SUPPLEMENTARY METHODS

ChIPseq analysis

ChIPseq was performed on H3K4me1 (C15410194), H3K9me3 (C15410193), H3K27me3 (C15410195) and H3K36me3 (C15410192) for LoVo, and H3K4me1 and H3K9me3 for HT29, using antibodies obtained from Diagenode. Briefly, after cell lysing, sonication of nuclei was performed (UCD-300, BioRuptor) to obtain 150-500bp fragments. ChIP reaction was performed on a Diagenode SX-8G IP-Star Compact using Diagenode automated Ideal Kit reagents (C01010011). Protein A beads were incubated for 10 hours with 3-6µg of antibody and 2-4 million of sonicated cell lysate. ChIP samples were de-crosslinked at 65°C for 4 hours and subsequently treated for 30 minutes with RNAse Cocktail and proteinase K. DNA was then purified (MiniElute PCR purification kit, Qiagen), followed by library preparation according to manufacture (HTP Illumina library preparation kit, KAPA Biosystems). Fourteen cycles of PCR were performed, followed by size selection for 200-400bp fragments and final library purification (GeneRead Size Selection kit, Qiagen). ChIP libraries were sequenced using HiSeq 2000 (Illumina) with 100bp single-ended reads. Generated raw reads were filtered for quality (Phred33 ≥ 30) and length (n ≥ 32), and adapter sequences were removed using Trimmomatic v0.22[1]. Reads passing filters were then aligned to the human reference (hg19) using BWA v0.6.1. Peak calls are obtained using MACS2 v 2.0.10.07132012[2].

Topological associated domains

Topological association domains (TADs) were determined from the Hi-C library. Two independent biological replicates were used for each cell line. Read alignment to GRCh37 and identification of valid di-tags were performed using Bowtie2 v2.2.6[3] and HiCUP v0.5.9[4], respectively. The replicates were merged together, and the interaction matrices were determined with Homer v4.8.3[5] adopting recommended settings. The TADs were identified using the domain calling algorithm described by Dixon *et al.*[6], which is based on directionality index, and used the default bin size of 40kb and window size of 2Mb. Median TAD size was 600kb for HT29 and 840kb for LoVo. For each cell line, capture Hi-C (CHi-C) interactions on the same chromosome (*cis*-interactions) were classified as within-TAD if the promoter and *cis*-regulatory element (CRE) fragments were both completely contained within the TAD boundaries. Those interactions where either the promoters or the CREs were partially overlapping any TAD boundaries were excluded[7].

TCGA colorectal cancer whole genome sequencing data

Whole-genome sequencing (WGS) data on 56 colon adenocarcinoma (COAD) and 17 rectal adenocarcinoma (READ) samples (12 microsatellite instable [MSI] and 61 microsatellite stable [MSS]) were downloaded from the NCI Genomic Data Commons Data Portal as tumour and matched-normal BAM files aligned to GRCh37 (accessed 7 July 2016). These samples had > 20x coverage and had matched tumour RNAseq data. Mutations were called across the whole genome using MuTect v1.1.7[8] according to best practice. Data from dbSNP v147 and COSMIC noncoding variants v77[9] were used for additional support. We excluded regions overlapping open reading frames (ORFs) (extended by 5bp to also account for splice sites) and 5′ UTRs and 3′ UTRs as defined by Ensembl v73[10]. To reduce false positives, the Duke excluded and HiSeqDepth top 5% regions defined by UCSC Genome Browser were omitted from the analysis[11]. Furthermore, we excluded immune system-coupled somatic hypermutation regions corresponding to 429 annotated immunoglobulin and the MHC loci (with each region extended by 50kb; Ensembl v73)[10]. FoxoG was used to remove any mutations that may have been caused by oxidation during sample preparation[12]. In addition, we ensured that variants were supported by a minimum of one alternative read in each strand direction, a mean Phred base quality score > 26, mean mapping quality ≥ 50, and an alignability site score of 1.0 when using the alignability of 100mers by GEM from ENCODE/CRG. For downstream analysis we excluded samples with: (i) > 500,000 mutations; (ii) *POLE*-mutated cancers; and (iii) MSS cancers with > 100,000 mutations, providing a final dataset of 12 MSI and 50 MSS (36 COAD and 14 READ) cancers.

Functional annotation and motif analysis in the identified CREs

We annotated the identified mutated and copy number variation (CNV)-affected CREs using LoVo and HT29 histone ChIPseq peaks. We determined whether somatic mutations were predicted to disrupt transcription factor (TF) binding using the MEME suite[13], in conjunction with the human TF motif database (HOCOMOCO v10[14], $P < 1.0 \times 10^{-4}$) and DeepSEA[15].

qRT-PCR

Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen) and cDNA was synthesised using SuperScript™ III Reverse Transcriptase (ThermoFisher Scientific) according to manufacturer's instructions. qRT-PCR was performed using TaqMan Fast Advanced Master Mix (ThermoFisher Scientific) and TaqMan Gene Expression Assay probes according to manufacturer's protocols. Each qPCR reaction was performed with three technical replicates on the 7900HT Fast Real-Time PCR System (ThermoFisher Scientific). To assess the relative mRNA level of *ETV1* and *RASL11A* the comparative C_T method was used: values were firstly normalised to *GAPDH* as endogenous control, then to mean target gene intensity within replicates and finally to the mean control sample intensity across replicates. Statistical significance was calculated using two-tailed *t*-tests over three independent experiments. TaqMan probe unique identifiers are detailed in **Supplementary Table 19**.

siRNA transfections

To test siRNA efficiency, 1.5×10^5 HT29 cells were seeded in 6-well plates and transfected 24 hours later using 10µl Lipofectamine RNAiMAX (ThermoFisher Scientific) and 50pmol siRNA (Eurofins Genomics) per well. Cells were collected after 24 hours and knockdown efficiency was assessed by qRT-PCR. siRNA sequences are reported in **Supplementary Table 19**. To perform cell viability and proliferation assays, $1.5\text{-}2.5 \times 10^3$ HT29 cells were seeded in 96-well plates and transfected using 0.3µl Lipofectamine RNAiMAX and 1.5pmol siRNA per well.

Survival analysis

To examine the relationship between gene expression and relapse-free survival (RFS) and overall survival (OS) in colorectal cancer, we made use of data from three independent patient cohorts from de Sousa *et al.*[16], Marisa *et al.*[17] and the TCGA[18] (**Supplementary Table 16**). Patients were censored based upon their last known clinical follow up. For RFS, patients that died without a relapse were censored at their time of death. For each series, gene expression was first treated as a continuous variable in Cox-regression with inclusion of age at diagnosis, sex, MSI/MSS status and tumour stage as covariates. If the MSI/MSS status of cancers had not been reported, status was retrieved from Hause *et al.*[19]. Analysis was performed using the log-rank test to estimate expression-associated hazard ratio (HR) and the 95% confidence interval (CI). The Wald test was used to determine statistical significance. Meta-analyses of the independent patient cohorts were performed under a fixed-effects model. We also stratified cancers by the expression of the gene. A tumour was said to have high or low expression of a gene if the expression value was within the top or bottom third of expression values for the gene across all cancers respectively. Kaplan-Meier analysis was then performed using this tumour stratification. The log-rank test was used to assess the differences between these survival distributions. Statistical tests were conducted using R v3.1.3[20].

REFERENCES

1. Lohse, M. *et al.* RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res* **40**, W622-7 (2012).
2. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-40 (2012).
3. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
4. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
5. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).

6. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

7. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).

8. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).

9. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-11 (2015).

10. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-5 (2014).

11. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-21 (2015).

12. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **41**, e67 (2013).

13. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-8 (2011).

14. Kulakovskiy, I.V. *et al.* HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res* **44**, D116-25 (2016).

15. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-4 (2015).

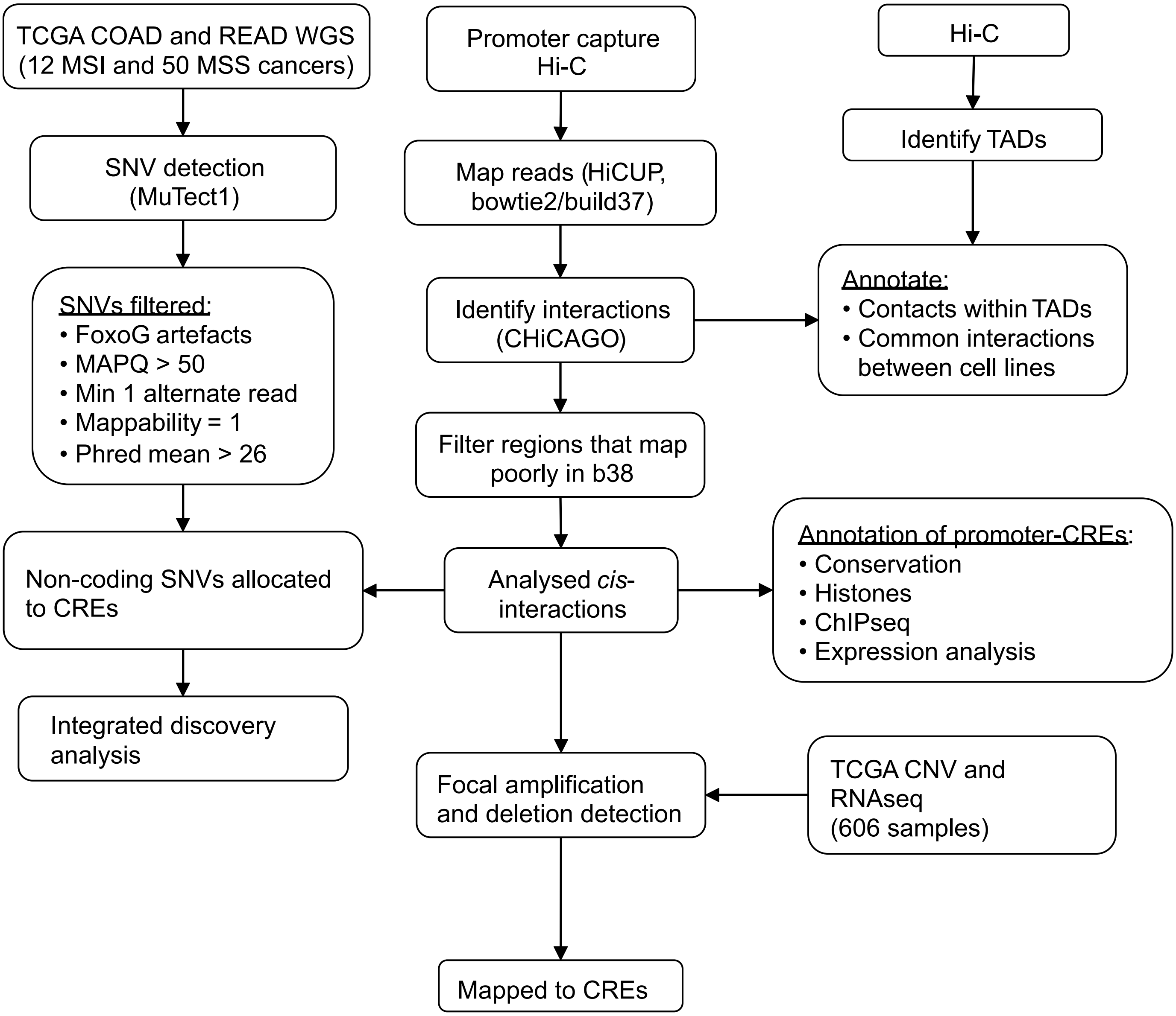
16. de Sousa, E.M.F. *et al.* Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. *Cell Stem Cell* **9**, 476-85 (2011).

17. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* **10**, e1001453 (2013).

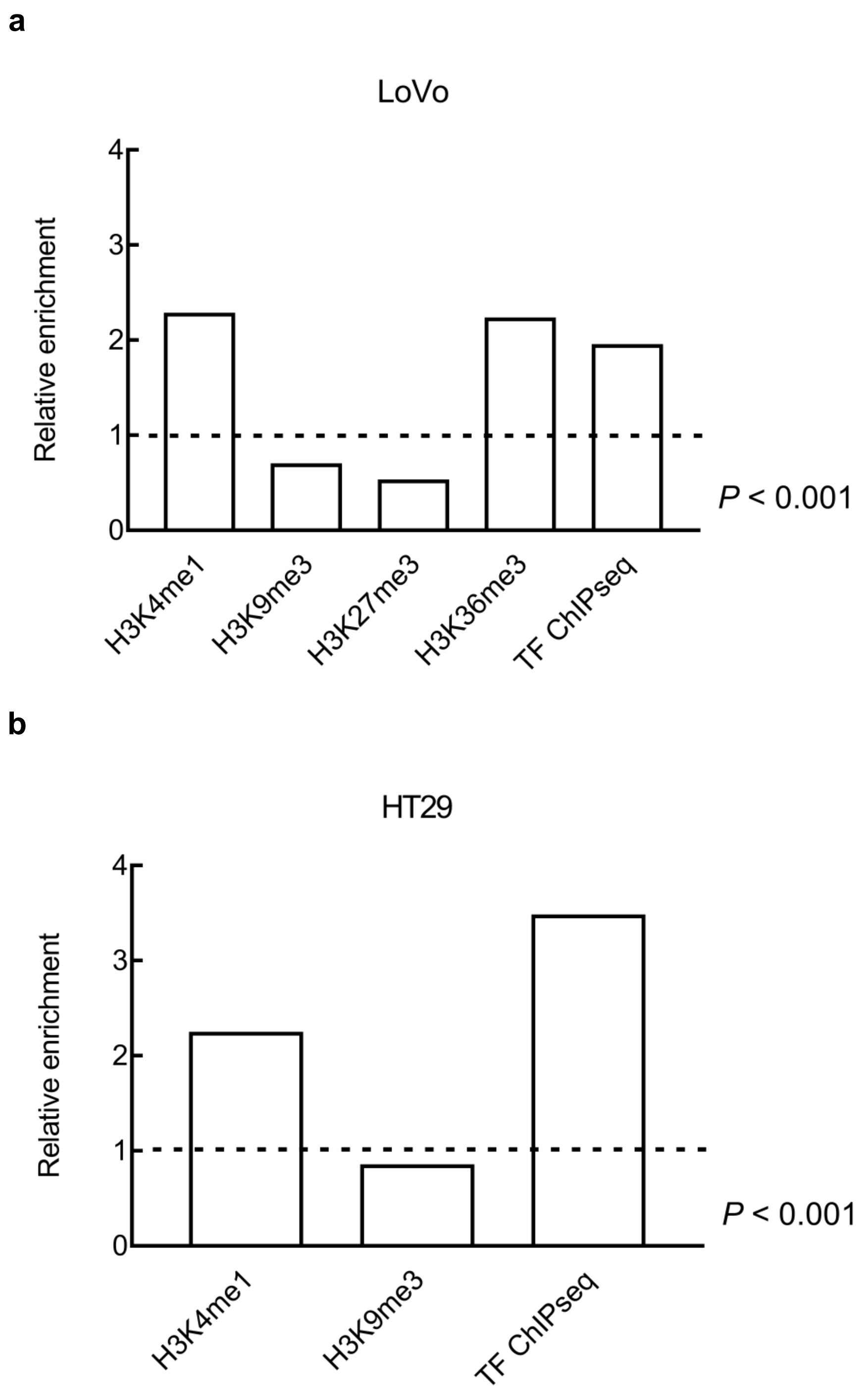
18. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-7 (2012).

19. Hause, R.J., Pritchard, C.C., Shendure, J. & Salipante, S.J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med* **22**, 1342-1350 (2016).

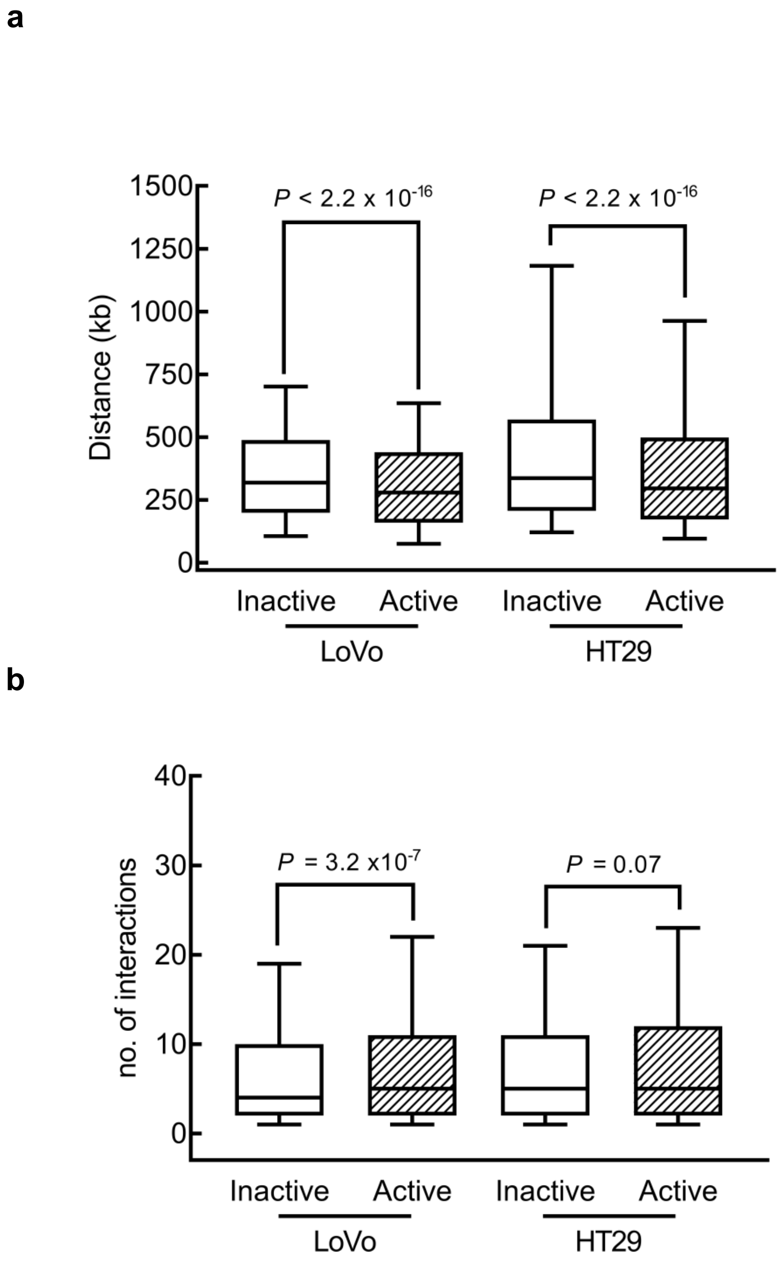
20. R Development Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2015).


