nature genetics

Pan-cancer analysis of somatic copy-number alterations implicates *IRS4* and *IGF2* in enhancer hijacking

Joachim Weischenfeldt[1–3,23], Taronish Dubash[4,5,23], Alexandros P Drainas[1,23], Balca R Mardin[1], Yuanyuan Chen[6], Adrian M Stütz[1], Sebastian M Waszak[1], Graziella Bosco[7], Ann Rita Halvorsen[8], Benjamin Raeder[1], Theocharis Efthymiopoulos[1], Serap Erkek[1,9], Christine Siegl[4,5], Hermann Brenner[10], Odd Terje Brustugun[8,11], Sebastian M Dieter[4,5], Paul A Northcott[12], Iver Petersen[13], Stefan M Pfister[9], Martin Schneider[14], Steinar K Solberg[15], Erik Thunissen[16], Wilko Weichert[17–19], Thomas Zichner[1], Roman Thomas[7,18], Martin Peifer[7,20], Aslaug Helland[8,11], Claudia R Ball[4,5,19], Martin Jechlinger[21], Rocio Sotillo[6], Hanno Glimm[4,5,19] & Jan O Korbel[1,22]

Extensive prior research focused on somatic copy-number alterations (SCNAs) affecting cancer genes, yet the extent to which recurrent SCNAs exert their influence through rearrangement of *cis*-regulatory elements (CREs) remains unclear. Here we present a framework for inferring cancer-related gene overexpression resulting from CRE reorganization (e.g., enhancer hijacking) by integrating SCNAs, gene expression data and information on topologically associating domains (TADs). Analysis of 7,416 cancer genomes uncovered several pan-cancer candidate genes, including *IRS4*, *SMARCA1* and *TERT*. We demonstrate that *IRS4* overexpression in lung cancer is associated with recurrent deletions in *cis*, and we present evidence supporting a tumor-promoting role. We additionally pursued cancer-type-specific analyses and uncovered *IGF2* as a target for enhancer hijacking in colorectal cancer. Recurrent tandem duplications intersecting with a TAD boundary mediate *de novo* formation of a 3D contact domain comprising *IGF2* and a lineage-specific super-enhancer, resulting in high-level gene activation. Our framework enables systematic inference of CRE rearrangements mediating dysregulation in cancer.

Recent studies have provided numerous insights into the extent to which somatic DNA alterations affect protein-coding genes[1–9]. However, 98–99% of the genome is made up of noncoding regions, a substantial fraction of which contain CREs[10–13]. CREs, such as enhancers, can control gene expression over long distances—up to a megabase or more—accompanied by physical contact of enhancers with the promoters of their target genes[14–17]. Several recent studies have uncovered somatic point mutations modulating gene regulation in cancer cells[18–20], including those affecting CREs near *TERT*[18,19], *PAX5* (ref. 21) and *TAL1* (ref. 22).

By comparison, much less focus has been placed on characterizing the effects SCNAs may have on CREs, in spite of the relevance of SCNAs in cancer[4,23–31] and surveys suggesting that several common cancers are driven largely by SCNAs[32]. Individual studies focusing on the pediatric cancer entities medulloblastoma[28] and neuroblastoma[29,30], as well as leukemia[33,34], recently uncovered examples where recurrent SCNAs, including gains and losses, mediate gene overexpression by juxtaposing enhancers near cancer-related genes, a process termed enhancer hijacking. Importantly, the identification of enhancer hijacking events has challenged the previously widely followed principle that the type of SCNA can be used to define the function of putative cancer genes, with gains representing candidate oncogenic loci and losses indicating tumor suppressor loci[35]. The extent to which enhancer hijacking occurs in different cancers

[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. [2]The Finsen Laboratory, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. [3]Biotech Research and Innovation Centre (BRIC), Copenhagen, Denmark. [4]Department of Translational Oncology, National Center for Tumor Diseases (NCT), Heidelberg, Germany. [5]Division of Translational Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]Division of Molecular Thoracic Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [7]Department of Translational Genomics, Center of Integrated Oncology Cologne–Bonn, Medical Faculty, University of Cologne, Cologne, Germany. [8]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital–The Norwegian Radium Hospital, Oslo, Norway. [9]Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [10]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. [11]Department of Oncology, Oslo University Hospital–The Norwegian Radium Hospital, Oslo, Norway. [12]Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [13]Institute of Pathology, Jena University Hospital, Jena, Germany. [14]General Surgery, Heidelberg University Clinics, Heidelberg, Germany. [15]Department of Cardiothoracic Surgery, Oslo University Hospital–Rikshospitalet, Oslo, Norway. [16]Department of Pathology, VU University Medical Center, Amsterdam, the Netherlands. [17]Institute of Pathology, Technical University Munich, Munich, Germany. [18]Department of Pathology, University Hospital Cologne, Cologne, Germany. [19]German Consortium for Translational Cancer Research (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [20]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. [21]European Molecular Biology Laboratory (EMBL), Cell Biology Unit, Heidelberg, Germany. [22]European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, UK. [23]These authors contributed equally to this work. Correspondence should be addressed to H.G. (hanno.glimm@nct-heidelberg.de) or J.O.K. (jan.korbel@embl-heidelberg.de).

Received 26 April; accepted 19 October; published online 21 November 2016; doi:10.1038/ng.3722

remains unclear, as studies focused on identifying this process across different cancer types have been lacking. Relevant cancer driver genes acting in different cancers may thus far have been overlooked, as cancer genome analyses do not systematically search for this process.

Here we describe a computational framework termed *cis* expression structural alteration mapping (CESAM), which uses statistical concepts from expression quantitative trait locus mapping to integrate SCNAs, expression and chromatin interaction domain data[36] to systematically identify SCNAs mediating gene dysregulation in *cis*. Using CESAM, we determined an estimate for the incidence of enhancer hijacking among the jumble of DNA rearrangements occurring in cancer genomes. Here we report the first validated cases of enhancer hijacking in common solid tumors, and we describe new mechanisms by which recurrent SCNAs mediate gene dysregulation.

