The non-coding variant rs1800734 enhances DCLK3 expression through long-range interaction and promotes colorectal cancer progression

Ning Qing Liu1,†, Menno ter Huurne1, Luan N. Nguyen1, Tianran Peng1, Shuang-Yin Wang1, James B. Studd2, Onkar Joshi1, Halit Ongen3, Jesper B. Bramsen4, Jian Yan5,6, Claus L. Andersen4, Jussi Taipale5, Emmanouil T. Dermitzakis3, Richard S. Houlston2, Nina C. Hubner1 & Hendrik G. Stunnenberg1

Genome-wide association studies have identified a great number of non-coding risk variants for colorectal cancer (CRC). To date, the majority of these variants have not been functionally studied. Identification of allele-specific transcription factor (TF) binding is of great importance to understand regulatory consequences of such variants. A recently developed proteome-wide analysis of disease-associated SNPs (PWAS) enables identification of TF-DNA interactions in an unbiased manner. Here we perform a large-scale PWAS study to comprehensively characterize TF-binding landscape that is associated with CRC, which identifies 731 allele-specific TF binding at 116 CRC risk loci. This screen identifies the A-allele of rs1800734 within the promoter region of MLH1 as perturbing the binding of TFAP4 and consequently increasing DCLK3 expression through a long-range interaction, which promotes cancer malignancy through enhancing expression of the genes related to epithelial-to-mesenchymal transition.
A n individual’s risk to develop colorectal cancer (CRC) is affected by a broad spectrum of genetic variants that abolish the functions and/or alter the expression of target genes. In CRC, two types of genetic variants have been extensively discussed to contribute to disease onset and progression: (1) protein-coding mutations and (2) non-coding variants, in particular in DNA regulatory elements. To date, the majority of studies have focused on protein-coding mutations. Key coding mutations such as APC/CTNNB1, KRAS/BRAF, PIK3CA, TP53 and SMAD4 have been intensively characterized1. However, even though a great number of non-coding risk variants for CRC have been identified in genome-wide association studies (GWAS)2–9, their molecular functions have rarely been determined.

Functional genetic variants in distal DNA regulatory elements may alter transcription networks in several ways such as by affecting transcription factor (TF) binding. Single-nucleotide polymorphisms (SNPs) within TF-binding sites may affect the local chromatin accessibility10–12 and/or alter the expression of gene targets through mediating different chromatin interactions13–15. Therefore, identification of variant-specific TF interactors is of great importance to understand regulatory consequences of the variants. However, DNA oriented methods such as DNase I sequencing (DNase I-seq)16, systematic evolution of ligands by exponential enrichment sequencing17 and chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq)18 are biased by DNA motif knowledge or limited by the availability of antibodies, resulting in a biased identification of TF-binding dynamics at disease-associated loci. A recently developed proteome-wide analysis of disease-associated SNPs (PWAS) enables to identify DNA–TF interactions in an unbiased manner19. A similar approach has been used to characterize a protein–DNA interaction map for ultra-conserved elements20. Therefore, we performed a large-scale PWAS study to comprehensively understand TF-binding landscape related to CRC.

As the outcome from our PWAS screen, we further investigated the functions of a SNP located in the promoter region of MLH1 gene (MLH1-93G>A or rs1800734), which has been associated with the risk of several cancer types including CRC9,21,22, endometrial cancer23, glioblastoma24 and lung cancer25. Here we identified a molecular function of this SNP in promoting cancer malignancy through a novel gene target named DCLK3.

**Results**

PWAS identified TF occupancy switching at the 116 loci. We selected 116 SNPs associated with CRC risk2–9 for PWAS analysis including the following: (1) typed and imputed GWAS significant SNPs (for imputed SNPs, linkage disequilibrium (LD) $r^2 \geq 0.2$) from 8q24.21 (MYC-335), 15q13.3 (GREM1) and 18q21.1 (SMAD7); (2) SNPs with functional evidence rs16969681/15q13.3 (ref. 26), rs58920878/18q21.1 (ref. 27), rs16888589/18q23.3 (ref. 28) and rs4444235/14q22.2 (ref. 29); (3) 3 SNPs from 3p22.2 (MLH1 region)20; and (4) GWAS significant SNPs in DNSase I-seq peaks in minimal one of 15 fetal large intestine regions and 12 CRC cell lines (Supplementary Data 1). The PWAS analysis identified 731 TF-binding alterations between reference (Ref) and alternate (Alt) alleles ($P < 0.01$, A/B significance test) (Supplementary Data 2). Compared with proteome data, TF–DNA interactome data showed a clear enrichment for TFs (Supplementary Fig. 1a) and the altered binding events mediated by known TFs (Supplementary Data 3) showed stronger allele preference than other interactions ($P = 4.4 \times 10^{-6}$, Mann–Whitney U-test) (Supplementary Fig. 1b). Overlay of the pulldowns showed a consistent allele preference between two replicate experiments (Supplementary Fig. 1c). Many of the 731 TFs showed >8-fold affinity to one of the alleles at these loci (Fig. 1a), for example, TFAP4 at rs1800734 and RUNX1/RUNX2/CBFAP at rs1741640. As expected, top pathways associated with the 731 TFs included key CRC drivers such as WNT and transforming growth factor-$\beta$ pathways (Fig. 1b).

