–0.94)	1.05×10^{-3}
	Additive		0.88 (0.82–0.95)	3.76×10^{-4}		Additive		0.89 (0.85–0.93)	4.15×10^{-7}
Study III (*N*=11,469)					GECCO (*N*=37,740)				
	AA	1,192/1,757	1.00 (Ref)			AA	10,601/11,530	1.00 (Ref)	
	AT	2,140/3,582	0.91 (0.83–0.99)	4.01×10^{-2}		AT	6,111/7,047	0.96 (0.92–1.01)	8.37×10^{-2}
	TT	961/1,837	0.84 (0.76–0.94)	1.68×10^{-3}		TT	1,077/1,374	0.94 (0.90–0.98)	3.82×10^{-2}
	Dominant		0.88 (0.81–0.97)	5.80×10^{-3}		Dominant		0.95 (0.91–0.99)	1.44×10^{-2}
	Recessive		0.89 (0.82–0.98)	1.30×10^{-2}		Recessive		0.90 (0.82–0.97)	9.93×10^{-3}
	Additive		0.88 (0.83–0.93)	2.70×10^{-6}		Additive		0.95 (0.92–0.98)	2.80×10^{-3}
Combined study (*N*=20,814)					Combined study (*N*=82,995)				
	AA	1,994/2,775	1.00 (Ref)			AA	13,166/25,341	1.00 (Ref)	
	AT	4,113/6,507	0.90 (0.84–0.96)	2.11×10^{-3}		AT	9,111/25,798	0.87 (0.82–0.92)	3.17×10^{-6}
	TT	1,932/3,493	0.81 (0.74–0.87)	1.48×10^{-7}		TT	1,977/7,602	0.79 (0.74–0.84)	4.07×10^{-13}
	Dominant		0.86 (0.81–0.92)	1.37×10^{-5}		Dominant		0.90 (0.86–0.94)	8.56×10^{-7}
	Recessive		0.87 (0.81–0.93)	1.73×10^{-5}		Recessive		0.84 (0.78–0.91)	9.31×10^{-6}
	Additive		0.90 (0.86–0.93)	1.07×10^{-7}		Additive		0.83 (0.79–0.87)	7.58×10^{-14}

a) All *P* values were calculated by unconditional logistic regression models after adjusting for gender, age group, smoking status and drinking status, except for the GECCO datasets adjusted for gender and age group only.

dominant model (OR=0.90, 95%CI=0.86–0.94, $P=8.56\times10^{-7}$) and the recessive model (OR=0.84, 95%CI=0.78–0.91, $P=9.31\times10^{-6}$) also exhibited similar trends. These findings on functional annotation and population associations further support the putative causal role of rs1810503 in CRC.

The T allele of rs1810503 attenuates the binding affinity with the TF REST

We next explored the mechanisms of rs1810503 contributing to CRC risk by affecting *CEBPB* mRNA expression. Given that SNP-specific changes are thought to modify enhancer activity by altering TF binding (Leung et al., 2015), we examined whether rs1810503 directly altered the TF-binding motif by using HumanTFDB and JASPAR. According to the annotations, rs1810503 might bind to the TF REST in an allele-specific manner (Figure 4C). This observation was further supported by ChIP-seq data for REST in CRC cell lines (Figure 4A).

Thus, we carried out electrophoretic mobility shift assays (EMSAs) with nuclear proteins extracted from HCT116 and SW480 cells. As shown in Figure 4D, DNA sequences containing the rs1810503[T] allele exhibited a weaker affinity with nuclear proteins than the rs1810503[A] allele (lane 7 *vs.* lane 2). The binding signal of labeled probes containing the A allele was gradually attenuated by the A allele-containing unlabeled probes in a dose-dependent manner (lanes 2–4); however, such changes were not obvious for the T allele (lanes 7–9). In addition, we validated this observation using ChIP-qPCR assays in three CRC cell lines with different genotypes (HCT-15[AA], LoVo[AA], HCT116[AT]). In line with the motif analysis, we found a weaker REST binding in the rs1810503 region in HCT116 cells compared with HCT-15 and LoVo cells (Figure 4E), suggesting that REST was less prone to bind to the rs1810503[T] allele.

It is reported that *REST* plays a major role in tumorigenesis and metastasis in multiple types of cancer (Labrecque et al., 2019; Negrini et al., 2013). We then investigated the function of *REST* in CRC biology. As a result, in our 123 CRC-affected patients, *REST* was significantly overexpressed in tumor tissues compared with adjacent normal tissues (Figure 4F). The data from the genome-wide CRISPR-Cas9 screening in CRC SW48 cells revealed that *REST* was essential for cell viability (Figure 4G). To further elaborate the regulatory control of REST on *CEBPB*, we analyzed their expression pattern in CRC, and observed the positive correlation between the expression of *REST* and *CEBPB* in both our 123 CRC tumor tissues (Figure 4H) and the GTEx colon samples (Figure S14 in Supporting Information). When *REST* was knocked down in the aforementioned three CRC cell lines with different genotypes, *CEBPB* expression was decreased concomitantly in HCT-15 and LoVo cells, but not in HCT116 cells when delivered small interfering RNA (siRNA) with a low dose (Figure 4I–K), further suggesting that the regulatory effect of REST on *CEBPB* expression occurred in an allele-specific manner.

The T allele of rs1810503 weakens a promoter-enhancer interaction mediated by REST to downregulate *CEBPB* expression

To determine the regulatory mechanisms of rs1810503 affecting *CEBPB* expression, we performed luciferase reporter assays and found that the construct containing the rs1810503[T] allele exhibited lower enhancer activity than that containing the rs1810503[A] allele in both HCT116 and SW480 cell lines (Figure 5A and B). Additionally, the differences in luciferase activity between the two alleles of rs1810503 were dose-dependently attenuated when *REST* was knocked down at an increasing dose (Figure 5C and D), suggesting that the allele-specific effect of rs1810503 on *CEBPB* transcriptional activity was modulated by REST.

The SNP rs1810503 is located about 250 kb away from the *CEBPB* transcriptional start site (TSS). To evaluate whether there is a direct chromatin interaction between the region of rs1810503 and *CEBPB* promoter, we carried out allele-specific chromosome conformation capture (3C) assays in three CRC cell lines with different genotypes of rs1810503 (HCT116[AT], HCT-15[AA] and LoVo[AA]). As shown in Figure 5E, *CEBPB* promoter showed a stronger interaction with the region containing rs1810503 than any of the other neighboring NcoI sites tested. Notably, the interaction frequency was significantly lower in the cell line with the rs1810503[AT] genotype than in cell lines with the rs1810503[AA] genotype, suggesting that rs1810503 could mediate allele-specific long-range chromatin loops with *CEBPB* promoter. We further validated the long-range interaction using Hi-C assay in one CRC tumor tissue, and found that the region containing rs1810503 significantly interacted with *CEBPB* promoter (Figure 5F). The result was replicated using Hi-C data of the HCT116 cell line from Hi-C data Browser (Figure 5G). Altogether, we demonstrate that the T allele of rs1810503 attenuates the enhancer activity to decrease the expression of *CEBPB* by REST-mediated long-range enhancer-promoter interactions.