TCGA COAD and READ WGS (12 MSI and 50 MSS cancers)
SNV detection (MuTect1)
SNVs filtered:
• FoxoG artefacts
• MAPQ > 50
• Min 1 alternate read
• Mappability = 1
• Phred mean > 26
Non-coding SNVs allocated to CREs
Integrated discovery analysis
Promoter capture Hi-C
Map reads (HiCUP, bowtie2/build37)
Identify interactions (CHiCAGO)
Filter regions that map poorly in b38
Analysed cis-interactions
Focal amplification and deletion detection
Mapped to CREs
Hi-C
Identify TADs
Annotate:
• Contacts within TADs
• Common interactions between cell lines
Annotation of promoter-CREs:
• Conservation
• Histones
• ChIPseq
• Expression analysis
TCGA CNV and RNAseq (606 samples)

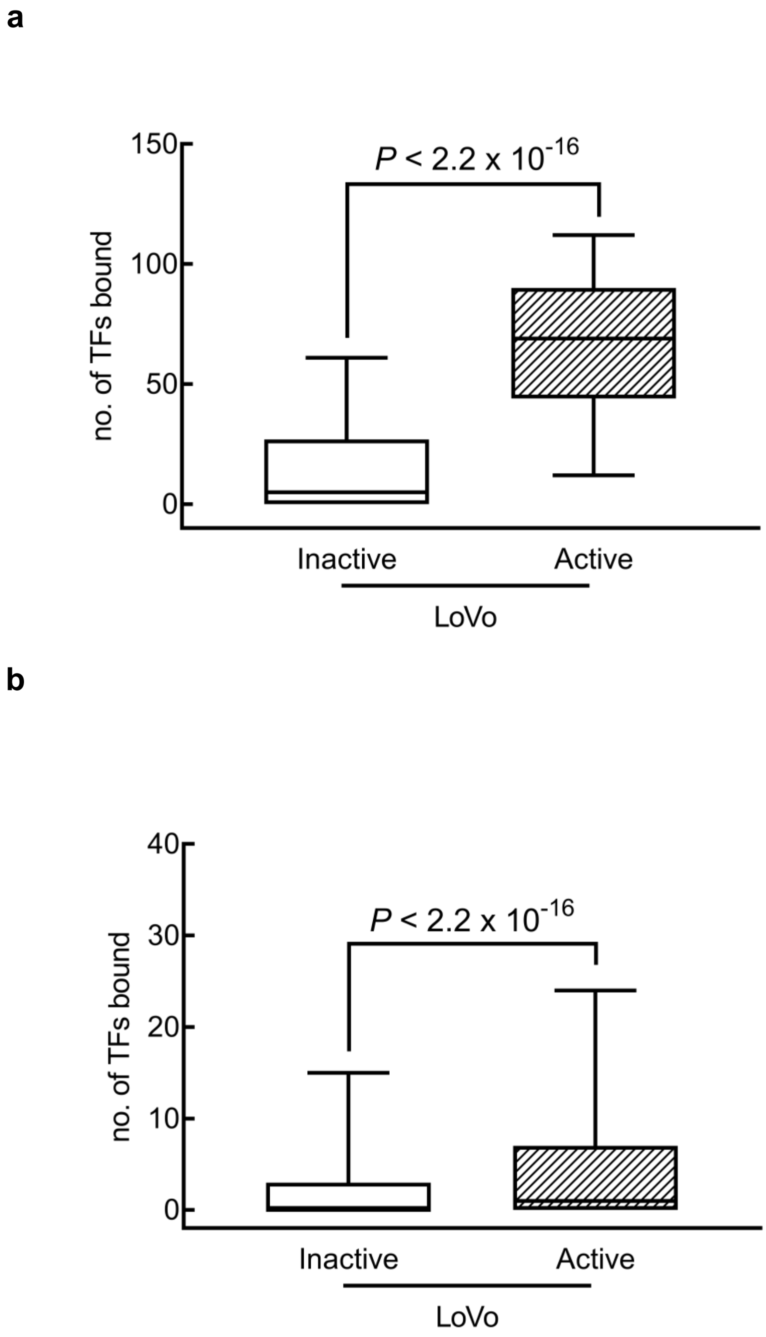
Supplementary Figure 1. Detailed workflow for the identification of mutated CREs in CRC. COAD, colon adenocarcinoma; READ, rectal adenocarcinoma; CRE, *cis*-regulatory element; TAD, topologically associated domain; CNV, copy-number variation; SNV, single nucleotide variant**.**



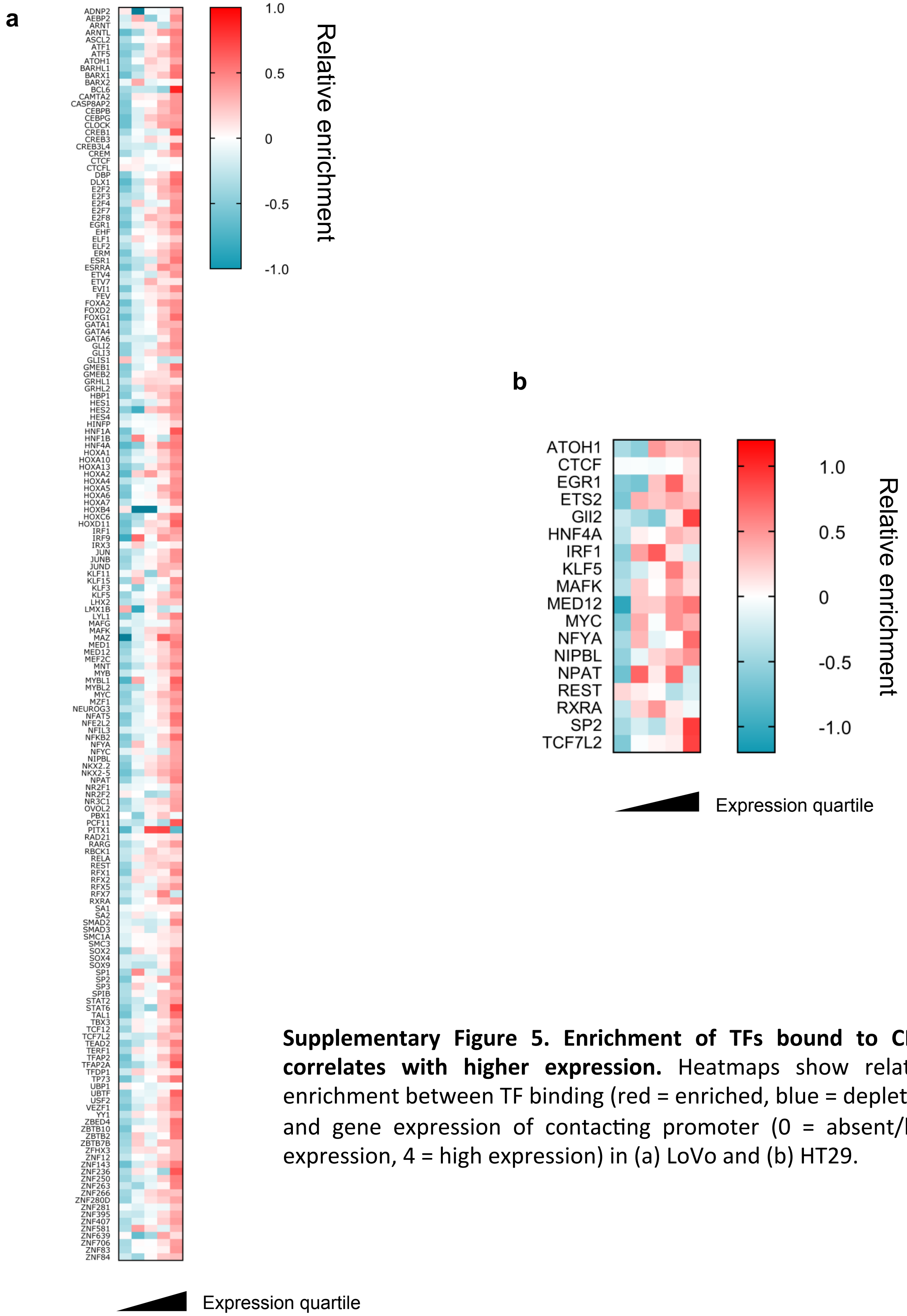
Supplementary Figure 2. CREs are enriched for active histone marks and TF ChIPseq peaks. Relative enrichment for epigenetic features in (a) LoVo and (b) HT29.