It is well-known that DNA hypersensitive sites coincide with regulatory elements and are the hotspots for TF binding8. To better predict in vivo TF binding, it is necessary to consider the hypersensitivity of tested regions. Therefore, we ranked all the SNP–TF interactions based on PWAS fold change (Ref/Alt allele) and DNAse I hypersensitivity (DHS) of the SNP loci. A total of 27 significant allele-specific SNP–TF binding events were considered to be important (Fig. 1c and Supplementary Table 1). Many selected TF–SNP interactions (Fig. 1d and Supplementary Fig. 1d) were validated using ChIP-seq data (Supplementary Fig. 1e). Based on this selection, the top candidate SNP is rs1800734 (MLH1-93G>A) in the 5' untranslated region (UTR) region of MLH1 gene. Therefore, we decided to focus on this SNP.

**PWAS identified specific interactors of rs1800734.** Our PWAS screen identified TFAP4 as an allele-specific interactor with an almost 16-fold higher affinity for G-allele, whereas ELF1 showed higher affinity for the A-allele of rs1800734 in LoVo cells (Fig. 1d). The A-allele abolishes TFAP4 binding due to a point mutation at the last position of the E-box (Fig. 2a), which simultaneously creates an E26 transformation-specific (ETS) family binding motif (Supplementary Fig. 2a). We corroborated and extended this finding by label-free-based DNA pulldown with SNU175 and COLO320 extracts (Fig. 2b and supplementary Fig. 2b). Notably, other basic helix-loop-helix (bHLH) and ETS family members also displayed allele-specific binding, including MYC and ELF2 (Supplementary Fig. 2b,c), indicating that these TFs can compete with TFAP4 and ELF1 at this locus. Using ChIP-seq as an orthogonal technique, we validated TFAP4 and ELF1 binding at the rs1800734 in the SNU175 and COLO320, cell lines heterozygous for this locus. Consistent with PWAS, TFAP4 ChIP-seq showed a dominant preference for G-allele binding in the two cell lines (Fig. 2c). ELF1 did not show significant allele-binding preference (Supplementary Fig. 2d), which may be due to the competitive binding interference by other ETS proteins.

**DCLK3 is a novel target of rs1800734.** Given the position of the SNP in the promoter of MLH1, we further investigated whether predisposition of rs1800734 in CRC is due to DNA methylation of MLH1 promoter as proposed9,30. We tested hypersensitivity and transcription of G- and A-allele in the two heterozygous cell lines, which showed the two alleles are equally accessible and transcribed (Fig. 2d). The neighbouring gene EPM2A1P1, also reported to be regulated by rs1800734 (ref. 31), was similarly unaffected by rs1800734 (Fig. 2d). Hence, we conclude that rs1800734 does not result in allele-specific epigenetic silencing of either MLH1 or EPM2A1P1 in these cell lines. We sought to confirm our observation in the Systems Biology of Colorectal Cancer (SYSCOL) cohort of paired healthy and tumour tissues (healthy: $n = 236$, tumour: $= 289$). A strong expression quantitative trait loci (eQTL) between rs1800734 and MLH1 was observed in the healthy tissues ($P = 0.001$, linear regression model) but the A-allele was associated with increased MLH1 expression. This eQTL was lost in the tumours and remained only a weak association in microsatellite stable (MSS) tumours ($P = 0.025$, $n = 236$, linear regression model) (Supplementary Fig. 3a). Intriguingly, we identified a correlation between rs1800734 and DCLK3 expression in the healthy ($P = 0.029$, linear regression model)
Figure 1 | PWAS screen systematically identified allele-specific TF binding at the selected CRC risk loci. (a) Allele-specific binding of the 731 candidate TFs at the 116 CRC risk loci. Chromatin environment of the 116 SNPs was described by DHS of these loci in the 15 fetal large intestine tissues and 12 CRC cell lines, and histone modifications (H3K4me3, H3K4me1 and H3K27ac) within ±1 kb regions around these loci in the LoVo and HCT116 cell lines. The TFs with $P$-value $<10^{-30}$ and absolute $\log_2$(fold change) > 3 was listed in the bubble plot. Bubble size represents $-\log_{10}(P$-values) of the interactors in the pull-down screen ($n=2$ pulldowns per SNP, $P$-values: A/B significance test). (b) Pathway annotation of the 731 TFs (HR, gonadotropin-releasing hormone receptor pathway; WNT/PDGF/CCKR/transforming growth factor (TGF)-β, Wnt/PDGF/Cholecystokinin/TGF-β signalling pathways). (c) TF–SNP interactions ranked by fold changes in the PWAS screen and DHS at the SNP loci. Bubble size indicates the $-\log_{10}(P$-values) of the TF–SNP interactions ($n=2$ pulldowns per SNP, red and blue bubbles: $P$-value $<0.05$, Z-test). (d) The top three candidate TF–SNP interactions ($n=2$ pulldowns per SNP, red and blue dots: $P$-values $<0.01$, A/B significance test).
Importantly, the eQTL association between rs1800734 and identical in the three isogenic lines (Supplementary Fig. 3c). Confirmed higher TFAP4 binding to the G-allele (Supplementary).

