Common variation in a long non-coding RNA gene modulates variation of circulating TGF-*β*2 levels in metastatic colorectal cancer patients (Alliance)

Julia C.F. Quintanilha[1], Alexander B. Sibley[2], Yingmiao Liu[3], Donna Niedzwiecki[2,4,5], Susan Halabi[2,4,5], Layne Rogers[2], Bert O'Neil[6], Hedy Kindler[7], William Kelly[8], Alan Venook[9], Howard L. McLeod[10,11], Mark J. Ratain[7], Andrew B. Nixon[3], Federico Innocenti[12†] and Kouros Owzar[2,4,5*†]

Abstract

Background Herein, we report results from a genome-wide study conducted to identify protein quantitative trait loci (pQTL) for circulating angiogenic and inflammatory protein markers in patients with metastatic colorectal cancer (mCRC). The study was conducted using genotype, protein marker, and baseline clinical and demographic data from CALGB/SWOG 80405 (Alliance), a randomized phase III study designed to assess outcomes of adding VEGF or EGFR inhibitors to systemic chemotherapy in mCRC patients. Germline DNA derived from blood was genotyped on whole-genome array platforms. The abundance of protein markers was quantified using a multiplex enzyme-linked immunosorbent assay from plasma derived from peripheral venous blood collected at baseline. A robust rank-based method was used to assess the statistical significance of each variant and protein pair against a strict genome-wide level. A given pQTL was tested for validation in two external datasets of prostate (CALGB 90401) and pancreatic cancer (CALGB 80303) patients. Bioinformatics analyses were conducted to further establish biological bases for these findings.

Results The final analysis was carried out based on data from 540,021 common typed genetic variants and 23 protein markers from 869 genetically estimated European patients with mCRC. Correcting for multiple testing, the analysis discovered a novel *cis*-pQTL in *LINC02869*, a long non-coding RNA gene, for circulating TGF-*β*2 levels (rs11118119; AAF = 0.11; *P*-value < 1.4e-14). This finding was validated in a cohort of 538 prostate cancer patients from CALGB 90401 (AAF = 0.10, *P*-value < 3.3e-25). The analysis also validated a *cis*-pQTL we had previously reported for VEGF-A in advanced pancreatic cancer, and additionally identified *trans*-pQTLs for VEGF-R3, and *cis*-pQTLs for CD73.

†Federico Innocenti and Kouros Owzar contributed equally to this work.

*Correspondence:
Kouros Owzar
kouros.owzar@duke.edu

Full list of author information is available at the end of the article

Conclusions This study has provided evidence of a novel *cis* germline genetic variant that regulates circulating TGF-*β*2 levels in plasma of patients with advanced mCRC and prostate cancer. Moreover, the validation of previously identified pQTLs for VEGF-A, CD73, and VEGF-R3, potentiates the validity of these associations.

Keywords Cancer, Common germline variation, Genome-wide analysis study, Circulating protein biomarkers, Protein quantitative trait loci, Inflammatory markers, Angiogenesis markers, TGF-*β*2, Long non-coding RNA

Background

The heritability of circulating protein abundance and evidence showing the influence of germline genetic variants in circulating protein levels have raised the interest in protein quantitative trait loci (pQTL) studies. pQTL studies have the objective of determining the impact of germline genetic variants on circulating protein levels. Circulating protein levels are involved in diverse biological processes, including disease development and response to medications. pQTL analyses can contribute notably to the discovery of new clinically relevant biomarkers and to better understanding the factors that regulate circulating proteins and the pathways involved in these biological processes [1, 2].

Colorectal cancer (CRC) is the third most common type of cancer and the second leading cancer-related death worldwide [3]. Many studies have attempted to identify circulating proteins as biomarkers in CRC patients, including biomarkers for the early detection of CRC [4, 5], prognosis [6, 7], treatment response [7], regional tumor localization [6], and disease dissemination [6]. Thus, the assessment of the impact of germline genetic variants on circulating protein levels through pQTL analyses in CRC patients can potentially lead to insights into the mechanisms involved in CRC development and treatment outcome.

Herein, we report results from a study of common genetic variation with respect to variation in circulating proteins with putative inflammatory or angiogenic function in patients with metastatic (m) CRC. Specifically, we sought to identify functional *cis-* and *trans*-pQTL variants using genome-wide germline genotyping data and circulating protein levels measured using a custom panel of putative cancer-related angiogenic and inflammatory markers. To this end, we used clinical, genotyping, and pre-treatment candidate protein marker data obtained from patients with mCRC randomized to the Cancer and Leukemia Group B (CALGB, now part of the Alliance for Clinical Trials in Oncology (Alliance)) and the Southwest Oncology Group, CALGB/SWOG 80405. This was a phase III study randomizing mCRC patients to receive cetuximab, an epidermal growth factor receptor (EGFR) inhibiting monoclonal antibody, or bevacizumab, a vascular endothelial growth factor (VEGF) inhibiting monoclonal antibody, or the combination of the two in addition to systemic chemotherapy [8].

After accounting for multiple testing, our analysis discovered a novel *cis*-pQTL in the intronic region of *LINC02869* (alias *C1orf143*) for circulating TGF-*β*2 at the two-sided genome-wide level of 0.05. The novel *cis*-pQTL for TGF-*β*2 was then tested for validation in independent external cohorts of castration-resistant prostate cancer (CALGB 90401) and advanced pancreatic cancer patients (CALGB 80303) [9–13].

This analysis also validated a *cis*-pQTL for VEGF-A that we had previously identified in a cohort of patients with locally advanced or metastatic pancreatic cancer [13]. Finally, our study identified *trans*-pQTLs for VEGF-R3 and *cis*-pQTLs for CD73.

While the scope of our analysis was genome-wide, this paper is primarily focused on the presentation of its novel findings. In addition, we have provided a high-level summary of the other significant results from our analysis which reproduce our own previous findings and those of others for the sake of completeness, and to further establish the reliability of our data and approach.