RESULTS

CESAM: inference of SCNA breakpoints associated with expression alteration in *cis*

CESAM integrates SCNA breakpoint data with donor-matched transcriptome (mRNA-seq) data to identify candidate genes in *cis*, the altered expression of which is associated with SCNA-mediated rearrangements (**Fig. 1a**). This is achieved by linear regression of the mRNA-seq data on donor-matched SCNA breakpoint occurrence data (Online Methods). Thereby, CESAM relates gene expression values to binned SCNA breakpoints occurring in the vicinity of each gene. Breakpoint binning is achieved through the use of published data on TADs[36] (**Fig. 1a**), 3D chromosomal domains with a mean size of 830 kb, which are largely invariant across cell types[36–38]. TADs can confine physical and regulatory interactions between enhancers and their target promoters[38–42] and if disrupted can result in ectopic expression[34,42].

For TADs that are recurrently affected by SCNAs, expression association is tested independently for each gene located within the given TAD (**Fig. 1a** and **Supplementary Fig. 1**). CESAM further pursues independent filtering[43] to avoid testing genes that (i) have low expression, (ii) display minor levels of expression variance, or (iii) are recurrently deleted or amplified to a copy number ≥ 4 (**Fig. 1b** and Online Methods). To adjust for multiple testing, CESAM controls the false discovery rate (FDR) at 5%. Finally, CESAM summarizes functional annotations to facilitate inspection of proximal CREs.

Pan-cancer analysis of SCNAs affecting gene expression in *cis*

We used CESAM to analyze 7,416 previously published cancer genomes involving 26 tumor types from The Cancer Genome Atlas

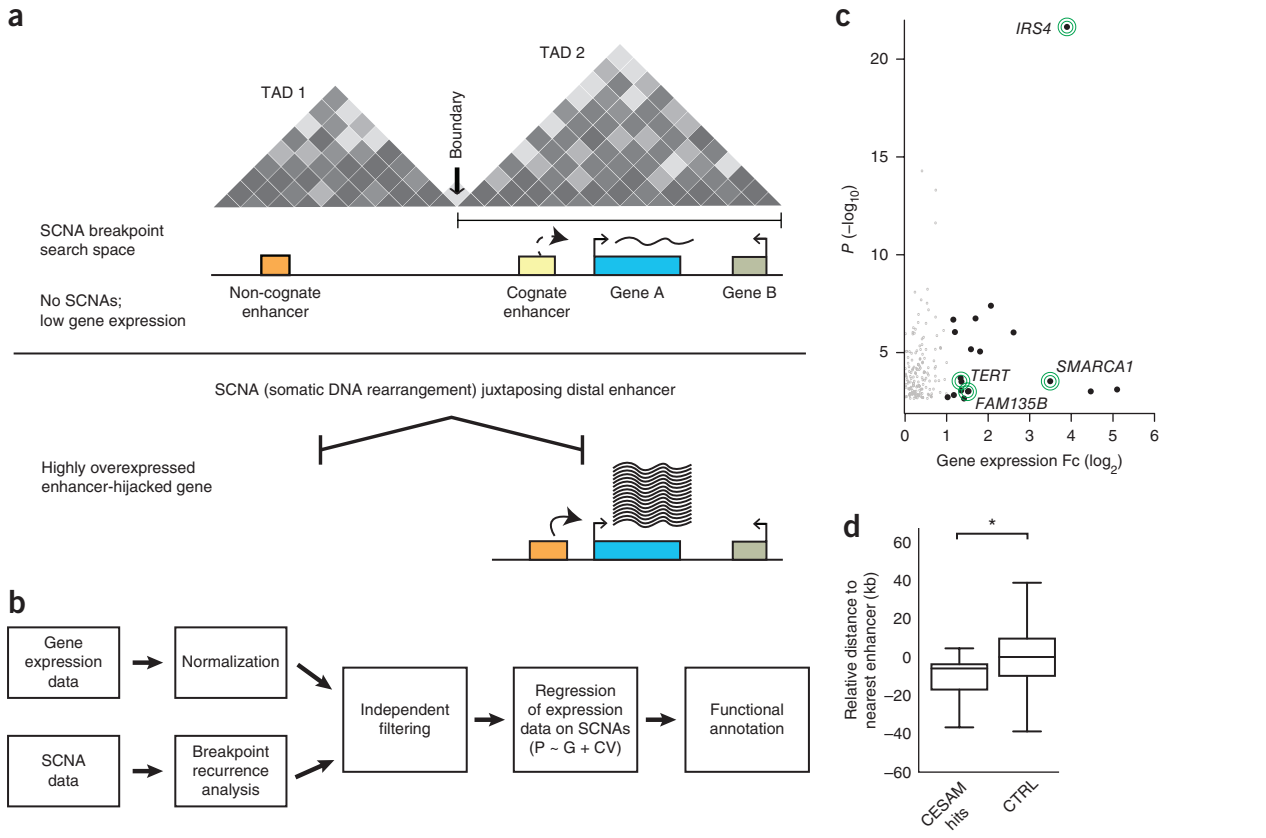

Figure 1 CESAM: framework for uncovering SCNAs driving gene dysregulation *in cis*. (**a**) The principle underlying CESAM. TADs are depicted as Hi-C-based contact maps[36] with gray shading indicating locus interactions (darker shading indicates stronger interactions as measured by Hi-C). SCNA breakpoints are binned within each TAD (SCNA breakpoint search space). (**b**) Detailed analysis workflow of CESAM. P, phenotype; G, SCNA genotype; CV, covariate. (**c**) Volcano plot of CESAM hits in a pan-cancer setting, with nominal *P* values plotted versus the expression fold change (Fc). Candidate genes identified by CESAM are shown as black dots (genes discussed in the text are highlighted by green circles). Gray dots denote loci removed on the basis of CESAM's filtering criteria (including removal of expression alterations driven by gene dosage change). (**d**) Relative distance to the nearest annotated enhancers at distal breakpoints of SCNAs identified by CESAM (CESAM hits) versus SCNAs not implicated by CESAM, which here were used as controls (CTRL) (*$P = 0.001$ based on 1,000 permutations using the s.d. of the observed proximity). Negative values indicate closer proximity to genomic features relative to background. In the box plots, the center line indicates the median, and the boxes denote the interquartile range (IQR) and extend from the first (Q1) to the third (Q3) quartile (edges). Whiskers represent Q1 − 1.5 × IQR and Q3 + 1.5 × IQR, respectively.