DISCUSSION

GWASs have identified thousands of genetic variants that are associated with diseases and traits of medical importance in humans. However, the genes or functional DNA elements through which the genetic variants exert their effects on diseases and traits remain largely unknown. Although several methods such as eQTL analysis and TWAS (Cao et al., 2020; Codrich et al., 2021; Vangala et al., 2021) have exhibited remarkable advantages, identifying target genes of risk SNPs remains a big challenge due to the complex patterns of regulatory programs. To bridge this gap, we used a multi-dimensional integration approach, iRIGS, to pinpoint risk genes from a massive pool of candidates across the CRC loci. As a result, we provided multiple lines of strong evidence to establish a gene-centric view of the genetic architecture of CRC. Moreover, we identified a key HRG, *CEBPB*, in the 20q13.13 region, and unraveled the biological implications of *CEBPB* and its regulatory SNP in CRC tumorigenesis, further deepening the understanding of the etiology and pathogenesis of CRC.

Traditionally, the nearest genes to lead SNPs were recognized as risk genes at CRC GWAS loci, but distinct studies have identified multiple lead SNPs and sometimes reported different candidates within the same locus. For example, at the 15q13.3 locus, the nearest genes identified across three studies were *GREM1*, *CRAC1*, *BMP4*, *BMP2* (Fortini et al., 2021; Jaeger et al., 2008; Tomlinson et al., 2011). Moreover, considering that a large proportion of phenotype-associated SNPs identified by GWAS lied within noncoding regions of the genome, the strategy of choosing the closest gene to each index SNP might not explain

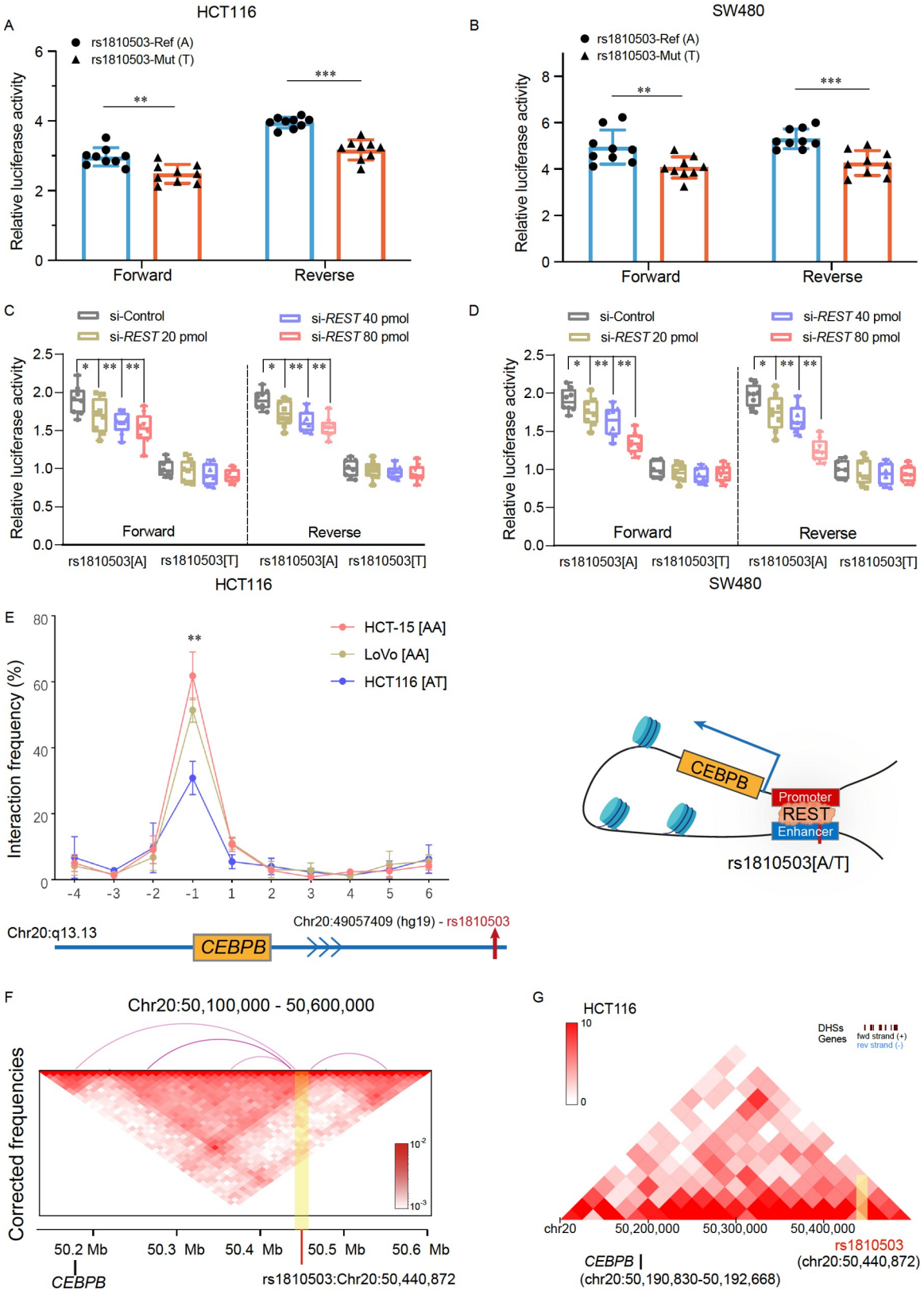

Figure 5. The mutated allele of rs1810503 attenuates a promoter-enhancer interaction mediated by REST to downregulate *CEBPB* expression. A and B, Relative reporter gene activity of the constructs containing the rs1810503[A] or rs1810503[T] allele in CRC HCT116 (A) and SW480 (B) cells. Data are shown as mean±SD. C and D, Effect of *REST* knockdown on the relative luciferase activity of the constructs containing the rs1810503[A] or rs1810503[T] allele in HCT116 (C) and SW480 (D) cells. E, The 3C profiles in multiple CRC cell lines show the relative interaction frequencies between the *CEBPB* promoter (the anchor) and representative NcoI sites, including a fragment containing rs1810503. Data are shown as mean±SEM. **, $P<0.01$ (two-sided Student's *t*-test). F and G, Hi-C plots reveal the interaction of the region containing rs1810503 with *CEBPB* promoter in our CRC tumor tissue (F) and the HCT116 cell line from Hi-C data Browser (G).

the complicated regulatory mechanisms well (Smemo et al., 2014; Wang et al., 2007). Noticeably, accumulated evidence has demonstrated that long-range interactions between regulatory elements and gene promoters play key roles in transcriptional regulation (Sanyal et al., 2012). Thus, there is a need to apply multi-omics approaches rather than physical proximity alone into the functional interpretation of GWAS findings.

The iRIGS framework is designed to take advantage of the high-dimensional omics data, and the more relevant genomic features are included, the more accurate the prediction is (Wang et al., 2019a). In this study, through integrating genomic features including differential expression, mutation frequency, DRE-promoter links, and distance to index SNP, we identified a total of 105 HRGs with the highest posterior probability for each CRC GWAS locus. We observed that HRGs captured more genomic features that were characteristics of CRC risk genes compared with LBGs and the nearest genes. Therefore, iRIGS was proved to provide a powerful way to probabilistically rank candidate genes at each GWAS locus. Clinically, identifying risk genes at the associated loci is essential for preventing oncogenesis and designing therapeutic interventions that can halt tumor growth (Hormozdiari et al., 2016). Given this, we found that HRGs were relevant to cancer signaling pathways, high mutation burdens, immune infiltration, and pharmaceutical targets, which were quite important and beneficial for clinical applications. These findings highlight the promise of the identified risk genes for drug repositioning for CRC. Although the results might be not a direct proof that these genes were genuine cancer genes, it was encouraging to see that they corresponded to genes that, when altered, significantly affected CRC cell growth. In fact, both CRISPR-based and RNAi-based dependency scores suggested that HRGs were essential for the growth of CRC cells. These findings strongly suggest that integrating multi-omics data is able to propose candidate risk genes which exhibit high confidence and show promise to be further experimentally validated.