Sanger sequencing (Fig. 3b). No other mutation was observed in (Fig. 3a). Successful targeting of rs1800734 was confirmed by

were generated from COLO320 using CRISPR-CAS9 technique (Fig. 3a). Successful targeting of rs1800734 was confirmed by

expression, two isogenic cell lines homozygous for G- or A-allele were generated from COLO320 using CRISPR-CAS9 technique (Fig. 3a). Successful targeting of rs1800734 was confirmed by Sanger sequencing (Fig. 3b). No other mutation was observed in the surrounding region. ChIP–quantitative PCR (qPCR) confirmed higher TFAP4 binding to the G-allele (Supplementary Fig. 3b) and the level of MLH1 and EPM2AIP1 transcription was identical in the three isogenic lines (Supplementary Fig. 3c).

Importantly, the eQTL association between rs1800734 and DCLK3 expression was replicated and a fivefold difference in DCLK3 expression was observed between G- and A-homozygotes (Fig. 2e). By sequencing the inter-exonic reverse transcriptase–cDNA products of DCLK3, we confirmed the transcription of DCLK3 in these cell lines (Supplementary Fig. 3d).

Figure 2 | Identification of allele-specific interactors for rs1800734 and its potential gene targets. (a) Motif analysis interpreted that A-allele of the loci perturbs E-box motif. (b) The specific binding of TFAP4 (P1 = 2.0 × 10−6, P2 = 3.1 × 10−5, P3 = 7.8 × 10−5, Student’s t-test) and ELF1 (P1 = 2.1 × 10−4, P2 = 0.011, P3 = 6.9 × 10−4, Student’s t-test) on G- and A-allele was observed in the three different cell lines (data are represented as mean and error bars indicate s.d., n = 3 pulldowns per SNP). ZFP62 binding on A-allele, as shown in Fig. 1d, was not consistent in all the three cell lines and was hence not considered as the general TF regulator at this locus. (c) This allele-specific binding was validated by ChIP-seq of TFAP4. This SNP did not change (d) the local chromatin accessibility and the expression of two cis-regulated genes (MLH1 and EPM2AIP1) (G (G-allele) and A (A-allele) of rs1800734, wt (wild-type allele) and del (deletion allele) of rs35149869, P-values: Fisher’s exact test). (e) eQTL analysis revealed a novel gene targets DCLK3 of this SNP (P-values were calculated in FastQTL software based on linear regression model).

and tumour (P = 0.031, linear regression model) samples, and this association was highly significant in the MSS patients (P = 0.004, linear regression model) (Fig. 2e), indicating this locus may act as a distal enhancer and regulate DCLK3.

The A-allele positively regulates transcription of DCLK3. To establish the functional relation between rs1800734 and DCLK3 expression, two isogenic cell lines homozygous for G- or A-allele were generated from COLO320 using CRISPR-CAS9 technique (Fig. 3a). Successful targeting of rs1800734 was confirmed by Sanger sequencing (Fig. 3b). No other mutation was observed in the surrounding region. ChIP–quantitative PCR (qPCR) confirmed higher TFAP4 binding to the G-allele (Supplementary Fig. 3b) and the level of MLH1 and EPM2AIP1 transcription was identical in the three isogenic lines (Supplementary Fig. 3c). Importantly, the eQTL association between rs1800734 and DCLK3 expression was replicated and a fivefold difference in DCLK3 expression was observed between G- and A-homozygotes (Fig. 3c). By sequencing the inter-exonic reverse transcriptase–cDNA products of DCLK3, we confirmed the transcription of DCLK3 in these cell lines (Supplementary Fig. 3d).