Table 1 Ranked list of CESAM pan-cancer candidate genes

Gene	Chromosome location of TAD	Fc	P_{adj}
IRS4	chrX: 107,720,001–108,600,000	15.0	2.47×10^{-18}
TBL1X	chrX: 8,640,000–10,080,000	4.4	1.15×10^{-5}
MTHFD1L	chr6: 151,200,000–151,760,000	3.4	4.01×10^{-5}
LIPA	chr10: 91,000,000–91,520,000	2.3	4.63×10^{-5}
PPP3CA	chr4: 101,080,000–103,480,000	2.4	0.000142
MLLT4	chr6: 167,400,000–169,000,000	6.2	0.000150
NCOR1	chr17: 15,880,000–16,440,000	3.1	0.000703
GIGYF2	chr2: 233,160,000–233,840,000	3.6	0.000822
BTD	chr3: 15,600,000–16,280,000	2.6	0.00862
SMARCA1	chrX: 128,400,000–129,200,000	11.5	0.0127
TERT	chr5: 40,000–1,720,000	2.7	0.0128
OSGIN1	chr16: 83,680,000–84,240,000	34.5	0.0250
TSC22D3	chrX: 105,560,000–107,240,000	2.7	0.0260
STUB1	chr16: 680,000–1,280,000	22.6	0.0275
FAM135B	chr8: 138,840,001–139,800,000	2.9	0.0298
KCNQ1	chr11: 2,160,000–3,600,000	2.3	0.0353
INSC	chr11: 16,760,001–17,520,000	2.1	0.0406
PTCHD1	chrX: 21,720,000–23,560,000	2.8	0.0463

The list is ranked by FDR-corrected *P* value and is continued as **Supplementary Table 1**. For each candidate gene, tumor types with at least three samples exhibiting upregulation were included. Fc, dosage-adjusted expression fold change for SCNA-carrier samples versus noncarrier controls; P_{adj}, adjusted *P* values according to the Benjamini–Hochberg procedure.

(TCGA) data portal ("URLs"). In this resource, SCNAs were defined on the basis of SNP6 arrays. Although these exhibit lower resolution than whole-genome sequencing for SCNA inference, presently the number of available specimens profiled using both mRNA-seq and SNP6 arrays markedly exceeds that of published whole-genome sequencing data sets with matched expression data (ICGC; see "URLs"). We first performed a pan-cancer analysis with CESAM and identified 18 gene loci with marked expression upregulation (fold change ≥ 2) in conjunction with *cis* SCNAs (**Fig. 1c**, **Supplementary Fig. 1**, **Table 1** and **Supplementary Table 1**). These encompassed several genes previously implicated in cancer, including *FAM135B* (found altered in esophageal cancer[44]); *SMARCA1* (ref. 45), a member of the SWI/SNF family of chromatin-remodeling proteins; and *TERT*, which encodes a catalytic subunit of telomerase. We observed a relatively high expression fold change at the pan-cancer level (>25-fold) for clustered deletions associated with upregulation of the insulin receptor substrate 4 gene (*IRS4*). Simulations demonstrated enrichment of annotated enhancers, clustered enhancers also referred to as super-enhancers[46], and promoters but not fragile sites at the distal end of SCNAs implicated by CESAM (**Fig. 1d** and **Supplementary Fig. 1**), in support of CRE-mediated activation mechanisms.

We first characterized the *TERT* locus, which CESAM identified in the greatest number of cancers, including kidney cancer, sarcoma and adrenocortical carcinoma (ACC). We observed the highest frequency in relation to cohort size (11.8%) for ACC (**Fig. 2a**,**b** and **Supplementary Table 1**). Pronounced clustering of SCNAs became evident at the *TERT* promoter, where overexpression-associated *cis* SCNAs were previously described in chromophobe kidney cancer[47]. *TERT*-overlapping SCNAs (i.e., gains), however, occurred more rarely in ACC (**Fig. 2a**). Across cancers we observed *TERT* expression fold changes of 2.7-fold in SCNA carrier versus pan-cancer noncarrier samples. Within individual cancer types (i.e., compared to noncarrier samples from the same tumor cohort), however, we frequently observed much higher fold changes—for example, >50-fold in ACC, kidney cancer and sarcoma. Both losses and gains contributed to overexpression, with deletions in *cis* occasionally resulting in even higher fold changes than high-level (copy number ≥ 4) *TERT* amplicons (**Fig. 2c** and **Supplementary Fig. 2**). Recent studies have implicated similar mechanisms of *TERT* upregulation in neuroblastoma[29] and chromophobe kidney cancer[47], lending support for common mechanisms of *TERT* activation involving *cis* SCNAs in different cancer types.