Importantly, we identified *CEBPB* as the HRG for CRC susceptibility at the 20q.13.13 locus. At this locus, rs1810502 was reported to be an index SNP in previous CRC GWAS conducted among European populations, and the C>T mutation was associated with a decreased risk of CRC (Schumacher et al., 2015). A previous study speculated that rs1810502 might contribute to CRC risk by affecting the nearest gene, *PTPN1* (Miki et al., 1994), the protein product of which is a regulator of endoplasmic reticulum unfolding proteins. The distance between the TSS of *PTPN1* and rs1810502 was approximately 69 kb. However, we found no evidence of an eQTL relationship between rs1810502 and *PTPN1* in GTEx colon tissues, or a significant interaction in any remote regulatory element-gene interaction datasets. Based on these findings, it appears that available evidence is insufficient to support the role of rs1810502 in regulating *PTPN1* expression in CRC. Noticeably, we pinpointed a new high-confidence risk gene, *CEBPB*, at the 20q13.13 locus, with high-dimensional genomic features boosting the inference accuracy. *CEBPB*, one of the CEBP family members, is a crucial transcription regulator of gene expression during innate immunity (Wang et al., 2019b), inflammatory responses and cancer development (Cheng et al., 2019; Liu et al., 2022). Through bioinformatics analysis and a series of experiments at the functional level, we found that *CEBPB* contributed to CRC cell proliferation likely by activating a panel of genes implicated in the MAPK, PI3K-Akt, and Ras pathways. These findings were consistent with previous studies on the critical importance of CEBPB as a key TF in tumorigenesis by regulating a multitude of oncogenic pathways. Zhou et al. (2022) reported that CEBPB bound to *TRIM2* promoter and activated its transcription, thereby decreasing the stability of p53 via promoting its ubiquitination in CRC. Tang et al. (2021) found that CEBPB could activate *UBQLN4*, which subsequently exerted oncogenic effects on CRC through the Wnt/β-catenin signaling pathway. Intriguingly, numerous researches elaborated that CEBPB was an essential "master" regulator of the biology of myeloid-derived suppressor cells (MDSCs), which were considered to contribute to the immunosuppressive tumor microenvironment and to be an obstacle to cancer immunotherapies (Fultang et al., 2020; Groth et al., 2019). All these findings indicate the significant role of CEBPB in the regulation of cancer and further support the promise of our HRG gene set in the understanding of CRC tumorigenesis. Besides, at the regulatory level, we performed experiments and illuminated the mechanisms of the variant, rs1810503, modifying CRC risk by regulating the expression of *CEBPB*. Briefly, the mutated allele of rs1810503, which was located in an enhancer of *CEBPB*, weakened a promoter-enhancer interaction mediated by REST to downregulate *CEBPB* expression, thus decreasing CRC risk (Figure 6).

There are still some limitations in this study. First, since transcriptional regulatory programs are commonly cell-type-specific, single-cell technology will be useful for dissecting variant-to-gene connections. In addition, despite multidimensional data supporting the prediction accuracy, the risk SNP-gene map still lacks experimental evidence to directly test these connections. Thus, the CRISPR-Cas9 system can be used for further validation.

In summary, the HRGs identified in our study greatly expand potential candidate targets of risk SNPs for CRC, which can be further verified by future studies to help elucidate the genetic etiology of CRC. Moreover, the critical roles of one particular HRG, *CEBPB*, and its regulatory variant in colorectal carcinogenesis, validated by functional experiments, further shed light on the pathogenesis of CRC. Together, we anticipate that this gene-centric map of genetic etiology of CRC is valuable for the refinement of GWAS association signals and can provide risk gene sets for future applications in precision oncology and beyond.

MATERIAL AND METHODS

CRC-associated loci

We included SNPs identified through GWAS and exome-wide association studies that showed an association with CRC at a significance level of $P<5.0\times10^{-8}$, as reported in the literature (Al-Tassan et al., 2015; Chang et al., 2018; Cui et al., 2011; Dunlop et al., 2012; Hofer et al., 2017; Houlston et al., 2010; Houlston et al., 2008; Huyghe et al., 2019; Jia et al., 2013; Jiang et al., 2015; Law et al., 2019; Lu et al., 2019; Orlando et al., 2016; Real et al., 2014; Schmit et al., 2019; Schmit et al., 2014; Schumacher et al., 2015; Tanikawa et al., 2018; Tenesa et al., 2008; Tomlinson et al., 2008; Tomlinson et al., 2011; Wang et al., 2016; Wang et al., 2017; Whiffin et al., 2014; Zeng et al., 2016; Zhang et al., 2014a; Zhang et al., 2014b). We excluded SNPs involved in SNP-SNP interactions or SNP-environment

https://doi.org/10.1007/s11427-023-2439-7

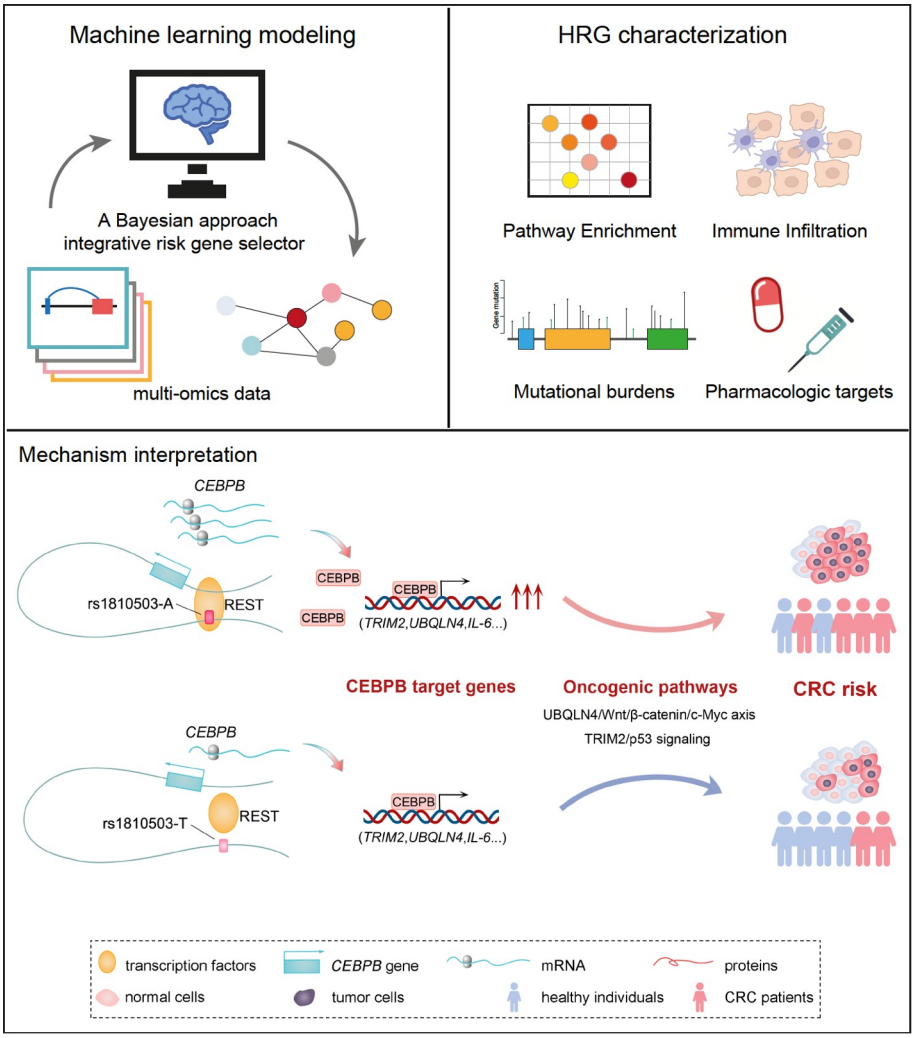

Figure 6. Graphical representation of the regulation and function of *CEBPB* in CRC. Compared with the rs1810503[A], the mutated rs1810503[T] allele attenuates the binding of the TF REST to the promoter of *CEBPB*, and weakens a promoter-enhancer interaction that reduces *CEBPB* expression to regulate a panel of downstream oncogenic pathways, thus decreasing CRC risk.