rs1800734 regulates DCLK3 through long-range interactions. A capture Hi-C study has suggested that rs3806624 in the promoter of EOMES affects AZI2, a gene 640 kbp downstream to this SNP, through a long-range chromatin interaction14. A similar constellation may apply to rs1800734. Therefore, Circularized Chromosome Conformation Capture with massively parallel DNA sequencing (4C-seq) was employed to search for long-range interactions between rs1800734 and other potential targets in the three isogenic cell lines. Using rs1800734 as a view point, a significantly enhanced interaction was observed in A-homozygote
with DCLK3 region (Fig. 3d). Increased interactions were found in the promoter and 3'-UTR region of DCLK3 in two independent experiments (Fig. 3e) and, in addition, the 3'-UTR interaction appeared to increase chromatin accessibility (Fig. 3f). In conclusion, the A-allele of rs1800734 increases the DCLK3 transcription through increased chromatin interaction and enhanced chromatin accessibility.

DCLK3 is a potential oncogenic and tumour progressive factor.

DCLK3 is one of three doublecortin-like kinases (DCLK1, DCLK2 and DCLK3). In this family, DCLK1 has been shown to be a cancer stem cell marker in intestinal tumours. The molecular function of the DCLK3 has not been characterized in depth. We therefore performed gene set enrichment analysis (GSEA) to identify DCLK3-associated gene sets in the SYSCOL RNA sequencing (RNA-seq) cohort. Interestingly, we found that epithelial-to-mesenchymal transition (EMT)-related genes were highly correlated to the expression of DCLK3 in the healthy tissues (normalized enrichment score (NES) = 2.50), which was enhanced in the tumour samples (NES = 3.10) (Fig. 4a and Supplementary Table 2). Using a cutoff of NES at 2.00, we also identified gene sets that are preferentially enriched in the healthy versus tumour tissues (Supplementary Table 2), for example ‘MYC targets’ and ‘G2/M checkpoint’ gene sets were enriched in the healthy tissue, whereas ‘Angiogenensis’ and ‘tumor necrosis factor alpha (TNFA) signalling via nuclear factor–κB’ gene sets were enriched in the tumour tissues (Fig. 4a). As expected, common EMT markers such as CALD1, FN1, SNAI1, SNAI2, TWIST1, VIM, ZEB1 and ZEB2 were highly significantly co-expressed with DCLK3 in the tumours (Fig. 4a), indicating that DCLK3 is an EMT regulator. Furthermore, we observed elevated DCLK3 expression in the tumour compared with healthy tissues (P = 2.2 × 10^-16, Mann–Whitney U-test) (Fig. 4b), but no difference between the microsatellite instable (MSI) and microsatellite stable (MSS) tumours (Fig. 4b). Interestingly, this elevation appeared to be correlated with the CRC progression: DCLK3 expression was at a comparable level in the healthy and adenoma tissues, and increased 2-fold in the stage I tumours and remained at this level during tumour malignancy (P = 2.2 × 10^-16, Kruskal–Wallis test) (Fig. 4b). Therefore, DCLK3 may promote EMT events and consequently drive tumour malignancy.
Discussion

Our study generated the first TF–SNP interaction map at presumed disease-relevant loci of CRC and determined TF-binding occupancy at the 116 upmost relevant CRC risk loci. Together with GWAS and epigenetic profiling data, our PWAS screens provide a comprehensive TF-binding landscape of these loci and yielded candidate interactions for further functional investigations (Fig. 1a,c,d and Supplementary Fig. 1d,e). As an
alternative approach to investigate allele-specific TF-binding events, the PWAS approach shows several advantages over the DNA-centric methods. In contrast to ChIP-seq technique, PWAS-based identification is a hypothesis-free approach, which does not require any knowledge on the possible binding TFs at risk loci. In addition, application of PWAS approach is not limited by the availability of high-quality ChIP-grade antibodies. Even though large ChIP-seq data sets have been generated, these data nevertheless cover only a small proportion of the entire repertoire of TFs. Other DNA-centric methods, such as DNase I-seq or assay for transposase-accessible chromatin with high throughput sequencing (ATAC-seq), are only capable of identifying DNA motifs present at risk loci, which makes it difficult to predict actual binding TF(s) from a family sharing similar motifs in a specific cell type. The outcome of PWAS is a reflection of TF abundance, TF ability to bind to the sequence (affinity), as well as synergistic and antagonistic effects due to binding of other TFs to adjacent or overlapping sequences. Therefore, actual binding TFs of a given cell type can be predicted using PWAS method. In addition, PWAS approach allows to identify TFs at the loci only partially matching consensus motif sequences and therefore cannot be predicted by motif prediction-based methods.