Methods

Clinical data

Patients registered to CALGB/SWOG 80405 were randomized to receive bevacizumab, cetuximab, or the combination of these two monoclonal antibodies, in addition to systemic chemotherapy. For the latter, the choice of a FOLFOX- or FOLFIRI-based regimen was at the discretion of the treating physician. The study was later amended by restricting participation to patients with wild-type KRAS tumors and by terminating the combination arm. Additional details on the design of the study, its amendments, and clinical baseline characteristics and outcomes for its patients have been reported in a primary clinical publication and its supplementary material [8]. Baseline demographic and clinical data used in the present analyses were obtained from the database used to generate the analyses reported in that publication.

Genotyping data

Germline DNA was extracted from peripheral blood. The genotyping was conducted in two separate batches using the Illumina Human OmniExpress (12v1) and the Illumina Human OmniExpressExome (8v1) platforms, respectively, by the Core of Genomic Medicine of the RIKEN institute in Yokohama, Japan. The genotyping design included the use of HapMap controls as well

as inter- and intra-plate replicates. The analysis data set was constructed on the basis of the intersection of the variants across these two platforms identified by their respective dbSNP Reference SNP ID (rsID). A number of quality control (QC) metrics, including genotype calling rate, AAF, Hardy-Weinberg *P*-values, were used to filter out variants. Additional technical details on the genotyping and QC processes have been previously reported [14].

Circulating protein markers
Levels of 23 soluble proteins (angiopoietin-2, HGF, ICAM-1, IL-6, OPN, PDGF-AA, PDGF-BB, PlGF, SDF-1, TGF-β1, TGF-β2, TIMP-1, TSP-2, VCAM-1, VEGF-A, VEGF-D, VEGF-R1, VEGF-R2, VEGF-R3, BMP-9, CD73, HER-3, TGFβ-R3) were measured in plasma from peripheral venous blood collected at baseline using multiplex enzyme-linked immunosorbent assay (ELISA). The plasma was double-spun, aliquoted, and frozen in liquid nitrogen. Additional technical details on this panel, including CVs, lower limits of quantitation, and limits of detection, have been previously reported [13, 15–17]. The analyses reported herein are based on measurements taken at baseline prior to any CALGB/SWOG 80405 protocol-directed treatment.

Statistical considerations
To ensure robustness against outliers and influential data points and deviations from normality assumptions, the Jonckheere-Terpstra statistic [18, 19] was used for the discovery of pQTLs. Then variance approximation provided in expression 6.19 in Hollander et al. [20], implemented by the fastJT package [21], was used to derive a standardized statistic whose null sampling distribution was approximated using a standard normal distribution.

To properly account for multiple testing in the discovery process, a conservative two-sided genome-wide significance level of 0.05/K, where K denotes the number of single nucleotide polymorphisms (SNPs) and protein marker pairs tested in the final analysis, was used. The potential confounding effects of baseline covariates, including age at time of registration (log base 10 transformed), self-reported gender, and global ancestry, were assessed using a robust linear regression rank-based approach implemented by the Rfit [22] package. The genotype effect was quantified on the additive scale as the number of copies of the alternate allele (additive genetic model), and global ancestry was inferred for patients previously identified as genetic Europeans [14] using the first three principal components estimated using the SNPRelate R package [23]. The Hodges-Lehmann-Sen estimator was used to estimate the location parameter for the distribution of the abundance of a protein conditional on the genotype. The per allele effect size was estimated as the ratio of the location parameter estimates. A 95%

exact confidence interval was calculated for each location parameter. These estimates were meant to serve as descriptive measures, and accordingly, the corresponding confidence levels were not adjusted for multiple testing. For each protein marker, the distribution of the unadjusted *P*-values was examined using Manhattan and QQ plots.

All statistical analyses were conducted using the R statistical environment [24] and its extension packages, including those from the tidyverse [25] ecosystem, foreach [26], SeqArray [27], kableExtra [28], knitr [29] and rmarkdown [30]. SNP and gene positions are reported per GRCh37.

Replication analysis
For a given pQTL pair, the Jonckheere-Terpstra statistic [18, 19] with the genotype effect quantified on the additive scale as the number of copies of the alternate allele (additive genetic model) was used to estimate the pQTL association in two independent external datasets, CALGB 90401 [9] and CALGB 80303 [11, 13]. CALGB 90401 included metastatic castration-resistant prostate cancer randomized to receive docetaxel in combination with prednisone on day 1 plus either placebo or bevacizumab every 21 days. CALGB 80303 included patients with advanced pancreatic cancer randomized to receive gemcitabine on days 1, 8, and 15 plus either placebo or bevacizumab on days 1 and 15. Additional details on the design of both studies, and clinical baseline characteristics and outcomes for its patients have been reported in primary clinical publications [9, 11].

Bioinformatics considerations
For a given pQTL pair, the extent of the signal, quantified by unadjusted *P*-values, relative to the positions of variants and their linkage disequilibrium (LD) within regions of annotated genes, was assessed visually using Locus Zoom ([31]; version 1.4) plots. The June 2010 release of The 1000 Genomes Pilot 1 EUR panel (November 2014; hg19 coordinates using GENCODE gene annotation [32]) was used as the reference. Putative functional effects were investigated using RegulomeDB [33], USCS Genome Browser [34], Haploreg [35], and SNPNexus [36]. AtSNP was used to quantify the impact of SNPs on transcription factor binding [37].

Results
The final analysis was conducted on the basis of a data set comprised of 540,021 SNPs, 23 baseline protein markers, and baseline demographic and clinical data from 869 genetically estimated European mCRC patients from CALGB/SWOG 80405 for whom protein marker data was available. The Consolidated Standards of Reporting Trials (CONSORT [38]) chart displayed in Innocenti,

et al. [14] provides additional details on the sample and variant selection process leading to the final analysis data set. Table 1 provides summaries of baseline demographic and clinical data for the analysis cohort of this study. The two-sided genome-wide level significance threshold was set to be 4.03e-09. At this level, 37 candidate pQTLs across four proteins, TGF-β2, VEGF-A, VEGF-R3, and CD73, were identified based on our pre-specified statistical decision rule. Overview and details for these candidates are illustrated in the Circos [39] plot in Fig. 1, and summarized in Supplementary Tables 1 and 2, Additional File 1. The Manhattan and quantile-quantile (QQ)-plots for TGF-β2, VEGF-A, VEGF-R3, and CD73 are shown in Supplementary Figs. 1–4, Additional File 1, respectively. Finally, for each of the 23 proteins, the top 100 pQTLs, ranked according to the corresponding unadjusted P-values are shown in Additional File 2.