interactions. Autosomal index SNPs were further filtered by linkage disequilibrium (LD, $r^2 \geqslant 0.2$), and the ones showing the most significant associations were selected. This resulted in a total of 148 independent index SNPs for subsequent analysis (Table S1 in Supporting Information). For each locus, candidate genes within the 2 Mb region centered on the index SNP were collected for further analysis.

Multi-omics integrative analysis based on iRIGS

We employed the powerful Bayesian algorithm, iRIGS (Wang et al., 2019a), to prioritize CRC-associated genes from the 148 risk loci, considering multi-omics supporting evidence and gene-gene networks. The gene-gene network was constructed based on the relationships inferred from the GO annotations. The connections between genes were determined by their shared annotations and functional relationships within the GO hierarchy, serving as a representation of the correlations among genes. It was utilized in iRIGS as a source of prior information for inferring risk genes from GWAS data. Moreover, multiple layers of genomic features were included in iRIGS in the study, including differential expression, mutation frequency, DRE-promoter links, and distance to index SNP. Gene expression data of 545 CRC individuals were downloaded from the TCGA database, quantified by RNA sequencing and represented by FPKM values. Differential expression between tumor tissues and adjacent normal tissues was determined using Wilcoxon rank-sum tests to calculate *P* values for each gene. Gene mutation data, containing mutation frequency information for 18,362 genes in 534 CRC patients from the TCGA database, were downloaded from the cBioPortal (Cerami et al., 2012). DRE-promoter links were collected from three sources: (i) Orlando et al. (2018) inferred chromatin contacts in cell lines by promoter capture Hi-C. The predicted DRE-promoter links were downloaded, from which we obtained 118,758 and 96,458 significant interactions in HT-29 and LoVo cell lines, respectively. (ii) Capture Hi-C data of human sigmoid colon tissue "SG1" were downloaded from the 3DIV database (Yang et al., 2018). The "P-O" interaction type was selected and 26,446 significant links ($P<0.05$) were obtained. (iii) The FANTOM5 project used the cap analysis of gene expression technology to infer enhancer-promoter interactions across multiple human tissues (Lizio et al., 2015). We

downloaded the "tss-enhancer associations" dataset from FANTOM5 and obtained 66,942 enhancer-promoter interactions. By integrating these multi-omics data into the iRIGS framework, we calculated the PP for each candidate gene, indicating its likelihood of being a risk gene. The candidate with the highest PP value for each locus was defined as an HRG. Genes with PP values lower than the median PP of all candidates were defined as LBGs.

Gene dependency analysis. CRISPR dependency data were derived from the 19Q2 Avana dataset, which contained gene dependencies estimated for each gene and cell line by the CERES algorithm (Meyers et al., 2017). RNAi dependency data were taken from Project DRIVE, which were reprocessed using the DEMETER2 algorithm (McFarland et al., 2018). Gene dependency information of 73 CRC cell lines in the CRISPR dataset and 44 in the RNAi dataset was subsequently obtained. For each dataset, the dependency scores for each gene were calculated by averaging the dependency scores in all CRC cell lines.

Functional characterization of iRIGS-identified genes

Gene sets from two sources were collected for gene set enrichment analysis. Firstly, the "H: hallmark gene sets" were downloaded from MSigDB. Secondly, a CRC-related gene set was generated from published literature in PubMed. Specifically, terms including "CRC", "colorectal cancer", "colorectal carcinoma", "colorectal tumor", "colon cancer", "colon carcinoma", "colon tumor", "rectal cancer" "rectal carcinoma", "rectal tumor", "risk", "tumorigenesis", "oncogenesis" and "gene" were searched, and the publication period was set from "1990/01/01" to "2022/12/31". A total of 21,672 articles were retrieved and the abstract texts were exported. By using pubmed.mineR, an R package that extracts gene names from texts and counts their frequency of occurrence (Rani et al., 2015), we obtained 2,065 genes with word frequency not less than five. Besides, two online tools, DAVID (Huang et al., 2009) and KOBAS (Xie et al., 2011), were also employed to perform gene set enrichment analysis.

To evaluate genomic variation of target genes, we downloaded masked copy number segment files from the TCGA data portal, in which probes containing germline mutations were removed, and determined significant focal copy number alterations using GISTIC 2.0. Masked somatic mutation calls identified by the MuTect2 pipeline were downloaded from TCGA, which detected not only somatic single-nucleotide variations but also small insertions and deletions. Visualization of the SCNA and mutation landscape was conducted by R package maftools.

Anticancer immune response is administered by tumor-infiltrating immune cells, and thus the quantification of various types of immune infiltrates can help to understand the mechanisms underlying immune regulation. We applied CIBERSORT (Newman et al., 2015) to estimate the immune cell infiltration levels for each CRC patient from gene expression profiles. The associations between immune infiltrates and target gene expression were also evaluated by the partial correlation coefficient with the inclusion of tumor purity as a covariable. We defined genes with the absolute value of correlation coefficient$\geqslant$0.15 and false discovery rate (FDR)<0.05 as immune infiltrate-related genes. The proportions of immune-related genes are the number of genes correlated with immune infiltrates divided by the total number of total target genes.

We next analyzed the associations between drug response and the expression of target genes. Expression profile and drug sensitivity data of human cancer cell lines were obtained from the GDSC (released in June 2020), containing the sensitivity data for 198 compounds over 809 cell lines. The dataset provides the drug response result (IC_{50} values) as a measure of drug sensitivity, and lower IC_{50} values indicate increased sensitivity to treatment. OncoPredict was used to impute drug response for TCGA cancer patients based on cancer molecular datasets from GDSC. Then, we calculated Spearman's correlation between target gene expression and predicted drug response in CRC cells, and defined FDR<0.05 and the absolute value of Spearman's correlation$\geqslant$0.15 as significant.

Prediction and characterization of CEBPB-binding genes

The ChIP-seq data for CEBPB in a CRC cell line, HCT116, was downloaded from the ENCODE database (FactorBook: ENCSR000BSD) (Davis et al., 2018). We took the intersection of the signal peaks from two repeated experiments, and then aligned the peaks to human genes (GRCh37.p13). To investigate the biological function of these CEBPB-binding genes, we performed KEGG pathway analysis via KOBAS, and subsequently constructed the coexpression network for CEBPB and genes in the top 20 pathways using the datasets from GEO (GSE9348) and TCGA. The genomic variation, immune infiltration and drug sensitivity of CEBPB-binding genes were evaluated as previously described.