TAD-boundary-intersecting deletions are associated with *IRS4* dysregulation in sarcoma and squamous cancers

We next turned our focus to *IRS4*, a locus that CESAM identified in diverse cancer types. SCNAs in *cis* of *IRS4*, a gene located on chromosome X, were most commonly seen in lung squamous carcinoma (LUSC; $n = 22$; **Fig. 3a**), sarcoma ($n = 7$) and cervical squamous carcinoma ($n = 3$), although overall 48 samples from ten tumor types exhibited *IRS4* overexpression (**Supplementary Fig. 3** and **Supplementary Table 2**). Although not yet implicated in these cancer types, *IRS4* was previously shown to have cell-cycle-promoting capabilities *in vitro*; for example, *IRS4* overexpression has been shown to enhance *IGF1*-induced cell proliferation in the 3T3 cell line[48] and to mediate proliferation and cell migration in hepatoblastoma cells[49]. The gene is presumed to act via the PI3K–AKT pathway[49–52], with *IRS4* overexpression inducing phosphatidylinositol 3,4,5-trisphosphate and AKT activation[49–53], and AKT inhibitors blocking the growth-promoting effect of *IRS4 in vitro*[49]. In spite of these prior findings, it is presently unclear whether *IRS4* has any tumor-promoting role *in vivo*, and the relatively high recurrence level (e.g., 4.4% in LUSC; **Supplementary Table 1** and **Supplementary Fig. 4**) of *cis* alterations associated with *IRS4* overexpression prompted us to investigate this locus in further detail.

We focused our analysis on LUSC, in which CESAM identified a set of recurrent deletions ($n = 20$) clustering 103 kb downstream of *IRS4* within a region demarcated by chrX: 107,549,609–107,872,288 (hg19) (**Fig. 3a**). *IRS4* expression was increased on average by 400-fold in comparisons of LUSC deletion carriers to noncarrier control LUSC samples, and by 25-fold when pan-cancer deletion carriers were specifically compared to pan-cancer noncarrier controls, whereas other genes *in cis* exhibited only modest expression alteration by comparison (**Fig. 3b** and **Supplementary Figs. 5** and **6**). We also observed focal high-level *IRS4* gene amplifications in two LUSC samples as well as in several samples from other tumor types exhibiting massive overexpression, supporting *IRS4* as the most plausible target of recurrent SCNAs at this genomic locus (**Fig. 3** and **Supplementary Fig. 6**). The *cis* deletions, notably, intersected with a TAD boundary downstream of *IRS4* that also coincides with CTCF binding sites at an inferred insulator region[13] (**Fig. 3a**). In sarcomas and, to a lesser extent, cervical squamous carcinoma, CESAM identified recurrent deletions at the exact same genomic interval in association with *IRS4* overexpression (**Supplementary Figs. 3** and **6**), an interval in which clustered deletions and *IRS4* expression are also seen in benign uterine leiomyoma[54].

Among TCGA lung cancer cohorts, SCNAs associated with *IRS4* overexpression were confined to LUSC, with none of the TCGA lung adenocarcinoma samples exhibiting such events. This is noteworthy because altered PI3K–AKT pathway signaling has been found to be particularly abundant in LUSC[55,56]. We also observed inversely correlated expression of *IRS4* and its paralog *IRS2* ($r = -0.11$; $P = 0.008$, Pearson product-moment correlation; **Supplementary Fig. 4**) in LUSC. The mutual exclusivity pattern suggests complementary roles in activating PI3K–AKT pathway signaling[50,53,57]. Additionally, we observed a significant co-occurrence of deletions *in cis* of *IRS4* and amplifications of the *FGFR1* cancer census gene on chromosome 8 (Pearson's chi-square test, $\chi^2 = 7.6$; $P = 0.006$; **Supplementary Fig. 4**). This is notable because IRS4 associates with FGFR1 and can promote