Our analysis identified a novel pQTL for TGF-β2: rs11118119 (chr1:218693872; A>G; alternate allele frequency (AAF)=0.11; P-value<1.4e-14) is in the intronic region of *LINC02869* (alias *C1orf143*). This variant is located 75,911 bases downstream from *TGFB2*. The genotypic Hodges-Lehmann-Sen estimates are 95.8

Table 1 Demographics and clinical characteristics of patients of genetically determined European ancestry included in the genome-wide pQTL analysis of CALGB/SWOG 80405

	Patients $n=869$
Age (years) – median (range)	59.7 (21.8–85.3)
Gender – n (%)	
Male	506 (58.2%)
Female	363 (41.8%)
Treatment arm – n (%)	
Chemotherapy/Bevacizumab	340 (39.1%)
Chemotherapy/Cetuximab	312 (35.9%)
Chemotherapy/Bevacizumab/Cetuximab	217 (25.0%)
Chemotherapy – n (%)	
FOLFOX	666 (76.6%)
FOLFIRI	203 (23.4%)
Prior adjuvant chemotherapy – n (%)	
No	744 (85.6%)
Yes	125 (14.4%)
Prior pelvic radiation – n (%)	
No	785 (90.3%)
Yes	84 (9.7%)
ECOG PS – n (%)	
0	545 (62.7%)
1	324 (37.3%)
Tumor location – n (%)	
Left	497 (57.2%)
Right	230 (26.5%)
Transverse	62 (7.1%)
Multiple	4 (0.5%)
Unknown	76 (8.7%)

(n=692; 95% CI=91.6, 100), 141 (n=164; 95% CI=128, 155) and 207 (n=12; 95% CI=140, 295) for genotypes AA, AG and GG respectively, while the median observed values were 87.2, 130.2, and 206.9, respectively. The estimated effect size in the rank-based linear model was 1.68 (CI=1.51, 1.88). This variant is in moderate LD (R^2=0.67 in the analysis data set) with rs725033 (chr1:218643940; G>A; AAF=0.12; P-value<1.6e-10) an intergenic variant 25,979 bases upstream from *TGFB2*. See Table 2; Fig. 2, and Supplementary Figs. 5 and 6, Additional File 1 for box and locus zoom plots of rs11118119 and rs725033, respectively.

Our analysis provided strong confirmatory evidence for a *cis*-pQTL we have previously reported for circulating VEGF-A in patients with locally advanced or metastatic pancreatic cancer: rs7767396 (chr6:43927050; A>G; AAF=0.48; P-value<2.6e-12), an intergenic variant 172,826 bases downstream from *VEGFA* and 41,272 bases upstream from *C6orf223*. See Table 2, and box and locus zoom plots in Supplementary Fig. 7, Additional File 1.

Our analysis identified *trans*-pQTLs for VEGF-R3 on chromosomes 3 and 9. These include rs10935473 (chr3:98416900; C>A; AAF=0.45; P-value<6.4e-39), an intergenic variant 16,277 bases upstream from *ST3GAL6-AS1*, intronic variants in *CPOX* (e.g., rs3804622; chr3:98303182; G>A; AAF=0.51; P-value<5.7e-23), intronic variants in *ST3GAL6-AS1* (e.g., rs844159; chr3:98443648; A>G; AAF=0.46; P-value<1.1e-19) and an intronic variant in *ABO* blood group gene (rs507666; chr9:136149399; G>A; AAF=0.20; P-value<9.1e-12). See Table 2, and Supplementary Figs. 8, 9, 10, and 11, Additional File 1 for box and locus zoom plots of rs10935473, rs3804622, rs844159, and rs507666, respectively.

Finally, we identified *cis*-pQTLs for CD73: rs2229523 (chr6:86199233; G>A; AAF=0.34; P-value<2.3e-15), a non-synonymous variant in *NT5E*, the gene that codes CD73, and an intergenic variant, rs494688 (chr6:86100089; G>A; AAF=0.11; P-value<1.9e-09), 59,712 bases upstream from *NT5E*. The estimated R^2 between rs2229523 and rs494688 was 0.04 in the analysis data set. See Table 2, and Supplementary Figs. 12 and 13, Additional File 1 for box and locus zoom plots of rs2229523, and rs494688, respectively.