SNP selection

We narrowed down candidate SNPs according to the following process: (i) including SNPs in LD with rs1810502 in Asians ($r^2 \geqslant 0.2$), (ii) excluding SNPs without DRE-promoter links with *CEBPB* in any dataset (HT-29, LoVo, SG1 and FANTOM5), (iii) excluding SNPs not associated with *CEBPB* expression ($P \geqslant 0.05$; "Colon_Sigmoid" from GTEx Analysis v7 (Maurano et al., 2012)), and (iv) ranking the SNPs based on the functional annotation scores from RegulomeDB (Kapoor et al., 2021), CADD (Zhu et al., 2016) and GWAVA (Chen and Schunkert, 2021).

Study subjects

Chinese populations. The association between rs1810503 and CRC risk was detected using the TaqMan genotyping platform in three independent case-control studies in Chinese populations. In study I, subjects were recruited from Cancer Hospital Chinese Academy of Medical Sciences in Beijing, China, including 923 CRC cases and 934 controls. A total of 2,823 CRC cases and 4,665 controls in study II were recruited from Tongji Hospital of Huazhong University of Science and Technology in Wuhan, China. In study III, 4,293 CRC cases and 7,176 controls were recruited from multiple centers including Cancer Hospital Chinese Academy of Medical Sciences (Beijing, China), Tongji Hospital of Huazhong University of Science and Technology (Wuhan, China), Zhongnan Hospital of Wuhan University (Wuhan, China), and Renmin Hospital of Wuhan University (Wuhan, China). The details are shown in Table S3 and Figure S15 in Supporting Information. All cases had histopathologically or cytologically confirmed CRC by at least two local pathologists, and blood was collected without chemotherapy or radiotherapy.

All controls were cancer-free, recruited from a community nutritional survey in the same region with the cases. For all subjects, peripheral blood samples were collected, and demographic characteristics inclusive of age, gender, smoking and drinking status, were obtained. Specifically, subjects who had smoked more than one cigarette per day for more than one year were defined as smokers. Subjects who had drunk more than twice a day for more than one year were defined as drinkers. Informed consent was obtained from all subjects. This study was approved by the ethics committee of each hospital.

Multi-ancestry populations. The effect of rs1810503 on CRC risk was further validated in multi-ancestry populations, mostly Europeans. The details are shown in Table S4 and Figure S15 in Supporting Information. In study I, genotype data for 1,219 CRC cases and 7,314 controls were obtained from PLCO (dbGaP accession code: phs000346.v2.p2, phs001554.v1.p1, phs001286.v2.p2 and phs001524.v1.p1). Details of the study have been described previously (Buys et al., 2011; Prorok et al., 2000). In brief, 1,219 cases had confirmed primary invasive colorectal cancer diagnosis during the trial and subjects with other cancer were excluded according to self-reported and questionnaire data. Controls were frequency matched to cases with a 6:1 ratio without replacement. Matching criteria were age at enrollment (two-year blocks) and gender. In study II, whole-genome sequencing data for 5,246 cases and 31,476 controls were obtained from UK Biobank. The study design, recruitment, cohort profile, and data collection have been described in detail on the website. The CRC cases were defined as a malignant neoplasm of all cancers based on the 10th Revision of International Classification of Diseases (ICD10, C18, C19, C20 (except C18.1)). The study included all patients diagnosed with colorectal cancer as a first primary malignancy. In study III, the association was tested in 17,789 CRC cases and 19,951 controls, data for whom were obtained from GECCO (dbGaP accession code: phs001315.v1.p1, phs001415.v1.p1 and phs001078. v1. p1).

SNP genotyping

Genomic DNA was extracted from whole blood samples using the RelaxGene Blood DNA System (Tiangen, Beijing, China). The SNP, rs1810503, was genotyped using the TaqMan genotyping platform (QuantStudio7, Applied Biosystems, USA). Several measures were implemented for quality control, including (i) mixing of case and control samples on the plates, (ii) blinding of the experimenters who performed the genotyping assays to the case/control status of samples, and (iii) inclusion of both positive and negative control samples on each 384-well plate.

Cell lines and cell culture

The human CRC cell lines HCT116, SW480, HCT-15 and LoVo were purchased from China Center for Type Culture Collection (Wuhan, China). Cells were maintained in 1× dulbecco's modified eagle's medium (DMEM; Gibco, USA) supplemented with 10% fetal bovine serum (FBS; PAN-Biotech, Germany) and 1% Penicillin-Streptomycin (Pen-Strep; Gibco) in a humidified atmosphere of 37°C and 5% CO_2. Cell lines were regularly tested for mycoplasma (MycoAlert; Lonza, USA), and authenticated using the AmpFLSTR Identifiler PCR Amplification Kit (Applied Biosystems, USA).

RNA interference

Three sets of siRNAs against *CEBPB* and *REST* (Table S5 in Supporting Information) and a scrambled siRNA used as a negative control were purchased from RiboBio (Guangzhou, China). Transfections with siRNAs were performed using Lipofectamine RNAiMAX (Invitrogen, USA) at a final concentration of 100 nmol L^{-1}.

Overexpression of *CEBPB*

The coding sequence of *CEBPB* (NCBI CCDS ID: CCDS13429.1) was synthesized by Genewiz (Suzhou, China), and then cloned into the pcDNA3.1(+) vector between the restriction sites NheI and KpnI. Transfections with plasmids were performed using Lipofectamine 3000 Reagent (Invitrogen) at a final concentration of 1 μg mL^{-1}.

Real-time reverse transcriptase polymerase chain reaction (RT-qPCR)

Total RNA of cells was isolated using Trizol Reagent (Invitrogen). Reverse transcription was performed using HiScript Reverse Transcriptase (Vazyme, Nanjing, China). Real-time PCR was performed with ChamQ SYBR qPCR Master Mix (Vazyme). Expression levels of *CEBPB* were quantified using the $2^{-\Delta\Delta Ct}$ method with *GAPDH* employed as an internal control. Primers used in qPCR can be found in Table S5 in Supporting Information.

Cell proliferation assays

HCT116 and SW480 cells were seeded in 96-well flat-bottomed plates, with each well containing 2,000 cells in 100 μL of cell suspension. After a certain time of incubation, cell proliferation was detected using the CCK-8 reagent (Dojindo, Japan), expressed as the optical density at 450 nm. Each experiment with three replicates was repeated three times.

Colony formation assays

Cells were seeded into 6-well cell culture plates at a density of 2,000 cells per well. After 10 days, the cells were washed with cold PBS twice and fixed with 100% methanol. Finally, cells were stained with crystal violet. The colony number in each well was counted.

Dual luciferase reporter assay

DNA fragments surrounding rs1810503[A] or rs1810503[T] were synthesized (Genewiz) and cloned into the pGL3-Promoter vector between the restriction sites BamHI and SalI. HCT116 and SW480 cells were seeded at 5×10^4 cells per well in 96-well plates, and 200 ng of reporter plasmid was cotransfected into cells with 2 ng of pRL-SV40 plasmid using Lipofectamine 3000 (Invitrogen). Cells were collected 36 h after transfection, and luciferase activity was measured using Dual-Luciferase Reporter Assay System (Promega, USA). Renilla luciferase activity was used to normalize firefly luciferase activity. Data were collected from triplicate wells and each experiment was repeated three times.

EMSAs

Nuclear proteins were extracted from HCT116 and SW480 cells using Nuclear and Cytoplasmic Protein Extraction Kit (Beyotime, Shanghai, China). Double-stranded oligonucleotides containing rs1810503[A] or rs1810503[T] were synthesized by Genewiz (Table S5 in Supporting Information). EMSA was performed according to the protocol accompanying Chemiluminescent EMSA Kit (Beyotime). Biotin-labeled oligonucleotides and nuclear proteins were incubated for 30 min. Competitive unlabeled oligonucleotides were added to the reaction mixture 25 min before the addition of labeled probes. SuperSignal West Femto Trial Kit (Thermo Fisher Scientific, USA) was used for image development.