Based on the PWAS screen, we investigated in great depths for a high-confidence CR-association SNP, MLH1-93G>A or rs1800734. It has been postulated that A-allele (risk allele) of MLH1 in a given cell type can be predicted using PWAS method. In adjacent or overlapping sequences. Therefore, actual binding TFs from a family sharing similar motifs in a specific cell type. The outcome of PWAS is a reflection of TF abundance, TF ability to bind to the sequence (affinity), as well as synergistic and antagonist effects due to binding of other TFs to adjacent or overlapping sequences. Therefore, actual binding TFs in a given cell type can be predicted using PWAS method. In addition, PWAS approach allows to identify binding TFs at the loci only partially matching consensus motif sequences and therefore cannot be predicted by motif prediction-based methods.

Methods

Cell culture and extraction of nuclear soluble fraction. Human CRC cell lines were cultured in DMEM or RPMI medium supplemented with 10% fetal bovine, 100 U ml⁻¹ penicillin and 100 μg ml⁻¹ streptomycin. LoVo and SNU175 cell lines were purchased from American Type Culture Collection and Korean Cell Line Bank (KCLB), respectively. Colo320 cell line was a generous gift of Dr Riccardo Fodde (Erasmus Medical Center Rotterdam, The Netherlands), which was originally from American Type Culture Collection. The authenticity of the cell lines were confirmed using microsatellite short tandem repeat (STR) assay by the suppliers of the cell lines. Mycoplasma infection was routinely tested in-house, to ensure that all the cell lines used for this study were free of mycoplasma contamination. Nuclear soluble fraction of LoVo, SNU175 and Colo320 were performed using a published protocol. Protein concentration of the obtained nuclear extract was quantified using Bradford assay. Each of the 3 mg extract (12 individual DNA pulldowns) was aliquoted, snap-frozen in liquid nitrogen and stored at −80°C.

High-throughput DNA pulldown. High-throughput DNA pull-down experiments were performed on 96-well filter plate format using our published method with minor modifications. To synthesize biontinated double-stranded DNA (dsDNA) oligo, we attached a non-genomic 15 bp sequence at the 3'-end of the anti-sense strand. Subsequently, a reverse complement biontinated primer was used to extend single-stranded DNA templates into dsDNA oligos. For each synthesis, 100 pmol of the biontinated primer and 200 pmol template were subjected to a PCR reaction using Hercule II Fusion Enzyme kit under the following conditions: 95°C for 3 min, followed by 30 cycles (n = 20) at 95°C for 1 min, 45°C for 1 min and 72°C for 1 min; 72°C for 3 min; infinite hold at 12°C. Briefly, high-throughput DNA pulldown was performed using a Multiscreen filter plate with 1.2 μm pores (Millipore, MBBV1210). The biontinated dsDNA oligos were immobilized on 20 μl of high-performance streptavidin sepharose (GE Healthcare, 17511301). Two-hundred and fifty micrograms of nuclear extracts and 15 μg of competitors (5 μg of poly(deoxyinosinic-deoxycytidylic) acid sodium salt (Sigma-Aldrich, P4929), poly(deoxyadenylc-thymidylc) acid sodium salt (Sigma-Aldrich, P0883) and Bakers yeast RNA (Sigma-Aldrich, R6750)) were added and incubated with immobilized oligos for 1.5 h at 4°C on a plate shaker for each of the pull-down experiments. The components of the competitors were incubated into ~ 300 bp fragments before use. The proteins unbound to DNA oligos were washed off using different washing buffers and the bound TFs were on-bead digested overnight using trypsin/lysC. The pull-down duplicates underwent dimethyl label swapping and measured by nanoscale liquid chromatography tandem mass spectrometry (LC-MS/MS) in a 2 h gradient.

Nuclear extract proteome. Deep proteome profile was generated from the nuclear extracts used for the pull-down experiments. An absolute quantification strategy was taken following a published method. In brief, 3.3 μg of universal protein standard 2 (Sigma-Aldrich) and 10 μg of the nuclear extracts were mixed and subjected to a filter-aided sample preparation digestion. In parallel, 30 μg of the nuclear extracts were also digested using a filter-aided sample digestion protocol and subsequently separated into six fractions using strong anion exchange. The universal protein standard 2 samples and the fraction was purified on C18 stage tips and profiled by nanoscale LC-MS/MS in a 4 h gradient.