Validation of the association between rs11118119 and TGF-β2 levels

In order to validate the novel *cis*-pQTL for TGF-β2 identified in our analysis (rs11118119, chr1:218693872; A>G), we tested the association between rs11118119 and TGF-β2 levels in 538 castration-resistant prostate cancer patients from CALGB 90401 and 216 advanced pancreatic cancer patients from CALGB 80303 [13]. Selected
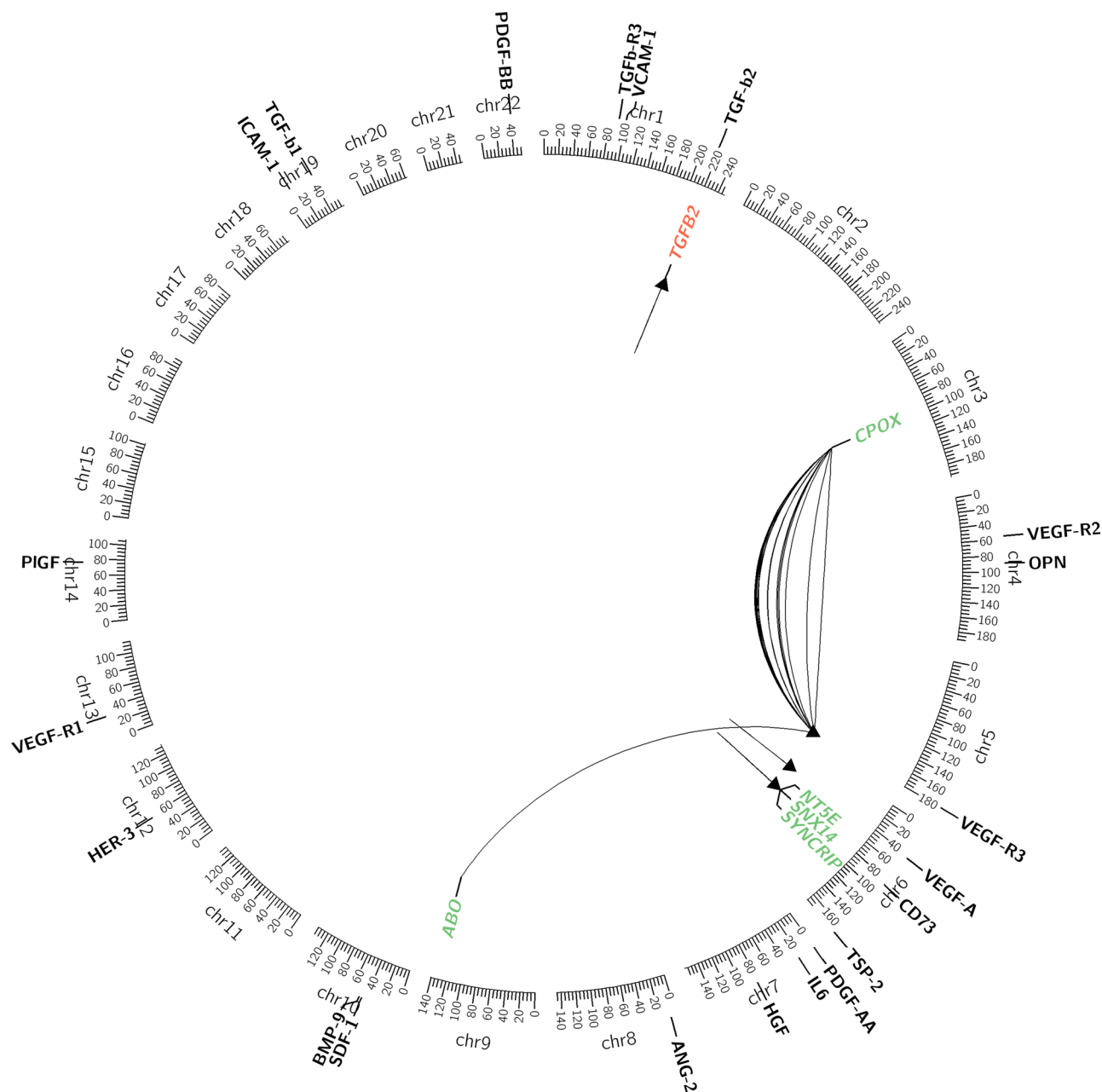
Fig. 1 Chromosome-based Circos plot for pQTL that passed the genome-wide threshold. The colors indicate if a gene contains one of the top SNPs (green) or is a flanking gene (red). Links with less curvature indicate smaller *P*-values

baseline characteristics for these two cohorts are summarized in Supplementary Table 3, Additional File 1.

The association was validated independently in CALGB 90401, where the G allele of rs11118119 (A > G) was associated with higher TGF-β2 levels (*P*-value < 3.3e-25, AAF = 0.10), similar to CALGB/SWOG 80405 (Fig. 2). We could not validate this association in CALGB 80303 (*P*-value = 0.500, AAF = 0.12, Fig. 2).

Bioinformatic analysis of rs11118119

Our bioinformatic analysis showed that rs1015275 (G > C), in high LD (R^2 = 0.91) with rs11118119 (A > G) and located 50 kb downstream from *TGFB2*, is located in the binding motif for the HAND1 transcription factor. Additional data from the JASPAR database [40] (using atSNP [37]) predicts preferential binding of HAND1 to the C allele of rs1015275 compared to the G allele (*p* = 0.0038, log-likelihood = -4.46 for the C allele, and *p* = 5.93×10^{-6}, log-likelihood = -27.45, *p* = 0.321 for the G allele) (Supplementary Fig. 14, Additional File 1). This

Table 2 Results and annotation for pQTLs that passed the genome-wide threshold

Protein/rsID	Chr	Base	Gene	N	AAF	HWE	Pvalue	Est	P.c	Est.c
CD73										
rs2229523	chr6	G > A	*NT5E*	869	0.34	0.82	2.3E-15	0.52	9.8E-17	0.52
rs494688	chr6	G > A	*DUTP5/ RP11-30P6.6*	869	0.11	0.73	1.9E-09	0.61	8.1E-10	0.60
TGF-*β*2										
rs11118119	chr1	A > G	*C1orf143*	868	0.11	0.48	1.4E-14	0.52	3.6E-18	0.51
rs725033	chr1	G > A	*C1orf143/ TGFB2*	869	0.12	0.74	1.6E-10	0.41	2.1E-12	0.40
VEGF-A										
rs7767396	chr6	A > G	*RP5-1120P11.1/ RP11-344J7.2*	868	0.48	0.95	2.6E-12	0.51	9.3E-12	-0.51
VEGF-R3										
rs10935473	chr3	C > A	*ST3GAL6-AS1/ WWP1P1*	869	0.45	0.49	6.4E-39	0.29	1.4E-43	0.29
rs3804622	chr3	G > A	*CPOX*	869	0.51	0.46	5.7E-23	0.21	2.0E-24	0.22
rs844159	chr3	A > G	*ST3GAL6-AS1*	869	0.46	0.58	1.1E-19	0.20	5.9E-20	0.20
rs507666	chr9	G > A	*ABO*	869	0.20	0.17	9.1E-12	0.19	2.3E-12	0.19
rs579459	chr9	A > G	*Y_RNA/ ABO*	869	0.23	0.08	9.2E-12	0.18	1.3E-11	0.18