ChIP qPCR

ChIP assays were performed with a ChIP assay kit (Millipore, USA) according to the manufacturer's instructions. Cells were crosslinked with 1% formaldehyde, and glycine was added to stop fixation. Genomic DNA was extracted from the fixed cells and sheared by sonication. Antibodies against REST (Proteintech, USA) and a nonspecific rabbit IgG (Santa Cruz, USA) were subsequently incubated with the cross-linked proteins and DNA overnight for immunoprecipitation with protein A/G magnetic beads. DNA fragments were purified and collected by a Dr. GenTLE Precipitation Carrier kit (TaKaRa, Japan). The purified DNA library was sequenced (BerryGenomics, Beijing, China) or analyzed by qPCR. The primers used for ChIP-qPCR are shown in Table S5 in Supporting Information.

3C assays

3C assays were performed according to the method described by Hagège et al. (2007), in three CRC cell lines (HCT116, HCT-15 and LoVo), which carried different genotypes of rs1810503. Cells were fixed with formaldehyde and lysed in lysis buffer. The cells were then digested with NcoI enzyme at 37°C overnight. After ligation, the cross-linked DNA fragments were extracted using phenol/chloroform and precipitated with ethanol. A bacterial artificial chromosome (BAC) clone that covered the genome segments of the target regions was applied to eliminate amplification efficiency differences among different primers. In addition, *GAPDH* was used to normalize cell background differences. The physical interactions between anchors and primers were measured by qPCR. All 3C-qPCR primers (Table S5 in Supporting Information) were synthesized by TSINGKE Biological Technology (Wuhan, China).

Hi-C and Hi-C data processing

The Hi-C libraries include crosslinking, chromatin digestion with four-cutter restriction enzyme MboI and marking of DNA ends, ligation and purification, shearing, and biotin pull down. A Hi-C map is a matrix of DNA-DNA contacts produced by the Hi-C experiment. The valid pairs after pooling were binned into 200 kb (100, 40, 20, 10, 5 kb) nonoverlapping genomic intervals to generate contact maps. Raw Hi-C contact maps can contain many different biases, such as map-ability, GC content and uneven distribution of restriction enzyme sites. The corresponding cumulative probability *P*-values and FDR *q*-values were calculated in Ay's Fit-Hi-C software for contacts between 5 kb bins for intrachromosomal interactions, and the interactions with *q*-values less than 0.1 were identified as significant interactions. The colorectum tumor sample was obtained from Zhongnan Hospital of Wuhan University in Wuhan, China.

Statistical analysis

For demographic characteristics, categorical variables were described as frequency (%), and differences between groups were compared by χ^2 tests. Continuous variables were described as mean±standard deviation (SD) or mean±standard error of mean (SEM), and differences between groups were compared by Student's *t*-tests. The association between the SNP and CRC risk was analyzed by logistic regression. The differences in differential expression, gene mutation, DRE-promoter links, and gene dependency were compared by Wilcoxon rank-sum tests. Gene set enrichment analysis was performed by χ^2 tests or Fisher's exact tests. In addition, two-sided *t*-tests were used to compare the differences between the control groups and the experimental groups in functional experiments. Statistical analyses were performed with R language (v3.5.3) and $P<0.05$ indicated that the results were statistically significant.

Web resources

TCGA data portal, http://portal.gdc.cancer.gov/; cBioPortal, http://cbioportal.org; 3DIV, http://www.3div.kr/; FANTOM5 "tss-enhancer associations" dataset, http://enhancer.binf.ku.dk/presets/; Depmap, https://depmap.org/portal/; MSigDB, https://www.gsea-msigdb.org/gsea/msigdb; DAVID, http://david.ncifcrf.gov/; KOBAS, http://kobas.cbi.pku.edu.cn/; RegulomeDB, https://www.regulomedb.org/regulome-search/; CADD, http://cadd.gs.washington.edu/score; GWAVA, http://www.sanger.ac.uk/sanger/StatGen_Gwava; UKBB, https://www.ukbiobank.ac.uk/; GDSC, https://www.cancerrxgene.org/.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (82103929, 82273713), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), Fundamental Research Funds for the Central Universities (WHU:2042022kf1205), Knowledge Innovation Program of Wuhan (whkxjsj011) and Translational Medicine and Interdisciplinary Research Joint Fund of Zhongnan Hospital of Wuhan University (ZNJC202207) for Jianbo Tian; Distinguished Young Scholars of China (81925032), Key Program of National Natural Science Foundation of China (82130098), the Leading Talent Program of the Health Commission of Hubei Province, Natural Science Foundation of Hubei Province (2019CFA009) and the Fundamental Research Funds for the Central Universities (2042022rc0026, 2042023kf1005) for Xiaoping Miao; the National Natural Science Foundation of China (82204128) for Xiaoyang Wang. The authors thank the numerous subjects, their families, and referring physicians that have participated in the studies.

Compliance and ethics

The author(s) declare that they have no conflict of interest. The procedures related to human subjects of our study were approved by the Chinese Academy of Medical Sciences Cancer Institute and the institutional review board of Tongji Medical College, Huazhong University of Science and Technology. The study was conducted in accordance with the Helsinki Declaration of 1975 (as revised in 2008).

Supporting information

The supporting information is available online at https://doi.org/10.1007/s11427-023-2439-7. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Al-Tassan, N.A., Whiffin, N., Hosking, F.J., Palles, C., Farrington, S.M., Dobbins, S.E., Harris, R., Gorman, M., Tenesa, A., Meyer, B.F., et al. (2015). A new GWAS and

meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci Rep 5, 10442.
Brænne, I., Civelek, M., Vilne, B., Di Narzo, A., Johnson, A.D., Zhao, Y., Reiz, B., Codoni, V., Webb, T.R., Foroughi Asl, H., et al. (2015). Prediction of causal candidate genes in coronary artery disease loci. Arterioscler Thromb Vasc Biol 35, 2207–2217.
Buys, S.S., Partridge, E., Black, A., Johnson, C.C., Lamerato, L., Isaacs, C., Reding, D.J., Greenlee, R.T., Yokochi, L.A., Kessel, B., et al. (2011). Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening randomized controlled trial. JAMA 305, 2295–2303.
Cao, R., Yang, F., Ma, S.C., Liu, L., Zhao, Y., Li, Y., Wu, D.H., Wang, T., Lu, W.J., Cai, W.J., et al. (2020). Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. Theranostics 10, 11080–11091.
Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2, 401–404.
Chang, J., Tian, J., Yang, Y., Zhong, R., Li, J., Zhai, K., Ke, J., Lou, J., Chen, W., Zhu, B., et al. (2018). A rare missense variant in TCF7L2 associates with colorectal cancer risk by interacting with a GWAS-identified regulatory variant in the MYC enhancer. Cancer Res 78, 5164–5172.
Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., and He, J. (2016). Cancer statistics in China, 2015. CA Cancer J Clin 66, 115–132.
Chen, Z., and Schunkert, H. (2021). Genetics of coronary artery disease in the post-GWAS era. J Intern Med 290, 980–992.
Cheng, P., Chen, Y., He, T., Wang, C., Guo, S., Hu, H., Ni, C., Jin, G., and Zhang, Y. (2019). Menin coordinates C/EBPβ-mediated TGF-β signaling for epithelial-mesenchymal transition and growth inhibition in pancreatic cancer. Mol Ther Nucleic Acids 18, 155–165.
Codrich, M., Dalla, E., Mio, C., Antoniali, G., Malfatti, M.C., Marzinotto, S., Pierobon, M., Baldelli, E., Di Loreto, C., Damante, G., et al. (2021). Integrated multi-omics analyses on patient-derived CRC organoids highlight altered molecular pathways in colorectal cancer progression involving PTEN. J Exp Clin Cancer Res 40, 198.
Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.
Cui, R., Okada, Y., Jang, S.G., Ku, J.L., Park, J.G., Kamatani, Y., Hosono, N., Tsunoda, T., Kumar, V., Tanikawa, C., et al. (2011). Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut 60, 799–805.
Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46, D794–D801.
Dunlop, M.G., Dobbins, S.E., Farrington, S.M., Jones, A.M., Palles, C., Whiffin, N., Tenesa, A., Spain, S., Broderick, P., Ooi, L.Y., et al. (2012). Common variation near *CDKN1A*, *POLD3* and *SHROOM2* influences colorectal cancer risk. Nat Genet 44, 770–776.
Fortini, B.K., Tring, S., Devall, M.A., Ali, M.W., Plummer, S.J., and Casey, G. (2021). SNPs associated with colorectal cancer at 15q13.3 affect risk enhancers that modulate *GREM1* gene expression. Hum Mutat 42, 237–245.
Fultang, N., Li, X., Li, T., and Chen, Y.H. (2020). Myeloid-derived suppressor cell differentiation in cancer: transcriptional regulators and enhanceosome-mediated mechanisms. Front Immunol 11, 619253.
Groth, C., Hu, X., Weber, R., Fleming, V., Altevogt, P., Utikal, J., and Umansky, V. (2019). Immunosuppression mediated by myeloid-derived suppressor cells (MDSCs) during tumour progression. Br J Cancer 120, 16–25.
Hagège, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forné, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc 2, 1722–1733.
Hofer, P., Hagmann, M., Brezina, S., Dolejsi, E., Mach, K., Leeb, G., Baierl, A., Buch, S., Sutterlüty-Fall, H., Karner-Hanusch, J., et al. (2017). Bayesian and frequentist analysis of an Austrian genome-wide association study of colorectal cancer and advanced adenomas. Oncotarget 8, 98623–98634.
Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. Am J Hum Genet 99, 1245–1260.
Houlston, R.S., Cheadle, J., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Spain, S.L., Broderick, P., Domingo, E., Farrington, S., et al. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet 42, 973–977.
Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., et al. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 40: 1426–1435.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57.
Huyghe, J.R., Bien, S.A., Harrison, T.A., Kang, H.M., Chen, S., Schmit, S.L., Conti, D. V., Qu, C., Jeon, J., Edlund, C.K., et al. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet 51, 76–87.
Jaeger, E., Webb, E., Howarth, K., Carvajal-Carmona, L., Rowan, A., Broderick, P., Walther, A., Spain, S., Pittman, A., Kemp, Z., et al. (2008). Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 40, 26–28.
Jia, W.H., Zhang, B., Matsuo, K., Shin, A., Xiang, Y.B., Jee, S.H., Kim, D.H., Ren, Z., Cai, Q., Long, J., et al. (2013). Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal cancer. Nat Genet 45, 191–196.
Jiang, K., Sun, Y., Wang, C., Ji, J., Li, Y., Ye, Y., Lv, L., Guo, Y., Guo, S., Li, H., et al. (2015). Genome-wide association study identifies two new susceptibility loci for colorectal cancer at 5q23.3 and 17q12 in Han Chinese. Oncotarget 6, 40327–40336.
Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B.J., Carlson, C.S., Chan, A.T., Chang-Claude, J., Chanock, S., et al. (2014). Estimating the heritability of colorectal cancer. Hum Mol Genet 23, 3898–3905.
Jin, T., Rehani, P., Ying, M., Huang, J., Liu, S., Roussos, P., and Wang, D. (2021). scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. Genome Med 13, 95.
Ju, W., Zheng, R., Zhang, S., Zeng, H., Sun, K., Wang, S., Chen, R., Li, L., Wei, W., and He, J. (2023). Cancer statistics in Chinese older people, 2022: current burden, time trends, and comparisons with the US, Japan, and the Republic of Korea. Sci China Life Sci 66, 1079–1091.
Kapoor, M., Chao, M.J., Johnson, E.C., Novikova, G., Lai, D., Meyers, J.L., Schulman, J., Nurnberger Jr, J.I., Porjesz, B., Liu, Y., et al. (2021). Multi-omics integration analysis identifies novel genes for alcoholism with potential overlap with neurodegenerative diseases. Nat Commun 12, 5071.
Labrecque, M.P., Coleman, I.M., Brown, L.G., True, L.D., Kollath, L., Lakely, B., Nguyen, H.M., Yang, Y.C., da Costa, R.M.G., Kaipainen, A., et al. (2019). Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer. J Clin Invest 129, 4492–4505.
Law, P.J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J., Farrington, S., Svinti, V., Palles, C., Orlando, G., et al. (2019). Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun 10, 2154.
Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518, 350–354.
Li, H.J., Qu, N., Hui, L., Cai, X., Zhang, C.Y., Zhong, B.L., Zhang, S.F., Chen, J., Xia, B., Wang, L., et al. (2020). Further confirmation of netrin 1 receptor (DCC) as a depression risk gene via integrations of multi-omics data. Transl Psychiatry 10, 98.
Liu, X.Z., Rulina, A., Choi, M.H., Pedersen, L., Lepland, J., Takle, S.T., Madeleine, N., Peters, S.D., Wogsland, C.E., Grøndal, S.M., et al. (2022). C/EBPB-dependent adaptation to palmitic acid promotes tumor formation in hormone receptor negative breast cancer. Nat Commun 13, 69.
Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16, 22.
Lu, Y., Kweon, S.S., Tanikawa, C., Jia, W.H., Xiang, Y.B., Cai, Q., Zeng, C., Schmit, S. L., Shin, A., Matsuo, K., et al. (2019). Large-scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. Gastroenterology 156, 1455–1466.
Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195.
McFarland, J.M., Ho, Z.V., Kugener, G., Dempster, J.M., Montgomery, P.G., Bryan, J.G., Krill-Burger, J.M., Green, T.M., Vazquez, F., Boehm, J.S., et al. (2018). Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat Commun 9, 4610.
Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet 49, 1779–1784.
Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. Science 266, 66–71.
Negrini, S., Prada, I., D'Alessandro, R., and Meldolesi, J. (2013). REST: an oncogene or a tumor suppressor? Trends Cell Biol 23, 289–295.
Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from