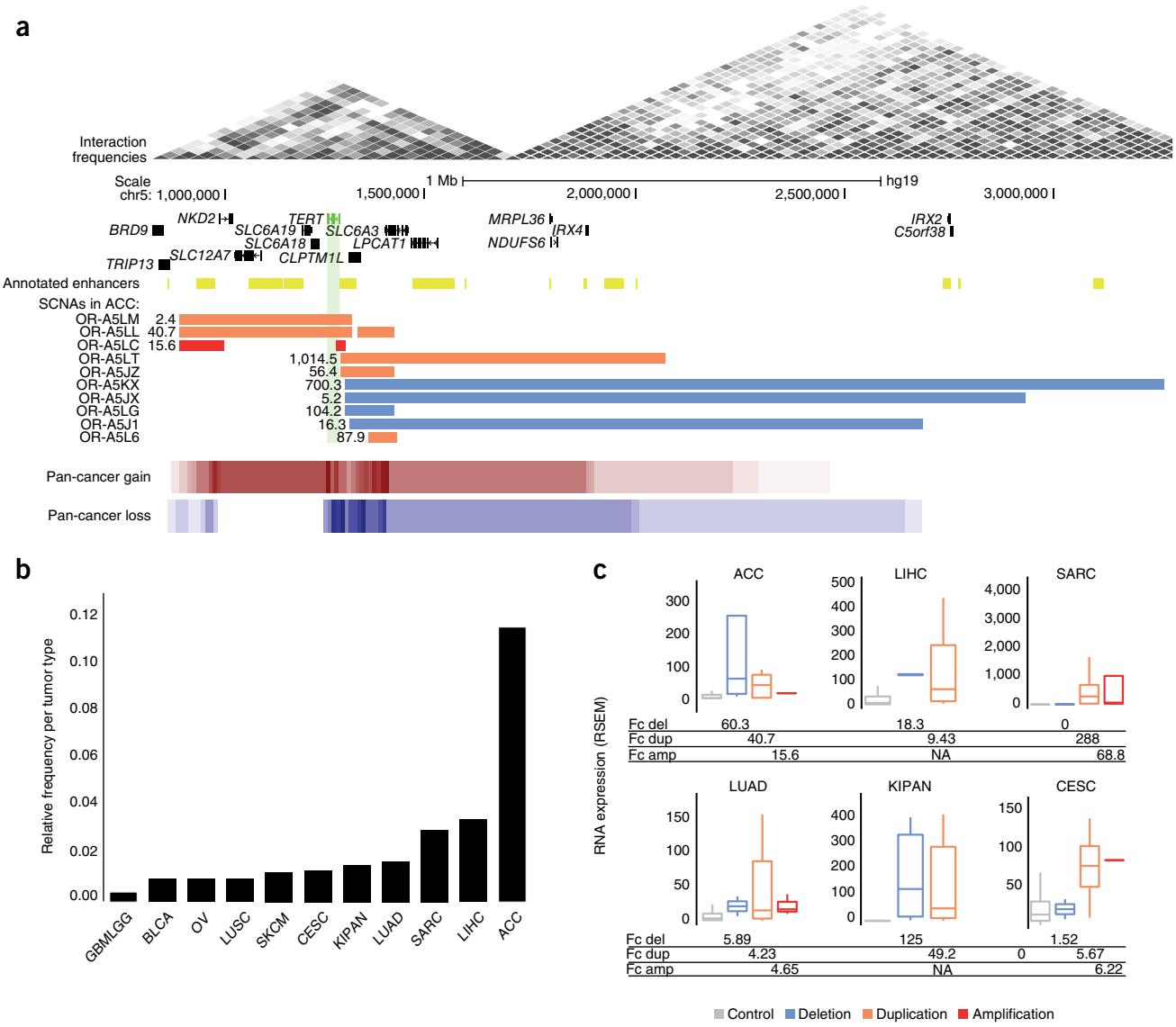

Figure 2 Analysis of the *TERT* locus: a CESAM pan-cancer hit. (**a**) Depiction of the *TERT* locus, the abnormal expression of which CESAM inferred to be mediated by *cis* SCNAs, for ACC and summarized across cancer types (pan-cancer copy-number gains and losses). Gene expression values reflecting fold changes versus noncarrier ACC samples are indicated for each SCNA. (**b**) Fraction of donors per tumor type for which CESAM inferred *TERT* dysregulation along with SCNAs in *cis* in at least three donors. GBMLGG, glioma; BLCA, bladder urothelial carcinoma; OV, ovarian serous cystadenocarcinoma; SKCM, skin cutaneous melanoma; CESC, cervical and endocervical cancers; KIPAN, pan-kidney cancer cohort; LUAD, lung adenocarcinoma; SARC, sarcoma; LIHC, liver hepatocellular carcinoma; ACC, adrenocortical carcinoma. (**c**) *TERT* expression values (unadjusted RSEM gene expression values) for different cancer types broken down by SCNA class. Del, deletion; dup, duplication; amp, amplification. In the box plots, the center line indicates the median, and the boxes denote the interquartile range (IQR) and extend from the first (Q1) to the third (Q3) quartile (edges). Whiskers represent Q1 − 1.5 × IQR and Q3 + 1.5 × IQR, respectively. The number of samples for each tumor type is provided in **Supplementary Table 1**.

FGFR1 signaling[58], and because FGFR1 can also activate PI3K–AKT pathway signaling[59]. Collectively these data implicate *IRS4* as a candidate genetic target in LUSC.

To investigate the tumor-growth-promoting effects of IRS4 *in vivo*, we subcutaneously injected a lung squamous cancer cell line, HCC-15, with and without an *IRS4*-overexpressing vector into athymic nude mice, performing two independent experimental replicates (with $n = 8$ and $n = 12$ mice, respectively; **Supplementary Note**). For this we introduced either a transgenic *IRS4* or an empty control lentivirus vector into HCC-15 cells. We observed palpable tumor formation in mice receiving transgenic *IRS4*-overexpression plasmids and in those receiving the empty control, albeit with a significantly increased growth of tumors harboring the *IRS4*-overexpression plasmids, in both experimental replicates ($P = 0.046$ and $P = 0.03$, respectively; two-tailed *t*-test; **Supplementary Fig. 7** and **Supplementary Table 3**). Resected tumors maintained *IRS4* overexpression, as shown by immunohistochemistry, quantitative reverse-transcription PCR (RT-qPCR) and flow cytometry (**Supplementary Fig. 7** and **Supplementary Table 3**), which strongly suggests a tumor-promoting effect of *IRS4* overexpression.

On the basis of the pronounced clustering of deletions downstream of *IRS4*, we hypothesized that alterations in chromatin structure or landscape may underlie *IRS4* dysregulation. To investigate this hypothesis, we performed experiments in primary LUSC specimens