Protein/rsID: protein marker/RefSNP ID of variant; Chr: chromosome of the variant according to hg19; Base: Illumina TOP reference > alternate alleles; Gene: gene symbol for intragenic variants or nearest downstream/upstream gene symbols for intergenic variants; N: number of samples for whom the variant was called; AAF: relative allelic frequency for the alternate allele; HWE: Hardy-Weinberg *P*-value; Pvalue: *P*-value for pQTL analysis; Est: parameter estimate for the variant effect from the rank-based linear regression; P.c: *P*-value for the variant effect from the rank-based linear regression adjusting for covariates; Est.c: parameter estimate for the variant effect from the rank-based linear regression adjusting for covariates. Variants are sorted by rsID within Protein and Chr. Note that *C1orf143* is an alias for *LINC02869*, and that the genotyping data included rs3812138, but this ID has been merged with rs2229523
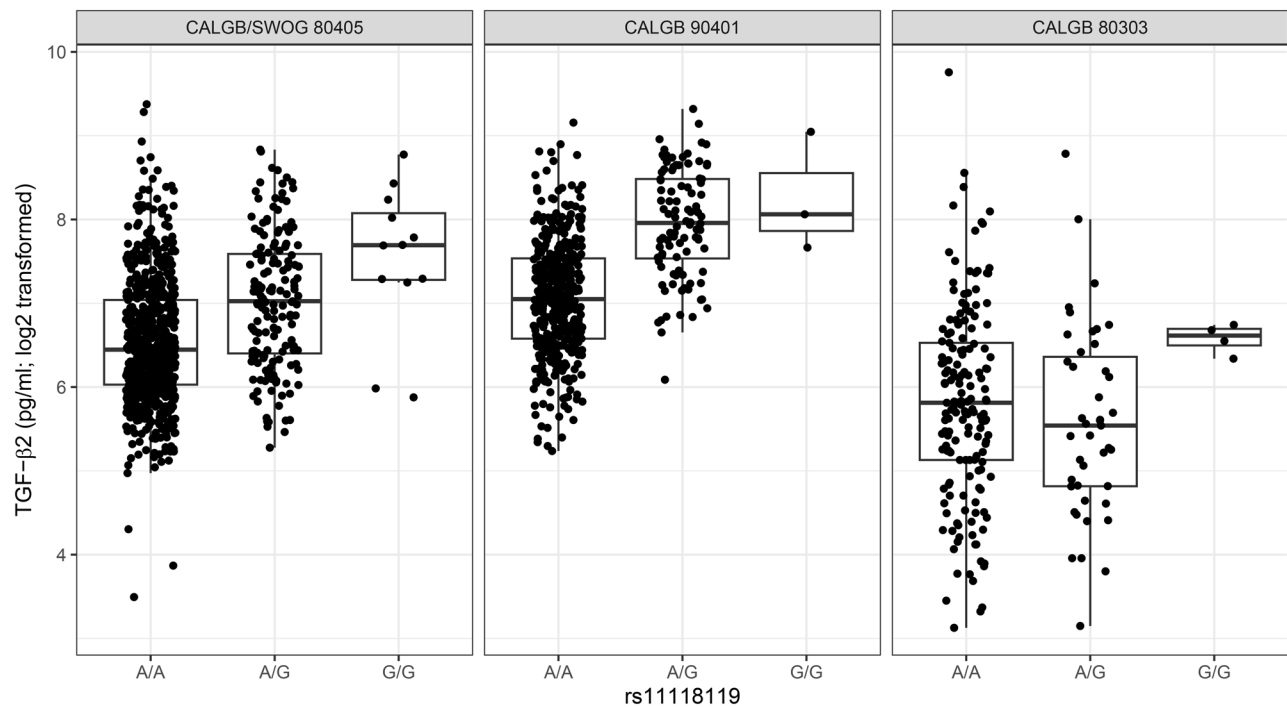

Fig. 2 Associations between rs11118119 (A > G) and TGF-*β*2 levels in CALGB/SWOG 80405, CALGB 90401, and CALGB 80303

evidence is derived from small-scale in vitro experiments for HAND1 in complex with TCF3 and TCF4. Moreover, the non-coding scores provided by SNPnexus [36] show that rs1015275 has a high predicted functionality according to EIGEN PC [41] (PC score = 1.33) and DeepSEA [42] scores (functional significance score = 0.010). High EIGEN PC scores and low DeepSEA scores indicate that

the SNP is predicted to be located in regions of open chromatin, accessible to transcription factors.

Discussion

The present study investigated the association between genetic markers and circulatory protein levels in mCRC patients and discovered a novel *cis*-pQTL for TGF-*β*2, rs11118119 (A > G) located in *LINC02869*. This finding

was supported by further confirmation in an independent external cohort of patients with castration-resistant prostate cancer. Moreover, this study also validated previously discovered *cis*-pQTLs for VEGF-A and CD73, as well as a *trans*-pQTL for VEGF-R3.

The TGF-*β* signaling pathway participates in different biological processes, including cell proliferation, differentiation, adhesion, migration, and apoptosis [43]. TGF-*β* acts as a tumor suppressor in normal epithelium cells and in the early stages of different types of cancer, including CRC [44, 45], prostate [46], and pancreatic [47]. However, in advanced cancers, TGF-*β* is abundantly expressed and acts as a tumor promoter. The TGF-*β* family consists of three members, TGF-*β*1, TGF-*β*2, and TGF-*β*3. Both TGF-*β*1 and TGF-*β*2 control the activity of stromal cells and tumor cells, affecting cancer progression [48, 49]. Higher TGF-*β*2 expression is correlated with the prognosis of different types of cancer, mainly CRC. Higher TGF-*β*2 expression has also been associated with lymph node metastasis in CRC patients and with the expression of several markers of immune cell subspecies in tumors. Thus, TGF-*β*2 expression is related to the magnitude of the tumor infiltration by immune cells, with the potential to serve as a prognostic biomarker in CRC [50].

We provided evidence of replication of the association between rs11118119 and TGF-*β*2 levels in metastatic castration-resistant prostate cancer patients from the CALGB 90401 study. Similar to mCRC patients in CALGB/SWOG 80405, the G allele of rs11118119 (A>G) was associated with higher levels of TGF-*β*2 (Fig. 2). The association was not replicated for advanced pancreatic cancer patients from CALGB 80303 (Fig. 2).