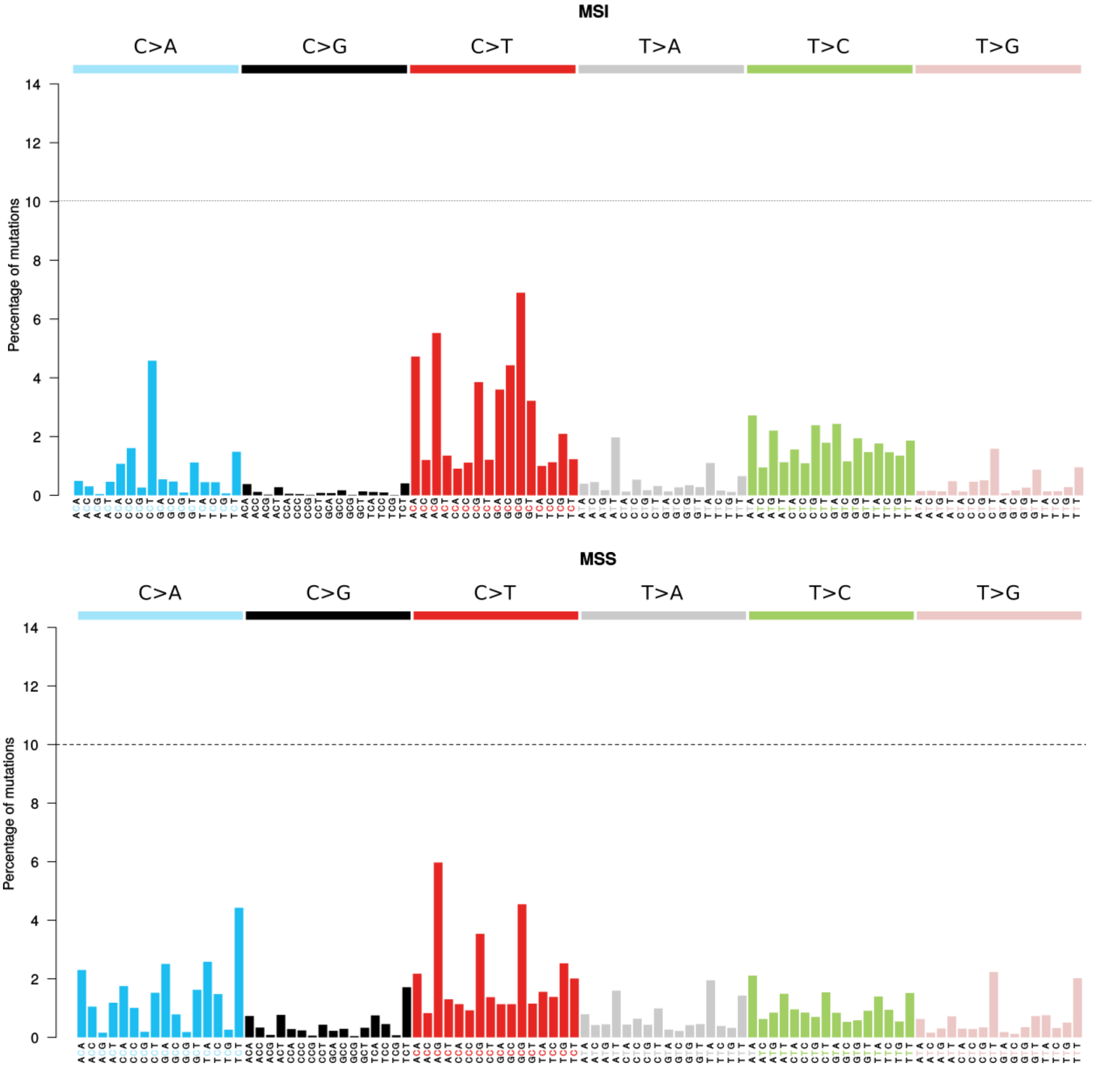
Supplementary Figure 3. Contacts between promoters and CREs show differences between active and inactive genes. (a) The distribution of distances (kb) between promoter and CREs and (b) the distribution of the number of contacts for each promoter, stratified by gene expression status (inactive or active) in LoVo and HT29 cell lines. Box plots denote quartiles. Whiskers correspond to the 10th and 90th percentiles. Differences assessed using two-sided Wilcoxon test.



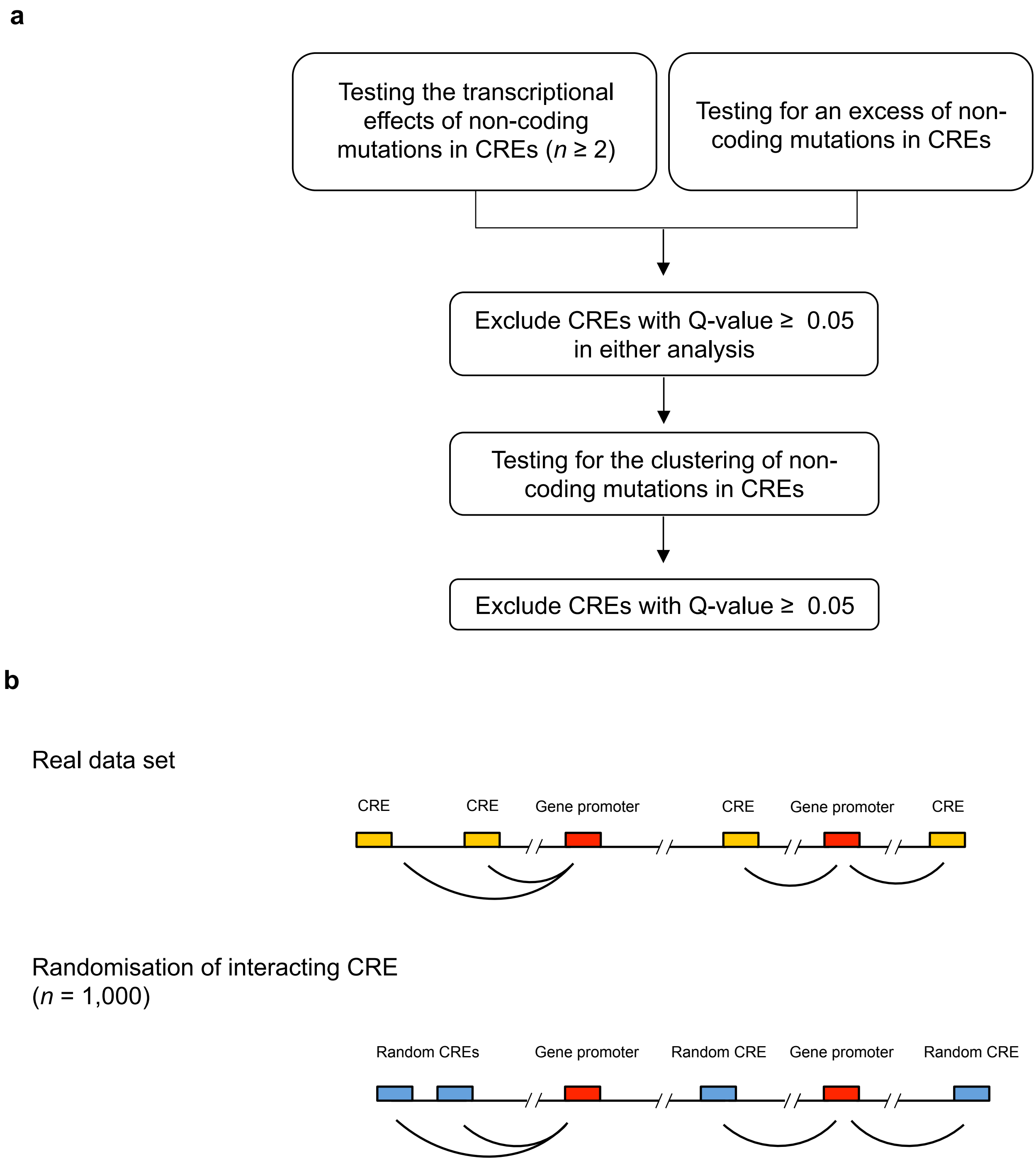
Supplementary Figure 4. Promoters and CREs of active genes are bound by an higher number of TFs. Distribution of TFs bound to (a) promoters and (b) CREs, according to gene expression status (inactive or active) in LoVo. Box-plots denote quartiles. Whiskers correspond to the 10th and 90th percentiles. Difference assessed using two-sided Wilcoxon test.



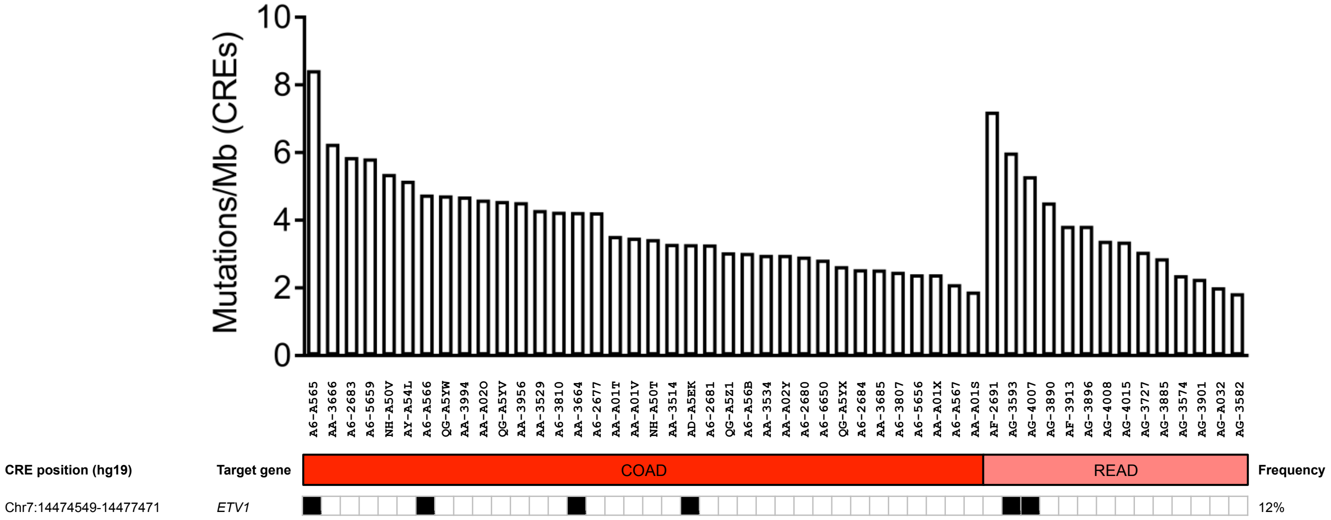
Supplementary Figure 5. Enrichment of TFs bound to CREs correlates with higher expression. Heatmaps show relative enrichment between TF binding (red = enriched, blue = depleted) and gene expression of contacting promoter (0 = absent/low expression, 4 = high expression) in (a) LoVo and (b) HT29.