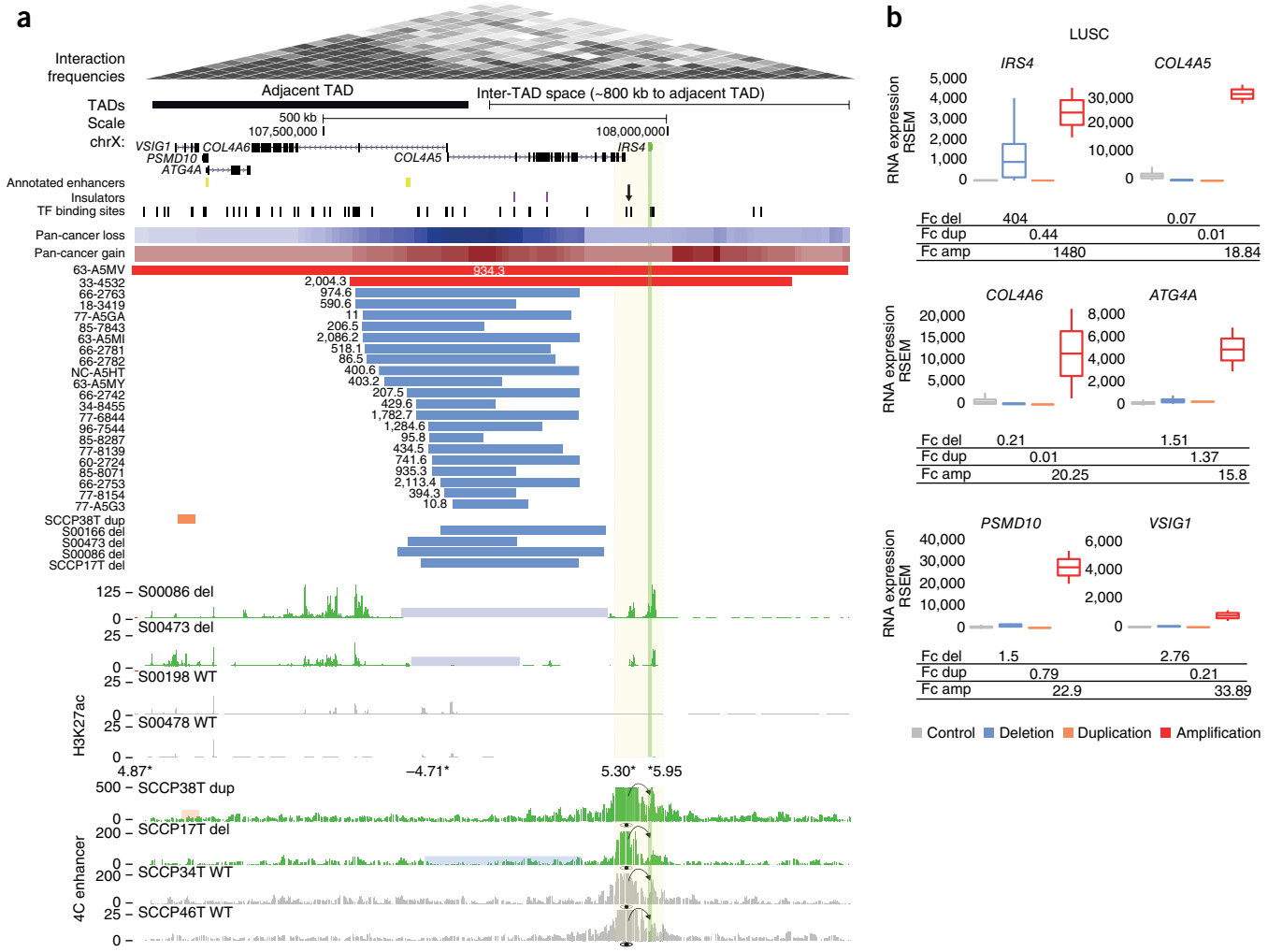

Figure 3 Recurrent SCNAs in *cis* are associated with a marked increase in *IRS4* expression. (**a**) Recurrent deletions at a TAD boundary near *IRS4*, and *IRS4* amplifications, are associated with *IRS4* dysregulation in LUSC. A region near *IRS4* exhibiting clustered transcription factor (TF) binding sites (candidate CRE) is indicated by an arrow. The recurrent deletions were evident both in male and female samples (indicating that both hemizygous and complete losses result in *IRS4* overexpression). Summarized SCNAs across cancer types (pan-cancer copy-number gains and losses) are shown as heat maps. The full list of pan-cancer SCNAs at the locus is presented in **Supplementary Table 2**. Deletion-carrier samples (del; in blue) exhibited marked H3K27ac[61] at the *IRS4* promoter and at the adjacent candidate CRE. SCNA carrier samples in which chromatin analyses were performed were confirmed to exhibit outlier expression by semiquantitative RT-PCR and qPCR (**Supplementary Fig. 3**, **Supplementary Table 3**, and data not shown). Asterisks indicate differentially occupied peaks identified by genome-wide H3K27ac analysis (values adjacent to asterisks indicate the fold change in H3K27ac signal for deletion carriers versus noncarriers). Results of 4C-seq experiments using the candidate CRE as a viewpoint in carrier versus noncarrier samples are depicted. dup, duplication; WT, wild-type locus. (**b**) LUSC expression measurements (unadjusted RSEM gene expression values) for carriers versus noncarriers revealing *IRS4* as the most plausible target. *IRS4* expression analyses revealed ~400-fold upregulation in deletion carriers and >1,000-fold upregulation for gene amplification carriers (number of controls, 470; del, 24; dup, 1; amp, 2). In the box plots, the center line indicates the median, and the boxes denote the interquartile range (IQR) and extend from the first (Q1) to the third (Q3) quartile (edges). Whiskers represent Q1 – 1.5 × IQR and Q3 + 1.5 × IQR, respectively.

(Online Methods). Expression analyses based on RT-qPCR in 94 primary LUSCs demonstrated greater than tenfold *IRS4* overexpression in 11 (12%) samples (**Supplementary Table 4**). We performed rearrangement screens in several samples using long-range paired-end sequencing[60] and identified *IRS4* proximal rearrangements in nine out of ten *IRS4*-overexpressing specimens (**Supplementary Fig. 8** and **Supplementary Table 4**). To investigate the chromatin landscape in these samples, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) in three deletion carriers and two controls (noncarrier LUSC samples both lacking the *cis* deletion and lacking *IRS4* overexpression). Several observations emerged from these experiments. First, we identified an accumulation of the active chromatin mark H3K27ac[61] on both sides of the commonly deleted region. Second, when comparing the deletion carriers to the controls, we observed four regions with differential H3K27ac marks (**Fig. 3a** and **Supplementary Note**). The strongest differential H3K27ac peaks within the wider genomic region of interest corresponded to *IRS4*, followed by a region 26 kb downstream of the gene exhibiting clustered transcription-factor-binding sites (**Fig. 3a** and **Supplementary Fig. 3**). None of the noncarriers exhibited measurable H3K27ac marks at this putative CRE, suggesting that its activity is confined to samples with *IRS4* upregulation. In addition, an H3K27ac peak at the bidirectional promoter of two nearby genes, *COL4A5* and *COL4A6*, encoding collagen type IV subunits, showed loss of signal consistent with deletion of these genes' promoter. Furthermore, we also observed modest differential H3K27ac signals near *VSIG1*, an