DNase I sequencing. DNase I library of the LoVo cell line was constructed following a reported protocol with some minor modifications. In short, 5×10⁶ nuclei were isolated using a buffer (15 mM Tris-HCl pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0 and 0.5 mM spermidine) supplemented with 0.05% Igepal CA-630 detergent. Subsequently, the isolated nuclei were digested with 80 U DNase I (Sigma-Aldrich, D4527) for 3 min and the
digestion was quenched by a step buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 0.1% SDS, 100 mM EDTA pH 8.0, 1 mM Spermidine and 0.5 mM Spermine). A 9% Sucrose gradient was applied to fractionate the samples for 24 h at 25,000 rpm at 16°C and the fractions with < 1 kb fragments were further purified and prepared according to the Illumina library preparation protocol.

DNase I library of other cell lines were prepared using a published protocol18 as described before. The cell lines were harvested under the confluence of 60% and washed with PBS. Nuclei were isolated with RSB lysis buffer (10 mM Tris-Cl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPA-CA-630) at 4°C for 10 min. Then the nuclei were treated with 0.12 unit of DNase I (Roche) in the provided buffer at 37°C for 15 min before being quenched by 50 mM of EDTA. Following RNase A (Sigma) treatment at 37°C for 15 min, proteinase K (NEB) was added for an additional hour at 56°C. DNA was extracted using phenol:chloroform; isoamylalcohol. Agarose Gel (2%) electrophoresis was applied to separate the released fragments (~ 100 bp) that were purified (Qiagen, MinElute Gel Extraction Kit), followed by Illumina TruSeq library preparation and Sequencing (HiSeq2000).

ChIP-seq and ChiP-qPCR analysis. ChiP assays were performed following a standard protocol. Cell lines were cross-linked by a final concentration of 1% paraformaldehyde for 10 min and subsequently cross-linking reaction was quenched using 1.5 M glycine. The harvested cell lines were then lysed and sonicated to obtain ~ 300 bp chromatin using Bioruptor Plus sonication device (Diagenode). The sonicated chromatin was pre-cleaned by Protein A/G magetic beads (ThermoFisher Scientific, 88802) and then incubated together with antibody conjugated beads overnight at 4°C. Antibodies against H3K27ac (Diagenode, C15410186), H3K4me3 (Diagenode, C15410196), H3K4me1 (Diagenode, C15410206), TFAP4 (Santa Cruz Biotechnology, sc-18593X, 6.4 g per ChIP assay) and ELF1 (Santa Cruz Biotechnology, sc-631X, 4.8 g per ChIP assay) were used in our ChiP experiments. Posterior to the incubation, captured chromatin was washed, eluted and de-crosslinked. Resulted DNA fragments were purified and prepared according to Illumina library preparation (H3K27ac ChiP) or KAPA Hyper Prep (TFAP4 and ELF1 ChiP) protocols before sequencing, or directly quantified using SYBR Green-based qPCR assays (Supplementary Table 3).

ATAC-seq analysis. ATAC libraries of the SNU175 and COLO320 cell lines were prepared by a documented protocol48 with some modifications. In brief, nuclei were isolated using a lysis buffer consisting of 10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPA-CA-630 detergent and then tagmented49 with 2 µl of Tn5 transposase and 12.5 µl 2 × TD buffer (Illumina, Nexterra DNA Library Preparation Kit). The resulted DNA fragments underwent two sequential nine-cycle PCR amplification, and in between two PCR reactions the libraries were selected for < 600 bp fragments using SPRI beads. The final PCR products were purified and quantified by KAPA Library Quantification Kits before sequencing.

Targeted RNA-seq and reverse transcriptase-qPCR. Total RNA was isolated from the SNU175 and COLO320 cell lines using a TRizol reagent (ThermoFisher Scientific, 159590618) based method. The yielded RNA was treated using DNase and then reversely transcribed into cDNA using random hexamers (ThermoFisher Scientific, 15596018) based method. The yielded RNA was treated using DNase and then reversely transcribed into cDNA using random hexamers (ThermoFisher Scientific, 15596018) based method. Then, the primer pairs to amplify target pre-defined genomic regions were designed using Perl programs (http://snp.cancer.gov/) and SYBR Green-based qPCR assays (Supplementary Table 3).