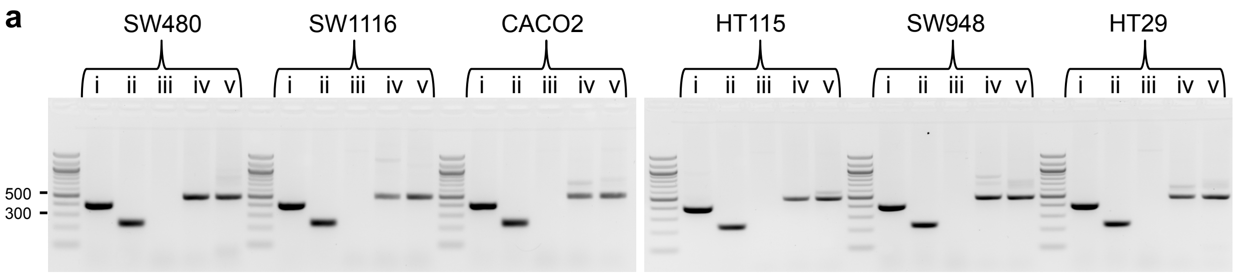
Supplementary Figure 6. Mutational signatures in the non-coding genome. MSI (*n*=12) and MSS (*n*=50).


a
Testing the transcriptional effects of non-coding mutations in CREs ($n \geq 2$)
Testing for an excess of non-coding mutations in CREs
Exclude CREs with Q-value ≥ 0.05 in either analysis
Testing for the clustering of non-coding mutations in CREs
Exclude CREs with Q-value ≥ 0.05
b
Real data set
CRE
CRE
Gene promoter
CRE
Gene promoter
CRE
Randomisation of interacting CRE
(n = 1,000)
Random CREs
Gene promoter
Random CRE
Gene promoter
Random CRE

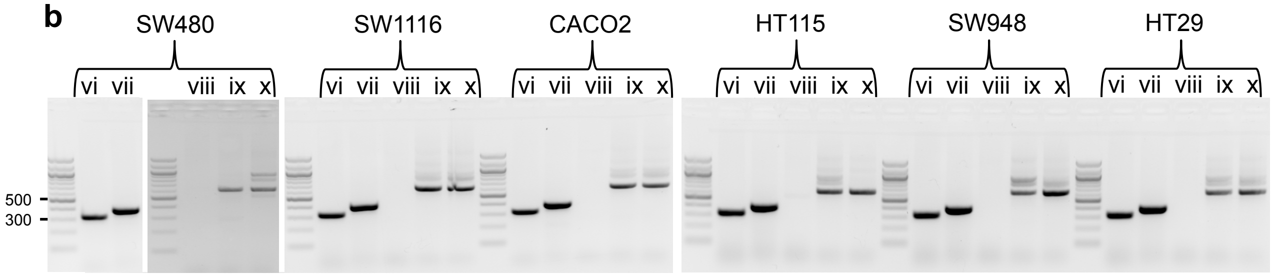
Supplementary Figure 7. Integrated discovery analysis. (a) Integrated driver discovery analysis applied to identify non-coding driver mutations. (b) Illustration of the CHi-C randomisation strategy in which randomly selected HindIII fragments contacting each gene were sampled, whilst maintaining the number of interactions each gene promoter is involved in. CRE, *cis*-regulatory element.



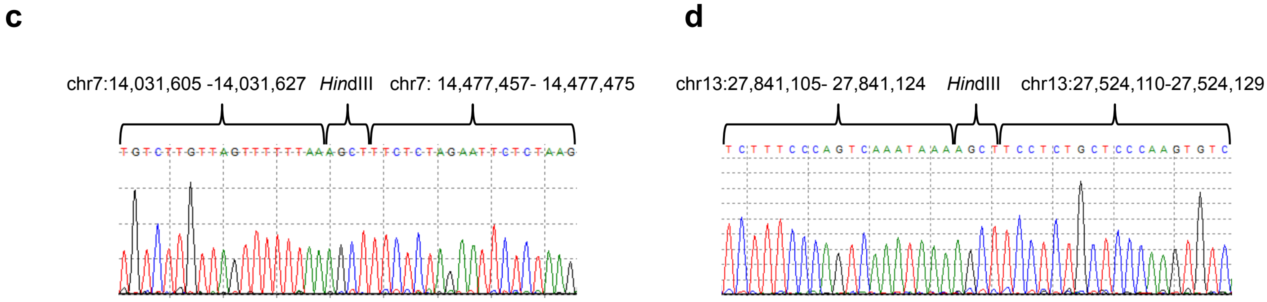
Supplementary Figure 8. Non-coding mutations in CREs. Upper panel shows the mutation rate across all CREs in each MSS cancer. Lower panel shows the presence of mutation in the *ETV1* CRE. COAD, colon adenocarcinoma; READ, rectal adenocarcinoma.



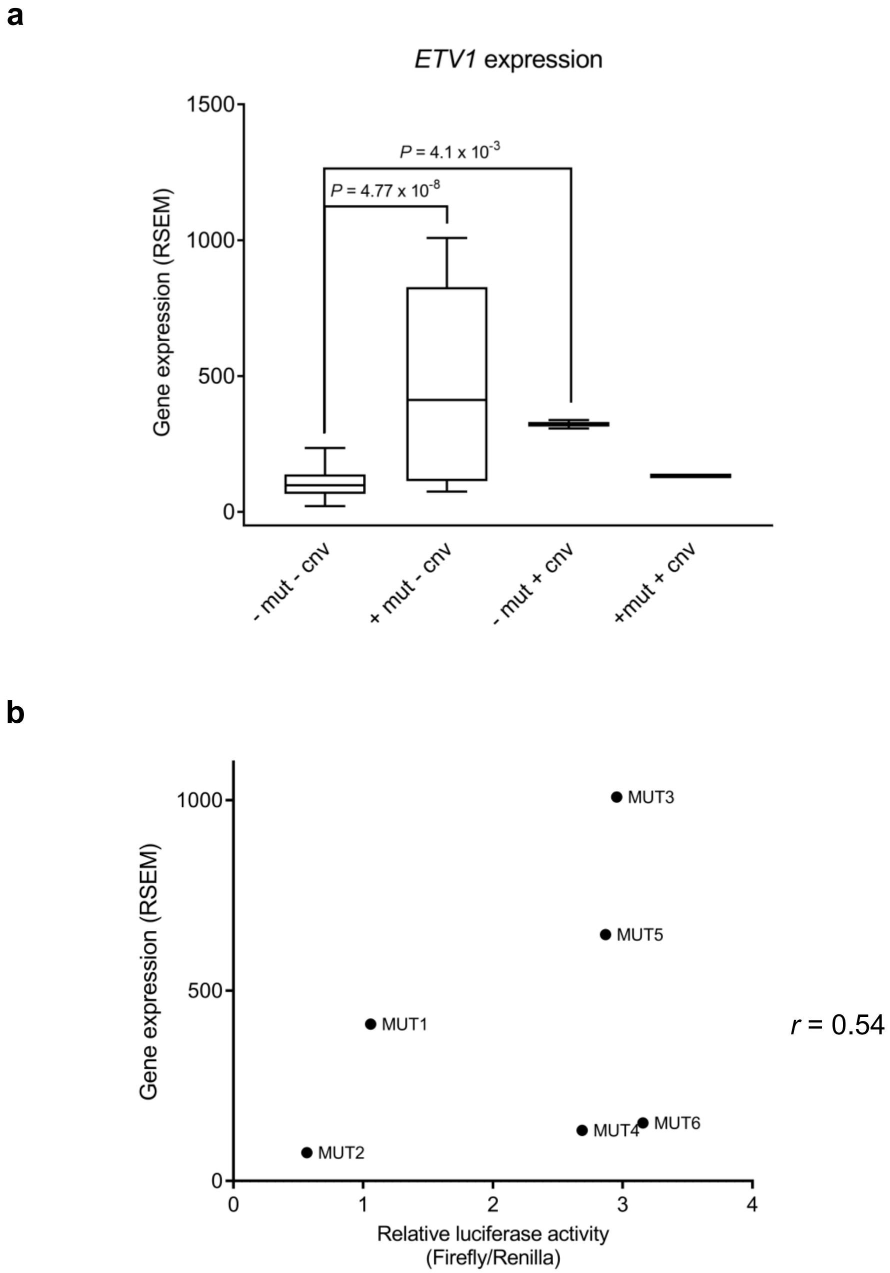
Fragment	Template	Region	Genomic Location (b37)	Size (bp)
i	Genomic	*ETV1* promoter	7:14031231-14031626	396
ii	Genomic	*ETV1* CRE	7:14477085-14477319	235
iii	Genomic	ETV1 promoter-CRE ligation	7:14031437-14031627(+)-AGCT-14477471-14477175(-)	496
iv	3C-1	ETV1 promoter-CRE ligation	7:14031437-14031627(+)-AGCT-14477471-14477175(-)	496
v	3C-2	ETV1 promoter-CRE ligation	7:14031437-14031627(+)-AGCT-14477471-14477175(-)	496