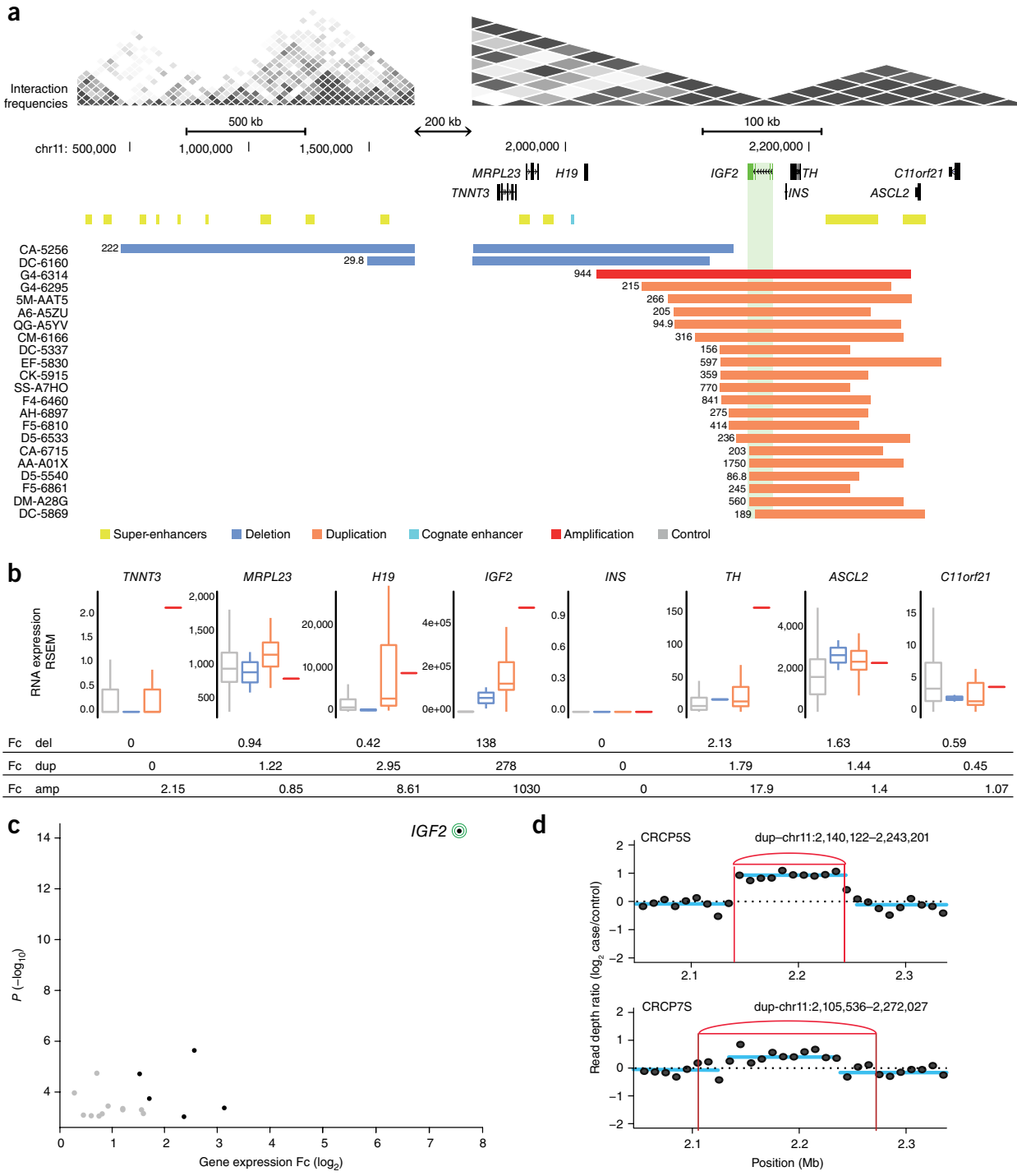

Figure 4 SCNAs associated with marked *IGF2* locus overexpression in *cis* in CRC. (**a**) Recurrent somatic duplications at the *IGF2* locus (green) associated with *IGF2* overexpression encompass a TAD boundary and a super-enhancer in the adjacent TAD but do not encompass the known *IGF2* cognate enhancer. Somatic deletions in *cis* extend over additional TAD boundaries. Although *IGF2* is positioned near a TAD boundary, the locus clearly falls into the TAD shown to the left (compare with **Fig. 5**). (**b**) Box plots depicting expression–SCNA relationships for all protein-coding genes within the respective TAD, with *IGF2* showing by far the most marked relationship, making it the most likely target of these recurrent SCNAs in *cis* (box plots are separated into deletion (del) carriers, duplication (dup) carriers, amplification (amp; >4 copies) carriers, and control samples lacking SCNAs in *cis*). $n = 356$ controls, 2 deletion carriers, 19 duplication carriers and 1 amplification carrier. The center line indicates the median, and the boxes denote the interquartile range (IQR) and extend from the first (Q1) to the third (Q3) quartile (edges). Whiskers represent Q1 $- 1.5 \times$ IQR and Q3 $+ 1.5 \times$ IQR, respectively. (**c**) Volcano plot of CESAM hits in CRC, with nominal *P* values plotted versus $\log_2$ expression change based on all samples with SCNAs in TAD (CESAM hits are depicted in black; *IGF2* is highlighted). (**d**) Structural variant detection by long-insert-size paired-end sequencing[4] followed by DELLY2 (ref. 75) analysis identified the presence of TAD-spanning *IGF2* locus tandem duplication in spheroid samples CRCP5S and CRCP7S (*IGF2* outlier expression was verified in both samples by qPCR; see **Supplementary Fig. 13** and **Supplementary Table 4**).

immunoglobulin-domain-containing gene that is expressed at only low levels in LUSC (and similarly in other cancers) and which exhibits no, or only modest, expression changes in conjunction with *cis* SCNAs (**Supplementary Figs. 5** and **6**).