CRISPR-CAS9-based SNP editing. CRISPR-CAS9 based SNP editing were performed according to a previously reported method50. Single guide RNAs (sgRNAs) were designed using an online tool (http://crispr.mit.edu/) and double-nicking strategy was taken to reduce undesirable off-target mutagenesis. The sgRNAs were then cloned into a U6-driven plasmid containing a green fluorescent protein (GFP) marker and the D10A mutant Cas9 nickase (Addgene, pSpCas9(9)-BB-2A-GFP, Plasmid 48140). The plasmid was then transformed into the DH5α competent E. coli strain and the products were purified for transfection. Subsequently, 400 ng (400 ng each) and a 199 bp single-stranded oligonucleotides (10 pmol single-stranded oligo donors/ssODNs (ssODNs), possessing G or A point mutations) were transfected into COLO320 cells following a standard lipofectamine LX Reagent protocol (ThermoFisher Scientific, 15338180). FACs analysis was used to sort 192 GFP-positive cells per cell line (G or A point mutations) into 96-well plates in 36 h after the transfection. In parallel, a wild-type cell line was treated in the same manner by without using sgRNAs, and this mockcd cell line was then sorted by FACs and served as the control. The desirable genotype at the rs1800734 was confirmed by Sanger sequencing. The oligos (sgRNAs and Single-Stranded Oligo Donors (ssODNs)) used in these experiments were listed in Supplementary Table 3.

SNP genotyping. SNP genotyping was performed by a standard Sanger sequencing-based method. The regions containing mutations were PCR amplified into ~ 500 bp fragments using specific primers (Supplementary Table 3) and the PCR products were purified using 1.5% agarose gel. A mixture of 10 ng purified PCR products and 6 pmol primers was used for Sanger sequencing and the data were analysed using CodonCode Aligner (V.5.0.2).

4C-seq analysis. The 4C experiments were carried out using a published protocol19 with some modifications. For each assay, 1 × 10⁷ cells were cross-linked and sequenced as in ChiP-seq experiments (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.2% IGEPA-CA-630 detergent and 1 × protease inhibitors). Subsequently, the nuclei were digested with 240 U NaI III enzyme (New England Biolabs Inc., R0215L) followed by a overnight in-nuclei ligation with 4,000 U T4 ligase (New England Biolabs Inc., M0202M) at 16°C. The ligated DNA was cross-linked, purified, digested with 90 U CviQV enzyme (New England Biolabs Inc., R06395) and circularized with 5,000 U T4 ligase. The circularized DNA (16 × 300 ng) was amplified with bait-specific inverse primers (Supplementary Table 3), pooled and purified, followed by KAPA Hyper Prep protocols.

Proteomics data processing. Recorded mass spectrometric (MS) files were analysed by MaxQuant software (version 1.3.5.7)22 using standard settings for dimethyl (N-term) and oxidation (M) for protein identification and with Lys-CAT and Max-Label Swiss-Prot human database (generated from version 06-2012). Batch effects of the pull-down data were removed using 'Combat' algorithm53. In dimethyl analysis, some proteins were consistently quantified in the pulldowns of only one allele (Ref- or Alt-allele) in two replicate experiments and the protein ratios can therefore not be obtained from MaxQuant output. Therefore, we imputed the missing value in these pull-down using ‘Replace missing values from normal distribution’ option in Perseus software (version 1.3.9.18)24, which allows to further calculate protein ratios and perform downstream analysis. TF interactors for each of the pulldown were identified using ‘Significance B’ function in Perseus software.

DNase I-seq ChiP-seq ATAC-seq data processing. Read mapping was performed using BWA-ALN (DNase I-seq and ChIP-seq) and BWA-MEM (ATAC-seq)55 against the hg19 reference human genome. PCR duplicates were removed for further data analysis. Peak calling was carried out using MACS2 (ref. 56) with default settings, except H3K27ac peaks were called using ‘broad’ option. Peaks were called at a q-value cutoff of 0.01. Overlapping peaks were merged for each of the different experiments before further analysis. Integrative Genomics Viewer57 was used to detect bi-allelic differential binding and hypersensitivity.

Targeted RNA-seq and RNA-seq data processing. Targeted RNA-seq data were mapped to the hg19 reference human genome using BWA-MEM55. Integrative Genomics Viewer57 was used to visualize the targeted RNA-seq data and detect bi-allelic differential expression. The SYSCOIL RNA-seq cohort was mapped to the human reference genome sequence (GRCh37 autosomes + X + Y + M) using GEM mapper56. The reads with mapping quality < 150 were excluded for further analysis. The genes were annotated using Ensemble 75. The reads of the genes were counted by ‘HTSeq’ framework59 and normalized using ‘DESeq’ package60.