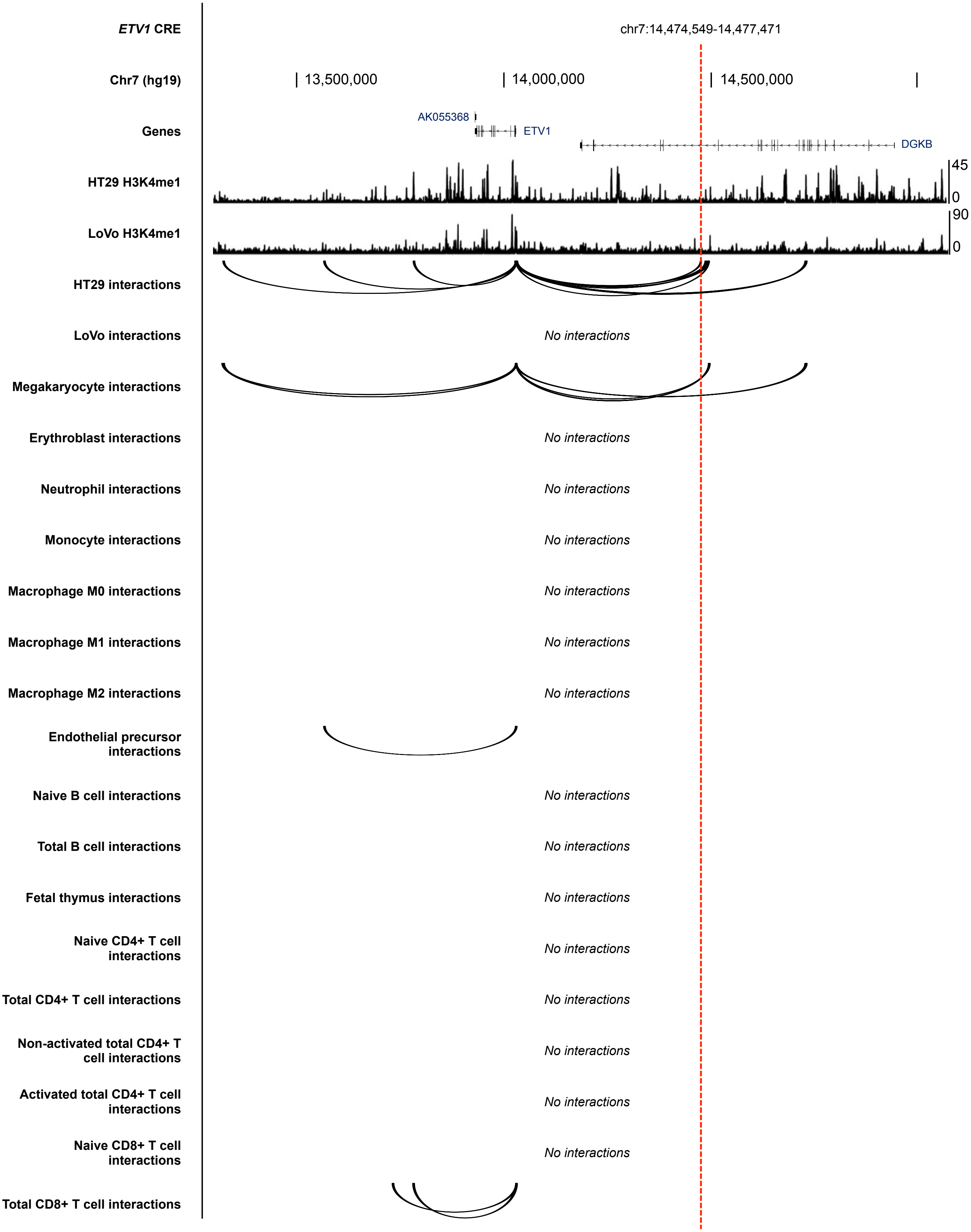
Fragment	Template	Region	Genomic Location (b37)	Size (bp)
vi	Genomic	*RASL11A* promoter	13:27841134-27841456	323
vii	Genomic	*RASL11A* CRE	13:27524159-27524555	397
viii	Genomic	*RASL11A* promoter-CRE ligation	13:27841364-27841105(-)-AGCT-27524110-27524473(+)	628
ix	3C-1	*RASL11A* promoter-CRE ligation	13:27841364-27841105(-)-AGCT-27524110-27524473(+)	628
x	3C-2	*RASL11A* promoter-CRE ligation	13:27841364-27841105(-)-AGCT-27524110-27524473(+)	628



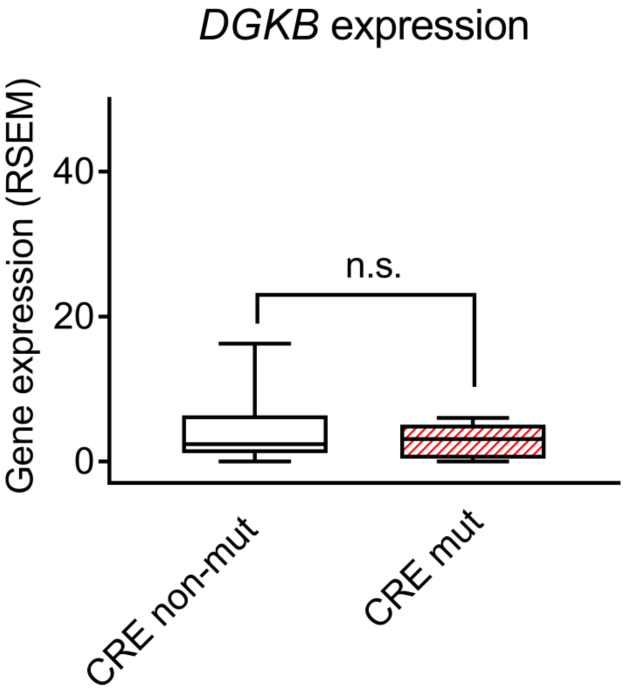
Supplementary Figure 9. Confirmation of *ETV1* and *RASL11A* promoter-CRE interactions in a panel of MSS CRC cell lines. Two cell culture independent replicates of *in situ* 3C libraries (3C-1, 3C-2) were used to validate (a) *ETV1* and (b) *RASL11A* promoter-CRE interactions. Genomic DNA was used as control. Primer pairs were designed to amplify promoter region, CRE region and across promoter-CRE ligation junction. (c and d) Sanger chromatograms of promoter-CRE fragments; promoters are shown to be ligated to their expected elements, separated by a HindIII cutting site.



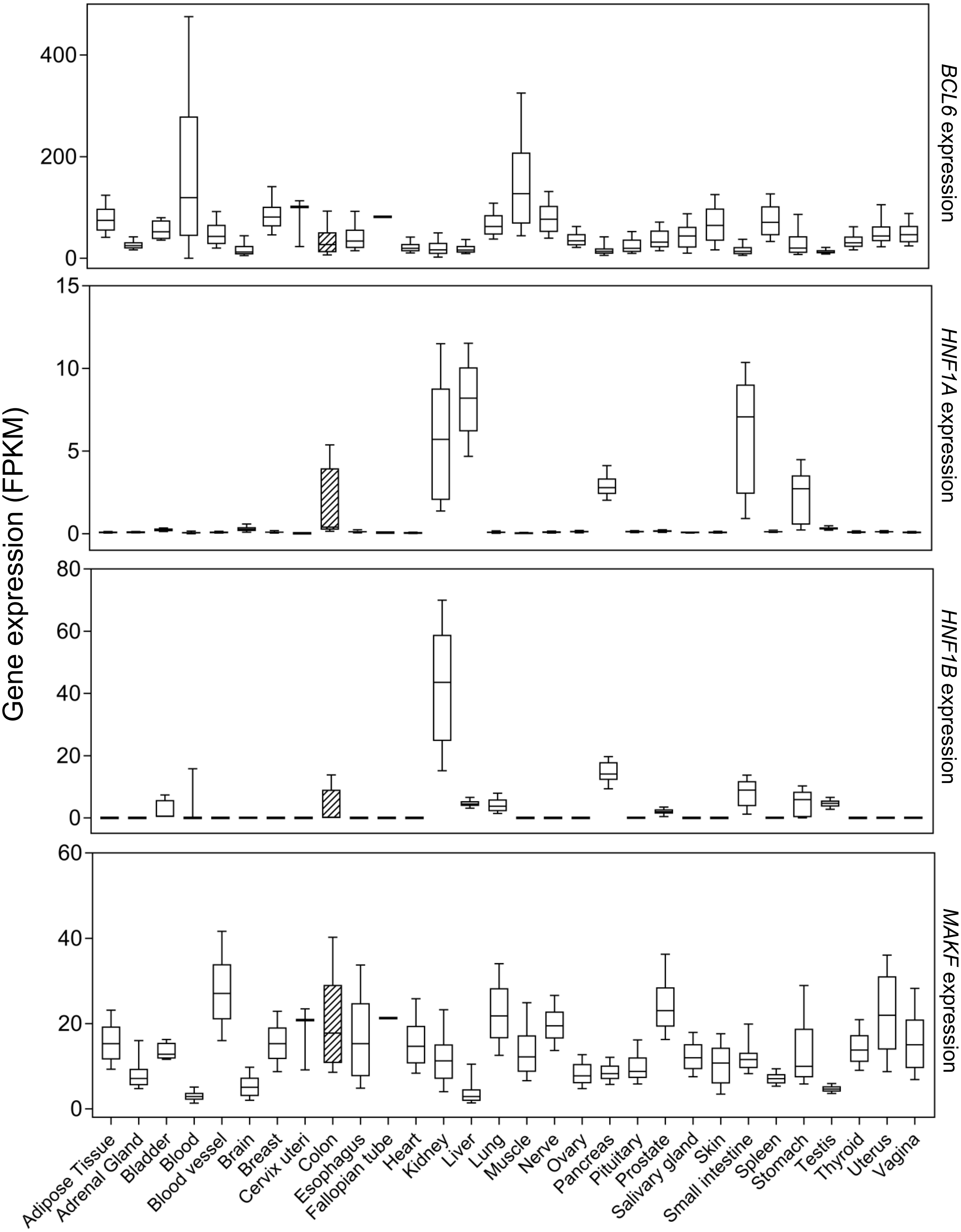
Supplementary Figure 10. Relationship between mutation status and *ETV1* expression in MSS cancers. (a) Samples are grouped by mutational status (-/+ mut) of the *ETV1* CRE (chr7:14,474,549-14,477,471) and *ETV1* coding region CNV status (-/+ cnv). Box-plots denote quartiles. Whiskers correspond to the 10th and 90th percentiles. Differences assessed by negative binomial test. - mut - cnv *n*=41, + mut - cnv *n*=5, - mut + cnv *n*=2, + mut + cnv *n*=1. (b) Correlation plot showing *ETV1* expression and relative luciferase activity for the identified six mutations. Spearman's rank correlation coefficient reported.