tissue expression profiles. Nat Methods 12, 453–457.
Orlando, G., Law, P.J., Cornish, A.J., Dobbins, S.E., Chubb, D., Broderick, P., Litchfield, K., Hariri, F., Pastinen, T., Osborne, C.S., et al. (2018). Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. Nat Genet 50, 1375–1380.
Orlando, G., Law, P.J., Palin, K., Tuupanen, S., Gylfe, A., Hänninen, U.A., Cajuso, T., Tanskanen, T., Kondelin, J., Kaasinen, E., et al. (2016). Variation at 2q35 (*PNKD* and *TMBIM1*) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. Hum Mol Genet 25, 2349–2359.
Peters, U., Jiao, S., Schumacher, F.R., Hutter, C.M., Aragaki, A.K., Baron, J.A., Berndt, S.I., Bézieau, S., Brenner, H., Butterbach, K., et al. (2013). Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. Gastroenterology 144, 799–807.e24.
Prorok, P.C., Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., David Crawford, E., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A., et al. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Control Clin Trials 21, 273S–309S.
Rani, J., Shah, A.R., and Ramachandran, S. (2015). pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. J Biosci 40, 671–682.
Rask, K., Thörn, M., Pontén, F., Kraaz, W., Sundfeldt, K., Hedin, L., and Enerbäck, S. (2000). Increased expression of the transcription factors CCAAT-enhancer binding protein-B (C/EBB) and C/EBPζ (CHOP) correlate with invasiveness of human colorectal cancer. Int J Cancer 86, 337–343.
Real, L.M., Ruiz, A., Gayán, J., González-Pérez, A., Sáez, M.E., Ramírez-Lorca, R., Morón, F.J., Velasco, J., Marginet-Flinch, R., Musulén, E., et al. (2014). A colorectal cancer susceptibility new variant at 4q26 in the Spanish population identified by genome-wide association analysis. PLoS ONE 9: e101178.
Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature 489, 109–113.
Schmit, S.L., Edlund, C.K., Schumacher, F.R., Gong, J., Harrison, T.A., Huyghe, J.R., Qu, C., Melas, M., Van Den Berg, D.J., Wang, H., et al. (2019). Novel common genetic susceptibility loci for colorectal Cancer. J Natl Cancer Inst 111, 146–157.
Schmit, S.L., Schumacher, F.R., Edlund, C.K., Conti, D.V., Raskin, L., Lejbkowicz, F., Pinchev, M., Rennert, H.S., Jenkins, M.A., Hopper, J.L., et al. (2014). A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. Carcinogenesis 35, 2512–2519.
Schumacher, F.R., Schmit, S.L., Jiao, S., Edlund, C.K., Wang, H., Zhang, B., Hsu, L., Huang, S.C., Fischer, C.P., Harju, J.F., et al. (2015). Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun 6, 7138.
Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507, 371–375.
Tang, X., Liang, Y., Sun, G., He, Q., Qu, H., and Gao, P. (2021). UBQLN4 is activated by C/EBPβ and exerts oncogenic effects on colorectal cancer via the Wnt/β-catenin signaling pathway. Cell Death Discov 7, 398.
Tanikawa, C., Kamatani, Y., Takahashi, A., Momozawa, Y., Leveque, K., Nagayama, S., Mimori, K., Mori, M., Ishii, H., Inazawa, J., et al. (2018). GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12. Carcinogenesis 39, 652–660.
Tenesa, A., Farrington, S.M., Prendergast, J.G.D., Porteous, M.E., Walker, M., Haq, N., Barnetson, R.A., Theodoratou, E., Cetnarskyj, R., Cartwright, N., et al. (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 40, 631–637.
Tomlinson, I.P., Carvajal-Carmona, L.G., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Palles, C., Broderick, P., Jaeger, E.E., Farrington, S., et al. (2011). Multiple common susceptibility variants near bmp pathway loci grem1, bmp4, and bmp2 explain part of the missing heritability of colorectal cancer. PLoS Genet, 7, e1002105.
Tomlinson, I.P., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A.M., Spain, S., Lubbe, S., Walther, A., Sullivan, K., et al. (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 40, 623–630.
Vangala, D., Ladigan, S., Liffers, S.T., Noseir, S., Maghnouj, A., Götze, T.M., Verdoodt, B., Klein-Scory, S., Godfrey, L., Zowada, M.K., et al. (2021). Secondary resistance to anti-EGFR therapy by transcriptional reprogramming in patient-derived colorectal cancer models. Genome Med 13, 116.
Wang, D., Yang, L., Yu, W., Wu, Q., Lian, J., Li, F., Liu, S., Li, A., He, Z., Liu, J., et al. (2019b). Colorectal cancer cell-derived CCL20 recruits regulatory T cells to promote chemoresistance via FOXO1/CEBPB/NF-κB signaling. J Immunother Cancer 7, 215.
Wang, H., Schmit, S.L., Haiman, C.A., Keku, T.O., Kato, I., Palmer, J.R., van den Berg, D., Wilkens, L.R., Burnett, T., Conti, D.V., et al. (2017). Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans. Intl J Cancer 140, 2728–2733.
Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81, 1278–1283.
Wang, M., Gu, D., Du, M., Xu, Z., Zhang, S., Zhu, L., Lu, J., Zhang, R., Xing, J., Miao, X., et al. (2016). Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. Nat Commun 7, 11478.
Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., Zhong, X., Tao, R., Wen, Z., Sutcliffe, J.S., et al. (2019a). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. Nat Neurosci 22, 691–699.
Whiffin, N., Hosking, F.J., Farrington, S.M., Palles, C., Dobbins, S.E., Zgaga, L., Lloyd, A., Kinnersley, B., Gorman, M., Tenesa, A., et al. (2014). Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. Hum Mol Genet 23, 4729–4737.
Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.Y., and Wei, L. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39, W316–W322.
Yang, D., Jang, I., Choi, J., Kim, M.S., Lee, A.J., Kim, H., Eom, J., Kim, D., Jung, I., and Lee, B. (2018). 3DIV: a 3D-genome Interaction Viewer and database. Nucleic Acids Res 46, D52–D57.
Yang, X., Zou, J., Cai, H., Huang, X., Yang, X., Guo, D., and Cao, Y. (2017). Ginsenoside Rg3 inhibits colorectal tumor growth via down-regulation of C/EBPβ/NF-κB signaling. Biomed Pharmacother 96, 1240–1245.
Yin, R., Song, B., Wang, J., Shao, C., Xu, Y., Jiang, H. (2022). Genome-wide association and transcriptome-wide association studies identify novel susceptibility genes contributing to colorectal cancer. J Immunol Res 2022: 5794055.
Yuan, Y., Bao, J., Chen, Z., Villanueva, A.D., Wen, W., Wang, F., Zhao, D., Fu, X., Cai, Q., Long, J., et al. (2021). Multi-omics analysis to identify susceptibility genes for colorectal cancer. Hum Mol Genet 30, 321–330.
Zeng, C., Matsuda, K., Jia, W.H., Chang, J., Kweon, S.S., Xiang, Y.B., Shin, A., Jee, S. H., Kim, D.H., Zhang, B., et al. (2016). Identification of susceptibility loci and genes for colorectal cancer risk. Gastroenterology 150, 1633–1645.
Zhang, B., Jia, W.H., Matsuda, K., Kweon, S.S., Matsuo, K., Xiang, Y.B., Shin, A., Jee, S.H., Kim, D.H., Cai, Q., et al. (2014b). Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat Genet 46, 533–542.
Zhang, B., Jia, W.H., Matsuo, K., Shin, A., Xiang, Y.B., Matsuda, K., Jee, S.H., Kim, D. H., Cheah, P.Y., Ren, Z., et al. (2014a). Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians. Int J Cancer 135, 948–955.
Zhou, Z., Shu, Y., Bao, H., Han, S., Liu, Z., Zhao, N., Yuan, W., Jian, C., and Shu, X. (2022). Stress-induced epinephrine promotes epithelial-to-mesenchymal transition and stemness of CRC through the CEBPB/TRIM2/P53 axis. J Transl Med 20, 262.
Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G. W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 48, 481–487.
Zou, J., Li, H., Chen, X., Zeng, S., Ye, J., Zhou, C., Liu, M., Zhang, L., Yu, N., Gan, X., et al. (2014). C/EBPβ knockdown protects cardiomyocytes from hypertrophy via inhibition of p65-NFκB. Mol Cell Endocrinol 390, 18–25.