The empirical relative allelic frequencies for the risk allele of rs11118119 in the genetically estimated European cohort of CRC, prostate cancer, and pancreatic cancer patients in our study are 0.11, 0.09, and 0.12, respectively. The corresponding reported putative relative frequency in the European (EUR) cohort from the 1000 Genomes database is 0.14 compared to a putative relative frequency of 0.56 in the African (AFR) cohort [51]. Effectively, the putative risk allele for this variant is the major allele in the latter population, and this finding might impact a significant proportion of patients with advanced tumors.

Bioinformatic analyses showed that rs1015275 (G>C), a SNP in high LD with rs11118119 (A>G) located around 50 kb downstream from *TGFB2*, has a high EIGEN PC score and a low DeepSEA score, which indicated that the SNP is predicted to be located in regions of open chromatin that are accessible to many transcription factors. Moreover, data from JASPAR database shows that rs1015275 (G>C) is predicted to alter the HAND1 binding motif, with the C allele increasing the likelihood of HAND1 binding compared to the G allele.

In addition, JASPAR database also shows that HAND1 can complex with TCF3 and TCF4. HAND1, TCF3, and TCF4 are transcription factors of the basic helix-loop-helix protein (bHLH) family, which bind to a consensus sequence, CAnnTG, that resides in *cis*-regulatory elements of downstream target genes [52]. Transcription factor interplay is intrinsically related to enhancer function [53], which might indicate higher *TGFB2* expression in patients with the C allele of rs1015275 (corresponding to the G allele of rs11118119), leading to higher circulating levels of TGF-*β*2.

The results of the present investigation validate one of our previous findings that identified rs7767396 as a *cis*-pQTL for circulating VEGF-A in patients with locally advanced pancreatic cancer from CALGB 80303 and in CRC patients in CALGB 80303 [13]. From the previous study, it is already known that the binding of NF-AT1 and ZBRK1 transcription factors may be altered by the presence of the G allele of rs7767396 (A>G), which can regulate VEGF-A plasma levels. Moreover, rs7767396, and SNPs in high LD with it (R^2>0.95, rs78355601, rs4513773, rs11757903), have been previously associated with VEGF-A plasma levels in several studies reported in the NHGRI-EBI genome-wide association studies (GWAS) catalog [54–60].

The results of this study also validated previously reported *trans*-pQTLs for VEGF-R3 on chromosomes 3 and 9. On chromosome 3, rs10935473 (C>A) has already been associated with plasma levels of VEGF-R3 in previous studies in patients with pre-diabetes or diabetes reported in the pGWAS database [61] and other studies reported in the NHGRI-EBI GWAS catalog [54]. Similar to our study, the A allele of rs10935473 (C>A) was associated with decreased levels of VEGF-R3. On chromosome 9, rs507666 (G>A) has also been associated with plasma levels of VEGF-R3 in a previous study in patients with pre-diabetes or diabetes reported in the pGWAS database [61]. Similar to our study, the A allele of rs507666 (G>A) was associated with lower levels of VEGF-R3.

Lastly, our analysis identified rs2229523 (G>A) as a *cis*-pQTL for CD73, with the A allele of rs2229523 in *NT5E* associated with higher plasma levels of CD73. The G allele of rs2229523 (G>A) was already reported as an eQTL decreasing the mRNA expression of *NT5E* in whole blood (p=1.1×10^{-5}, normalized effect size NES = -0.16) and many other tissues [62]. However, this is the first study reporting rs2229523 as a pQTL for the circulatory protein levels of CD73 in plasma.

This study has some limitations. The discovery process was limited to genetically estimated Europeans. The reported association between rs11118119 and TGF-*β*2 observed in CALGB/SWOG 80405 (advanced mCRC) and validated in CALGB 90401 (advanced prostate

cancer) failed to validate in CALGB 80303 (advanced pancreatic cancer). We note that the TGF-β2 assay used in CALGB/SWOG 80405 and CALGB 90401 was an improved version of the assay initially used in CALGB 80303. The first-generation TGF-β2 assay did not have as wide a dynamic range or level of sensitivity as the current TGF-β2 assay. Further, the TGF-β2 assay used in CALGB 80303 had much lower precision, exhibiting a coefficient of variation (CV) of 15.2% compared to 6.0% and 3.8% observed in CALGB/SWOG 80405 and CALGB 90401, respectively. The present analysis has been restricted to high quality typed variants at the genotype level. Imputation- and haplotype-based analyses may identify additional relevant sources of genetic variation. The mechanism proposed of how rs11118119 regulates the levels of TGF-β2 by bioinformatic analysis needs to be further validated in experimental models. Finally, the study results do not establish the link between this variant, the circulating markers, and clinically relevant outcomes, and do not consider potential for co-localization with other disease-trait loci.

Conclusions

In summary, this study has provided evidence of a novel *cis* germline genetic variant that regulates circulating TGF-β2 levels in plasma of patients with advanced CRC and prostate cancer. The putative reference relative allelic frequency for this variant ranges from 0.14 in the European population to over 0.5 in the African population. The discovery of a genetic variant that regulates the levels of TGF-β2 in circulation might have important implications for identification of prognostic biomarkers and mechanisms that shape disease heterogeneity in advanced tumors.

Abbreviations

AFR	African cohort from the 1000 Genomes database
CALGB	Cancer and Leukemia Group B
VEGF	Vascular endothelial growth factor
CONSORT	Consolidated Standards of Reporting Trials
CV	Coefficient of variation
EGFR	Epidermal growth factor receptor
ELISA	Enzyme-linked immunosorbent assay
EUR	European cohort from the 1000 Genomes database
GWAS	Genome-wide association study
AAF	Alternate allele frequency
mCRC	Metastatic colorectal cancer
pQTL	Protein quantitative trait loci
QC	Quality control
QQ	Quantile-quantile
SNP	Single nucleotide polymorphism
SWOG	Southwest Oncology Group

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10354-7.