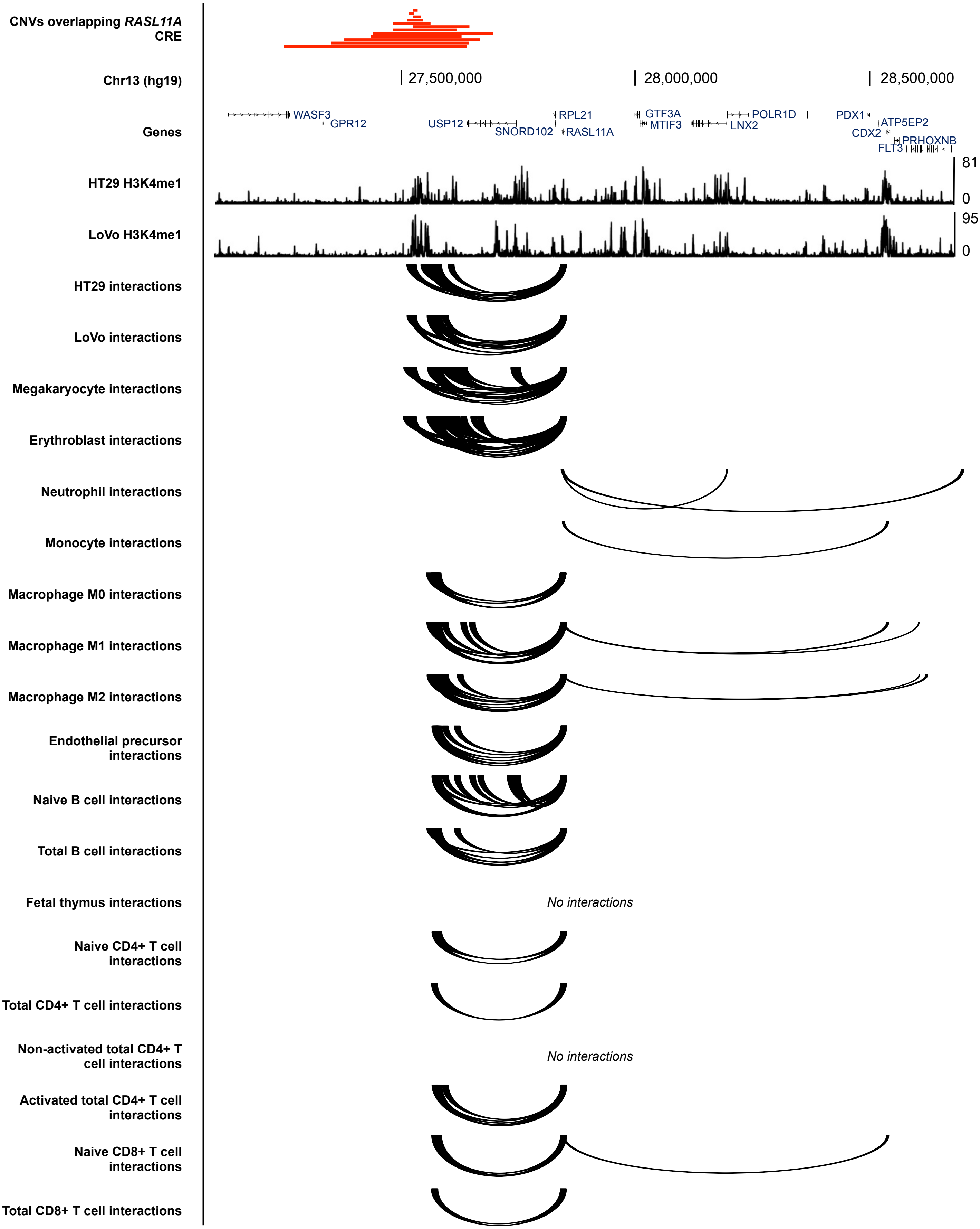
Supplementary Figure 11. *ETV1* promoter interactions in HT29, LoVo and 17 blood cell types and H3K4me1 mark in HT29 and LoVo. Red dotted line represents the position of *ETV1* CRE.



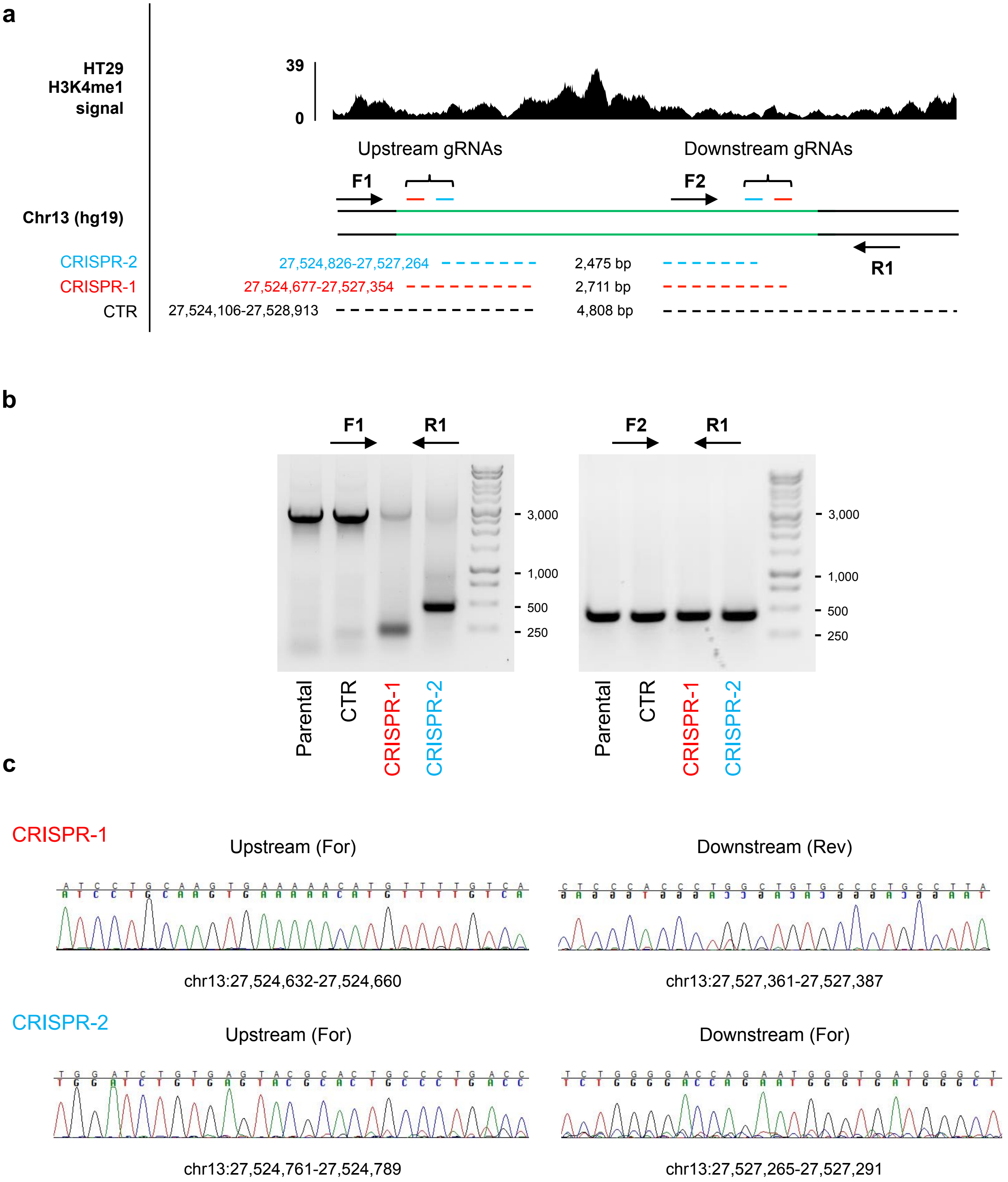
Supplementary Figure 12. Non-coding mutations in *ETV1* CRE intronic to *DGKB* are not associated with *DGKB* expression. Box-plots denote quartiles. Whiskers correspond to the 10th and 90th percentiles. Difference assessed by a negative binomial test, as implemented in edgeR (a two-sided approach; n.s., non-significant). CRE non-mut *n*=41, CRE mut *n*=5.



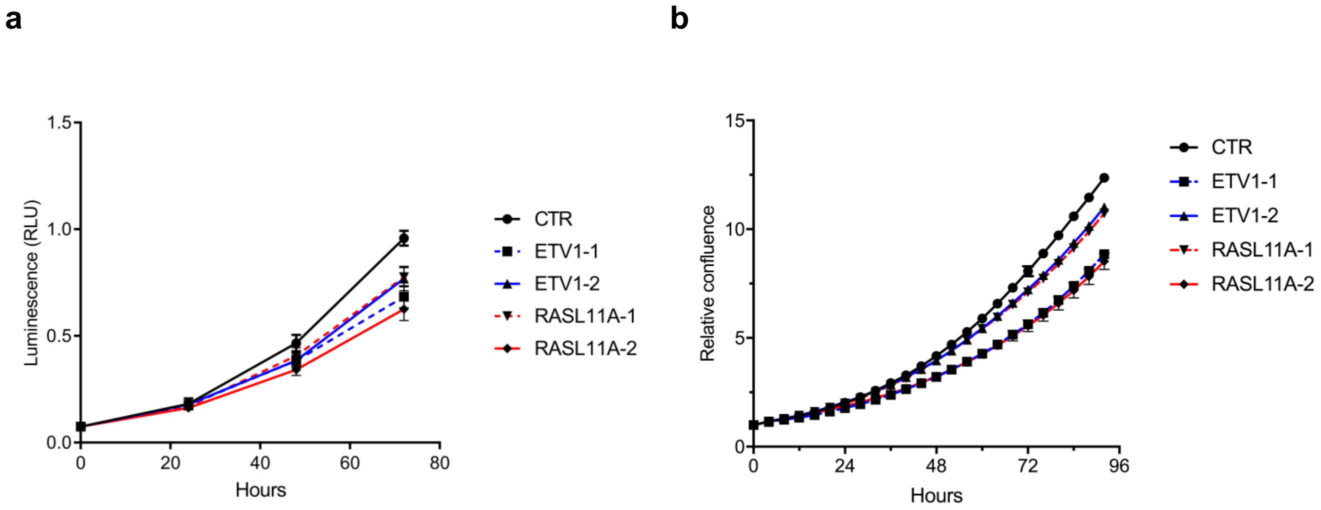
Supplementary Figure 13. Tissue-specific expression profiles for *BCL6*, *HNF1A*, *HNF1B* and *MAFK* using data from GTEx. Box-plots denote quartiles. Whiskers correspond to the 10th and 90th percentiles. Expression in colonic tissue shaded. Adipose Tissue *n*=403, Adrenal Gland *n*=97, Bladder *n*=4, Blood *n*=328, Blood Vessel *n*=465, Brain *n*=995, Breast *n*=135, Cervix Uteri *n*=3, Colon *n*=240, Esophagus *n*=456, Fallopian Tube *n*=1, Heart *n*=294, Kidney *n*=22, Liver *n*=81, Lung *n*=225, Muscle *n*=299, Nerve *n*=216, Ovary *n*=68, Pancreas *n*=107, Pituitary *n*=88, Prostate *n*=68, Salivary Gland *n*=46, Skin *n*=600, Small Intestine *n*=56, Spleen *n*=73, Stomach *n*=113, Testis *n*=112, Thyroid *n*=219, Uterus *n*=56 , Vagina *n*=62.