Calculate significance of the TF-SNP interactions. Importance of the TF–SNP interaction in our screen study was considered by combining DHS of the SNPs (read counts at the SNP position in 15 fetal large intestine tissues and 12 CRC cell lines) and the interaction strength (Log₂(fold change) of a TF between Ref and Alt allele pulldown). The detailed P-value calculation was performed using z-test based on following steps: (1) calculate average DHS (average DNase I-seq reads) of all the 116 loci in the DNase I-seq data from 15 fetal large intestine tissues and 12 CRC cell lines, and then z-score transform the hypersensitivity data of these loci; (2) calculate absolute Log₂(fold change) of every TF–SNP interaction between Ref and Alt alleles, and then z-score transform the absolute Log₂(fold change) of all the interactions; (3) average the z-scores of DHS and absolute Log₂(fold change), and convert the average z-scores into P-values based on normal statistical distribution.

eQTL analysis. The eQTL analysis was performed as described in a previous publication58. Germline genotypes of these patients were genotyped on the Illumina 25M Exome v1.0 and imputed to 1000 genomes phase 3 release using IMPUTE2. For eQTL analysis, we normalized gene quantification separately for healthy and tumor samples. Technical covariates were removed using the PEER program61 and 20 PEER factors were used in normalization. The cis region was defined as ± 1 Mb from the transcription start site for each gene. The associations between genotypes and gene quantification were obtained using the FastQTL software53.

4C-seq data processing. A reduced genome was generated by extracting the sequences flanking the NaIII restriction sites (30 bp on each strand from the NaIII restriction sites to downstream) using the hg19 reference human genome in order to remove the mappability of our 4C-seq data. Subsequently, the mappability of the reduced sequences from each strand was evaluated and the uniquely mappable NaIII restriction sites were kept for downstream analysis. The reads from each library were parsed based on the bait-specific primer sequence and mapped to the reduced hg19 genome using BWA-ALN with the default parameter settings. A Bioconductor package ‘r3Cseq’ with 2.5 kb sliding window was used to determine significant interactions and calculate interaction difference.
Pathway annotation and GSEA analysis. TFs binding differentially to the 116 loci were annotated to PANTHER pathways using PANTHER Classification System.52 GSEA analysis66 was applied to identify the gene sets correlated to DCLK3 expression. The search was performed against hallmark gene sets in Molecular Signatures Database (MsigDB, v5.1).51

Data availability. Raw and processed LC-MS/MS data and sequencing data are available at ProteomeXchange (https://www.ebi.ac.uk/pride/) and Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) under the accession numbers PXD0004435 and GSE89368, respectively. The publically available data used in this study were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) using the following accession numbers: (1) H3K4me1 and H3K4me3 ChIP-seq: GSM1240111, GSM945304, GSM1208810, GSM1208811; (2) DNase I-seq: GSM736493, GSM736600, GSM736560, GSM736587, GSM665815, GSM665818, GSM665826, GSM701490, GSM701495, GSM701513, GSM774213, GSM774214, GSM774241, GSM774242, GSM774273, GSM774323, GSM817186, GSM817188; (3) TF ChIP-seq: GSM1010902, GSM1208683, GSM1208952, GSM1208983, GSM1209042, GSM1209080, GSM1390354, GSM13010847, GSM12098763, GSM13010765, GSM13010790, GSM13010852, GSM1208598, GSM791411, GSM791412, GSM782123, GSM722708, GSM1122302, GSM1122303. The detailed information of in-house generated and public data sets used in this study are listed in Supplementary Table 4. All other remaining data are available within the article and Supplementary Information files, or available from the authors upon request.

References

Acknowledgements
This study was supported by the EU FP7 SYSCOL project. We gratefully thank Dr Xiaofei Zhang for sharing his knowledge on functional experiments, Ms Eva Janssen-Megens and Ms Bowon Kim for their sequencing service, and Dr Riccardo Fodde for providing the wild-type COLO320 cell line.

Author contributions
N.Q.L., N.C.H. and H.G.S. conceived and designed this study. N.Q.L., M.t.H., L.N.N., T.P. and O.J. performed the proteomics, genomics and cell biology experiments. N.Q.L., S.-Y.W., J.B.S., H.O., E.T.D., R.S.H. and N.C.H. carried out statistical and bioinformatics analyses. J.B.B. and C.L.A. provided the SYSCOL RNA-seq data, and J.Y. and J.T. provided some sequencing data of the LoVo cell line for this study. H.G.S. supervised the whole study. N.Q.L. and H.G.S. drafted the manuscript, which was reviewed by all the authors.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Liu, N. Q. et al. The non-coding variant rs1800734 enhances DCLK3 expression through long-range interaction and promotes colorectal cancer progression. Nat. Commun. 8, 14418 doi: 10.1038/ncomms14418 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/}

© The Author(s) 2017