Supplementary Material 1

Supplementary Material 2

Data and code availability

The code base to reproduce the statistical and replication analyses presented in this paper is available from a public source code repository (https://gitlab.oit.duke.edu/dcibioinformatics/pubs/calgb-80405-pqtl). The genotype and phenotype (clinical and protein) data for the CALGB/SWOG 80405 discovery cohort are available from the database of Genotypes and Phenotypes (dbGaP) through study accession: phs003428.v1.p1. The data for the CALGB 80303 and CALGB 90401 validation cohorts are available from the database of Genotypes and Phenotypes (dbGaP) through study accessions phs000250.v1.p1 and phs001002.v2.p1, respectively. The TGF-β2 protein marker for the CALGB 80303 validation cohort is available as supplementary material in Innocenti et al. [13].

Declarations

Ethics approval and consent to participate

The analyses were conducted using clinical, protein marker and germline DNA genotyping data from patients registered to CALGB/SWOG 80405 (registered under National Clinical Trial number NCT00265850) who had consented to participation in Institutional Review Board (IRB) approved pharmacogenomic substudy (CALGB 60501). The germline DNA and clinical data used in the validation analyses were obtained from patients from CALGB 90401 (NCT00110214) and CALGB 80303 (NCT00088894) who had provided consent to research under IRB approved pharmacogenomic substudies (CALGB 60404 and CALGB 60401 respectively). CALGB/SWOG 80405, CALGB 90401 and CALGB 80303 were multi-center clinical trials. Each of these three substudies were approved by the Cancer Treatment Evaluation Program (CTEP) of the National Cancer Institute (NCI) and the Institutional Review Board (IRB) of each participating center. The analyses presented in this paper were conducted

under protocol Pro00113199 approved by the Duke Health IRB of Duke University.

Consent for publication
Not applicable.

Author details
[1]UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Duke Cancer Institute, Duke University Medical Center, Durham, NC, USA
[3]Department of Medicine, Duke University Medical Center, Durham, NC, USA
[4]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA
[5]Alliance Statistics and Data Management Center, Duke University, Durham, NC, USA
[6]Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, IN, USA
[7]Department of Medicine, The University of Chicago, Chicago, IL, USA
[8]Department of Medical Oncology, Sidney Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA, USA
[9]Department of Medicine, University of California at San Francisco, San Francisco, CA, USA
[10]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[11]Present address: Utah Tech University, St George, UT, USA
[12]Division of Pharmacotherapy and Experimental Therapeutics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

References
1. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discovery. 2013;12(8):581–94.
2. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. Nat Rev Genet. 2021;22(1):19–37.
3. American Cancer Society. Key Statistics for Colorectal Cancer https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html2021 [.
4. Ward DG, Suggett N, Cheng Y, Wei W, Johnson H, Billingham LJ, et al. Identification of serum biomarkers for colon cancer by proteomic analysis. Br J Cancer. 2006;94(12):1898–905.
5. Engwegen JY, Helgason HH, Cats A, Harris N, Bonfrer JM, Schellens JH, Beijnen JH. Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry. World J Gastroenterol. 2006;12(10):1536–44.
6. Surinova S, Radová L, Choi M, Srovnal J, Brenner H, Vitek O, et al. Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. EMBO Mol Med. 2015;7(9):1153–65.
7. Liu Y, Lyu J, Bell Burdett K, Sibley AB, Hatch AJ, Starr MD, et al. Prognostic and predictive biomarkers in patients with metastatic colorectal Cancer receiving Regorafenib. Mol Cancer Ther. 2020;19(10):2146–54.
8. Venook AP, Niedzwiecki D, Lenz HJ, Innocenti F, Fruth B, Meyerhardt JA, et al. Effect of First-Line Chemotherapy Combined with Cetuximab or Bevacizumab on overall survival in patients with KRAS Wild-Type Advanced or metastatic colorectal Cancer: a Randomized Clinical Trial. JAMA. 2017;317(23):2392–401.
9. Kelly WK, Halabi S, Carducci M, George D, Mahoney JF, Stadler WM, et al. Randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer: CALGB 90401. J Clin Oncol. 2012;30(13):1534–40.
10. Hertz DL, Owzar K, Lessans S, Wing C, Jiang C, Kelly WK, et al. Pharmacogenetic Discovery in CALGB (Alliance) 90401 and mechanistic validation of a VAC14 polymorphism that increases risk of Docetaxel-Induced Neuropathy. Clin Cancer Res. 2016;22(19):4890–900.
11. Kindler HL, Niedzwiecki D, Hollis D, Sutherland S, Schrag D, Hurwitz H, et al. Gemcitabine plus Bevacizumab compared with gemcitabine plus placebo in patients with advanced pancreatic cancer: phase III trial of the Cancer and Leukemia Group B (CALGB 80303). J Clin Oncol. 2010;28(22):3617–22.
12. Innocenti F, Owzar K, Cox NL, Evans P, Kubo M, Zembutsu H, et al. A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in CALGB 80303. Clin Cancer Res. 2012;18(2):577–84.
13. Innocenti F, Jiang C, Sibley AB, Etheridge AS, Hatch AJ, Denning S, et al. Genetic variation determines VEGF-A plasma levels in cancer patients. Sci Rep. 2018;8(1):16332.
14. Innocenti F, Sibley AB, Patil SA, Etheridge AS, Jiang C, Ou FS, et al. Genomic Analysis of Germline Variation Associated with Survival of patients with Colorectal Cancer treated with Chemotherapy Plus Biologics in CALGB/SWOG 80405 (Alliance). Clin Cancer Res. 2021;27(1):267–75.
15. Nixon AB, Pang H, Starr MD, Friedman PN, Bertagnolli MM, Kindler HL, et al. Prognostic and predictive blood-based biomarkers in patients with advanced pancreatic cancer: results from CALGB80303 (Alliance). Clin Cancer Res. 2013;19(24):6957–66.
16. Quintanilha JCF, Liu Y, Etheridge AS, Yazdani A, Kindler HL, Kelly WK et al. Plasma levels of angiopoietin-2, VEGF-A, and VCAM-1 as markers of bevacizumab-induced hypertension: CALGB 80303 and 90401 (Alliance). Angiogenesis. 2021.
17. Nixon AB, Sibley AB, Liu Y, Hatch AJ, Jiang C, Mulkey F, et al. Plasma protein biomarkers in Advanced or metastatic colorectal Cancer patients receiving Chemotherapy with Bevacizumab or Cetuximab: results from CALGB 80405 (Alliance). Clin Cancer Res. 2022;28(13):2779–88.
18. Terpstra T. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Indagationes Math. 1952;55:327–33.
19. Jonckheere AR. A distribution-free k-Sample Test Against ordered Alternatives. Biometrika. 1954;41(1/2):133–45.
20. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. New York: John Wiley & Sons, Inc.; 2014.
21. Lin J, Sibley A, Shterev I, Nixon A, Innocenti F, Chan C, Owzar K. fastJT: an R package for robust and efficient feature selection for machine learning and genome-wide association studies. BMC Bioinformatics. 2019;20(1):333.
22. Kloke J, McKean J, Rfit. Rank-based Estimation for Linear models. R J. 2012;4:57–64.
23. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326–8.
24. R Core Team. R: a Language and Environment for Statistical Computing. https://wwwR-projectorg/; 2022.
25. Wickham H, Averick M, Bryan 1 J, Chang W, McGowan LDA, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4(43):1686.
26. RevolutionAnalytic WS. foreach: Foreach looping construct for R. R Packag Version 2015.
27. Zheng X, Gogarten SM, Lawrence M, Stilp A, Conomos MP, Weir BS, et al. SeqArray-a storage-efficient high-performance data format for WGS variant calls. Bioinformatics. 2017;33(15):2251–7.
28. Zhu H. Create awesome HTML table with knitr::kable and kableExtra. CRAN Repos; 2018.
29. Xie Y, Documentation. 2012;8.
30. Yihui X. Package Rmarkdown. Title Dynamic Documents for R. CRAN Repos; 2019.
31. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336–7.
32. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. Nucleic Acids Res. 2021;49(D1):D916–23.
33. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790–7.
34. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.

35. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 2016;44(D1):D877–81.

36. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic Acids Res. 2018;46(W1):W109–13.

37. Zuo C, Shin S, Keleş S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics. 2015;31(20):3353–5.

38. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996;276(8):637–9.

39. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.

40. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):D91–4.

41. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214–20.

42. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4.

43. Massagué J, Blain SW, Lo RS. TGFbeta signaling in growth control, cancer, and heritable disorders. Cell. 2000;103(2):295–309.

44. Xu Y, Pasche B. TGF-beta signaling alterations and susceptibility to colorectal cancer. Hum Mol Genet. 2007;16(Spec 1):R14–20.

45. Hoosein NM, McKnight MK, Levine AE, Mulder KM, Childress KE, Brattain DE, Brattain MG. Differential sensitivity of subclasses of human colon carcinoma cell lines to the growth inhibitory effects of transforming growth factor-beta 1. Exp Cell Res. 1989;181(2):442–53.

46. Cao Z, Kyprianou N. Mechanisms navigating the TGF-β pathway in prostate cancer. Asian J Urol. 2015;2(1):11–8.

47. Principe DR, DeCant B, Mascariñas E, Wayne EA, Diaz AM, Akagi N, et al. TGFβ signaling in the pancreatic Tumor Microenvironment promotes fibrosis and Immune Evasion to Facilitate Tumorigenesis. Cancer Res. 2016;76(9):2525–39.

48. Ma GF, Miao Q, Zeng XQ, Luo TC, Ma LL, Liu YM, et al. Transforming growth factor-β1 and -β2 in gastric precancer and cancer and roles in tumor-cell interactions with peripheral blood mononuclear cells in vitro. PLoS ONE. 2013;8(1):e54249.

49. Konrad L, Scheiber JA, Schwarz L, Schrader AJ, Hofmann R. TGF-beta1 and TGF-beta2 strongly enhance the secretion of plasminogen activator inhibitor-1 and matrix metalloproteinase-9 of the human prostate cancer cell line PC-3. Regul Pept. 2009;155(1–3):28–32.

50. Tu Y, Han J, Dong Q, Chai R, Li N, Lu Q, et al. TGF-β2 is a prognostic biomarker correlated with Immune Cell Infiltration in Colorectal Cancer: a STROBE-compliant article. Med (Baltim). 2020;99(46):e23024.

51. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.

52. Murre C, McCaw PS, Vaessin H, Caudy M, Jan LY, Jan YN, et al. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell. 1989;58(3):537–44.

53. Reiter F, Wienerroither S, Stark A. Combinatorial function of transcription factors and cofactors. Curr Opin Genet Dev. 2017;43:73–81.

54. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–12.

55. Sun W, Kechris K, Jacobson S, Drummond MB, Hawkins GA, Yang J, et al. Common genetic polymorphisms influence blood biomarker measurements in COPD. PLoS Genet. 2016;12(8):e1006011.

56. Nath AP, Ritchie SC, Grinberg NF, Tang HH, Huang QQ, Teo SM, et al. Multivariate Genome-Wide Association Analysis of a Cytokine Network Reveals Variants with widespread Immune, Haematological, and Cardiometabolic Pleiotropy. Am J Hum Genet. 2019;105(6):1076–90.

57. Sliz E, Kalaoja M, Ahola-Olli A, Raitakari O, Perola M, Salomaa V, et al. Genome-wide association study identifies seven novel loci associating with circulating cytokines and cell adhesion molecules in finns. J Med Genet. 2019;56(9):607–16.

58. Maffioletti E, Gennarelli M, Magri C, Bocchio-Chiavetto L, Bortolomasi M, Bonvicini C, et al. Genetic determinants of circulating VEGF levels in major depressive disorder and electroconvulsive therapy response. Drug Dev Res. 2020;81(5):593–9.

59. Debette S, Visvikis-Siest S, Chen MH, Ndiaye NC, Song C, Destefano A, et al. Identification of cis- and trans-acting genetic variants explaining up to half the variation in circulating vascular endothelial growth factor levels. Circ Res. 2011;109(5):554–63.

60. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73–9.

61. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.

62. The GTEx Consortium et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60. https://doi.org/10.1126/science.1262110.

Publisher's Note