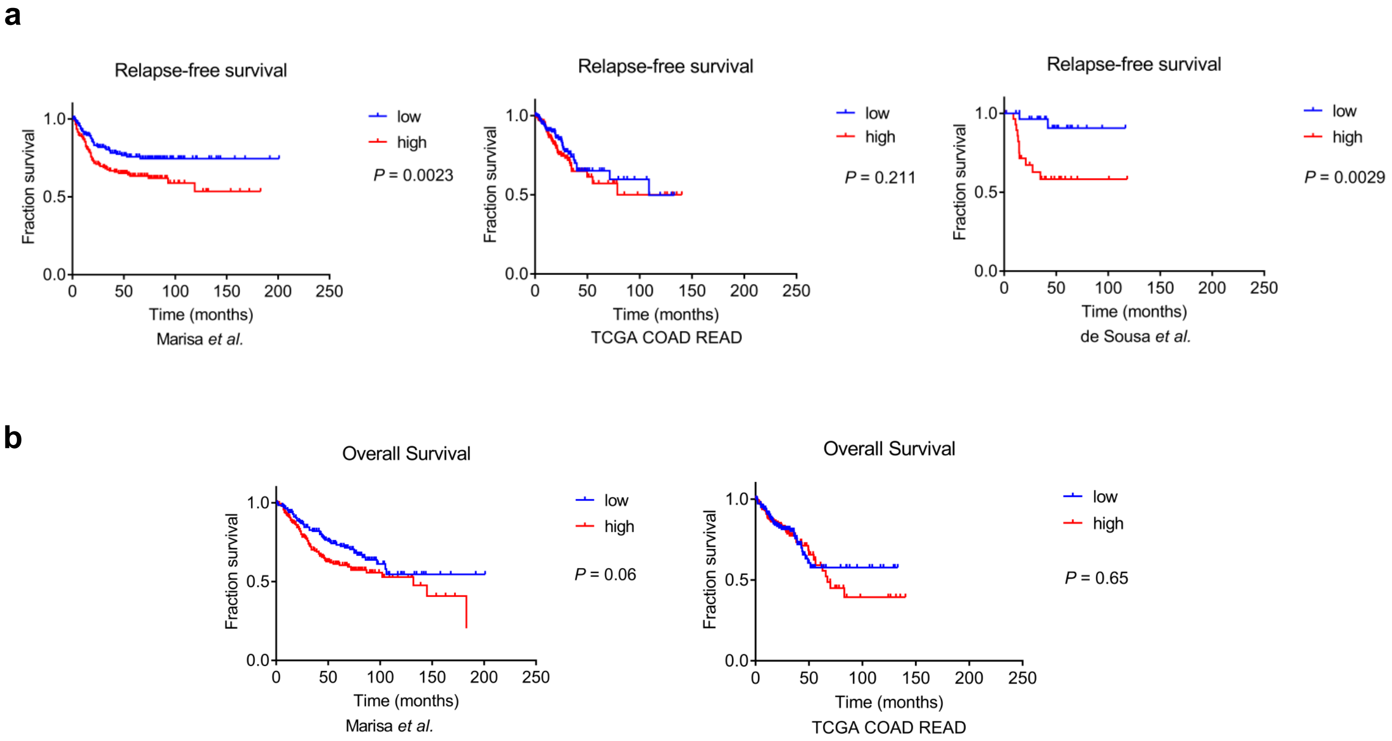
Supplementary Figure 14. *RASL11A* promoter interactions in HT29, LoVo and 17 blood cell types and H3K4me1 mark in HT29 and LoVo. Red lines represent positions of focal CNVs overlapping *RASL11A* CRE.



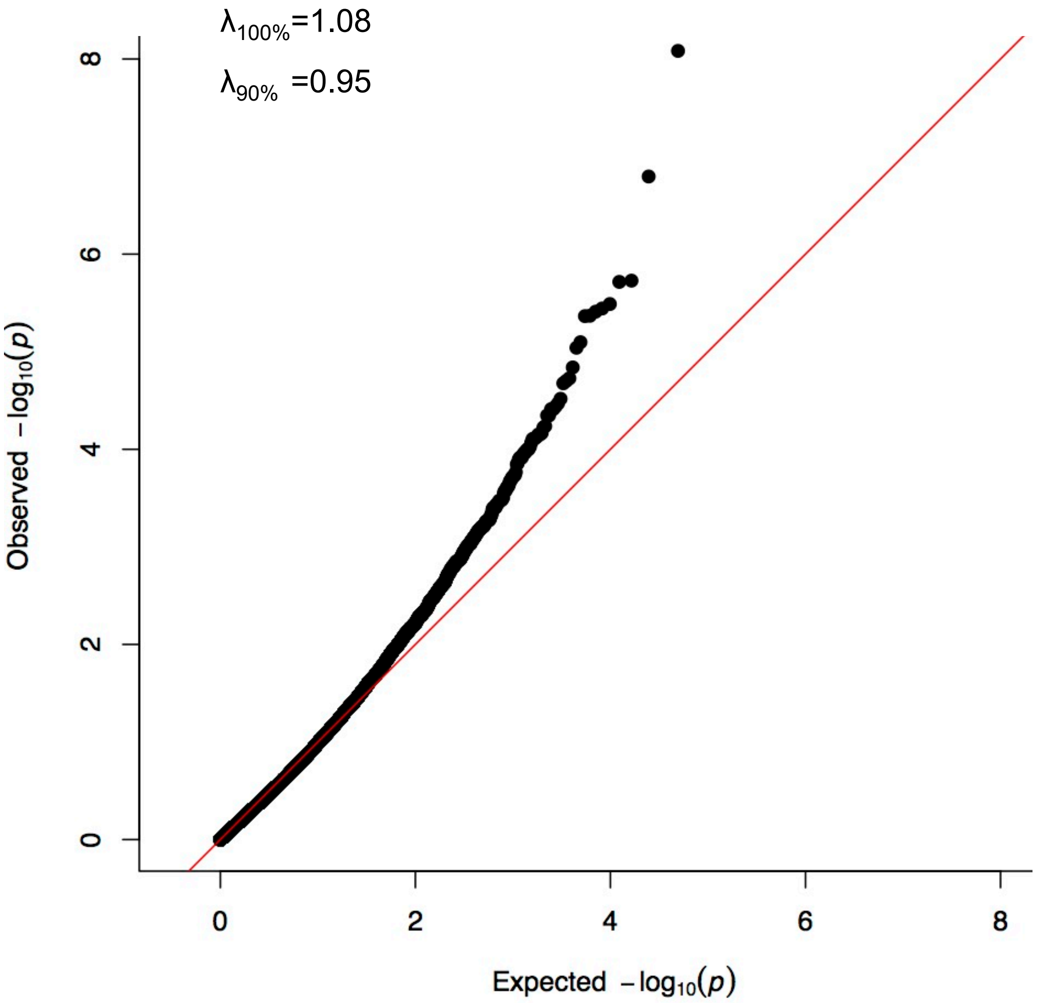
Supplementary Figure 15. CRISPR/Cas9-mediated deletion of the *RASL11A* CRE. (a) Schematic representation of the CRISPR/Cas9-mediated deletion and the primers used to validate the CRE disruption. The corresponding HT29 H3K4me1 ChIPseq signal is shown to demonstrate that functional elements of the CRE have been deleted. (b) Gel images show the PCR products obtained using either the primers at the boundaries (left) or the the internal primers (right); *n*=2 technical replicates. (c) Confirmation of the deletion is shown by the corresponding Sanger chromatograms.



Supplementary Figure 16. Cell growth curves. (a) Cell growth curves of siRNA-treated cells were assessed by measuring luminescence (RLU). Values normalised to the control (CTR) and plotted at 0, 24, 48 and 72 hours. (b) Cell growth curves of siRNA-treated cells were obtained by imaging each well every 4 hours. Values normalised to the 0h time point confluence and plotted as mean ± SEM using three independent experiments.



Supplementary Figure 17. *ETV1* expression levels are associated with patient prognosis. Kaplan-Meier plots of (a) relapse free-survival (Marisa *et al*: *n*=372, TCGA COAD READ: *n*=356, de Sousa *et al*: *n*=60) and (b) overall survival (Marisa *et al*: *n*=376, TCGA COAD READ: *n*=408) in patients stratified by high (top third) and low (bottom third) *ETV1* expression. Overall survival data only available for two series. Differences between survival distributions assessed with the two-sided log-rank test.



Supplementary Figure 18. Quantile-quantile plot comparing computed CRE mutational excess *P*-values to those expected under a uniform distribution. Red line represents the null hypothesis of no true mutational excess. Inflation factors were estimated using the regression approach, considering the 90% lowest *P*-values ($\lambda_{90\%}$) and 100% of the *P*-values ($\lambda_{100\%}$).