To investigate whether the differences we observed in the chromatin landscapes of deletion carriers versus noncarriers are accompanied by differences in 3D chromosome conformation, we additionally performed 4C-seq (chromosome conformation capture sequencing[62]) experiments using the putative CRE downstream of *IRS4* as a viewpoint. These experiments revealed tight physical proximity between the putative CRE and *IRS4*, indicating that this genomic region indeed interacts with and thus represents a candidate *IRS4* enhancer. Interestingly, the physical contacts between this CRE and the *IRS4* promoter were also present in tumor specimens without the *cis* SCNA (**Fig. 3a** and **Supplementary Fig. 3**), an observation verified by 4C-seq experiments using the *IRS4* promoter as the viewpoint (**Supplementary Fig. 3**). These results suggest that TAD boundary or insulator-loss-mediated spreading of active chromatin in the context of already established promoter–enhancer interactions results in *IRS4* overexpression (see our model in **Supplementary Fig. 9**).

IGF2: a CESAM hit in colorectal cancers exhibiting *IGF2* locus tandem duplication

We next carried out analyses focused on individual tumor types with CESAM, pursuing independent assessment across 26 cancer types. We identified between 1 and 14 candidates per cancer type, with a total of 98 genes implicated by CESAM in these tumor-type-focused analyses (**Supplementary Table 1** and **Supplementary Fig. 10**). A CESAM candidate catching our attention was the *IGF2* locus on chromosome 11, which CESAM implicated in colorectal cancer (CRC). *IGF2* was >250-fold overexpressed in CRCs harboring nearby SCNAs compared with CRC noncarrier controls, whereas other genes nearby showed no or only modest expression alterations (**Fig. 4a–c**). We found that 22 out of 378 (6%) CRCs from the TCGA resource exhibited *IGF2* upregulation in conjunction with *cis* SCNAs (**Fig. 4a**). Previously, *IGF2* high-level overexpression in CRC was thought to result from recurrent focal locus amplification[3,63,64], that is, elevated gene dosage of a locus encompassing both *IGF2* and the *MIR483* microRNA gene[3,47,48]. The microRNA gene, which is embedded within intron 8 of *IGF2*, was recently implicated as a driver oncogene[63,64]. Given the joint upregulation of *IGF2* and *MIR483* in CRC[3,63,64] and the fact that both have been implicated in dysplasia and tumorigenicity[63,64], we herein refer to this locus as the *IGF2* locus for simplicity.

Among the CRC samples exhibiting *IGF2* dysregulation, 20 harbored gains and 2 harbored focal deletions in *cis* (**Fig. 4a**). Detailed examination of the SNP6 data showed that the corresponding gains at this locus typically underlay single-copy duplications (copy-number ratio of 1.25–1.75; **Fig. 4a** and **Supplementary Fig. 11**), whereas only a single sample of the TCGA cohort showed higher-level locus amplification (copy number 6). The unusually high and consistent upregulation (>250-fold) in this context suggests that rather than gene dosage increases, specific locus rearrangements may drive *IGF2* dysregulation.

To further characterize the mechanism of *IGF2* activation, we next performed experiments with spheroid cultures derived from primary CRC samples (**Fig. 4d**, **Supplementary Fig. 12** and Online Methods). Expression profiling using RT-qPCR identified two CRC-derived spheroids overexpressing *IGF2*, termed CRCP5S and CRCP7S (**Supplementary Fig. 12** and **Supplementary Table 5**). Using long-range paired-end sequencing[60], we uncovered single-copy tandem duplications in both CRCP5S and CRCP7S that seamlessly overlapped

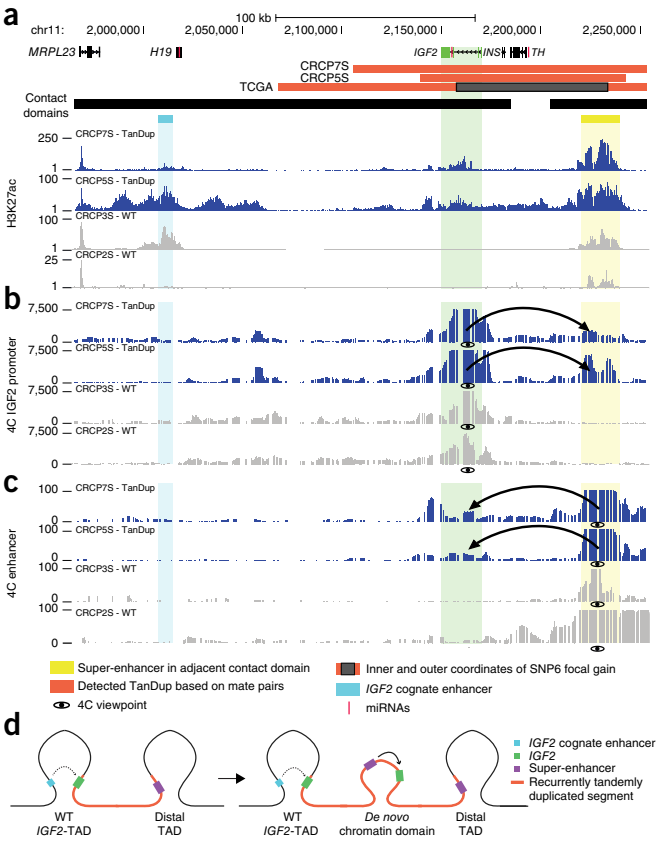

Figure 5 Verification of *IGF2* enhancer hijacking and model for mechanism involving *de novo* contact domain formation. (**a**) ChIP-seq for H3K27ac[61] yielded signals consistent with the activity of a previously annotated[76] lineage-specific super-enhancer in the TAD adjacent to the *IGF2* locus, but within the region accompanied by the recurrent somatic tandem duplication (TanDup) high-resolution contact domains from ref. 66. (**b**) 4C-seq experiments using the *IGF2* promoter region as the viewpoint demonstrated physical interaction between *IGF2* and the super-enhancer in TanDup-carrier samples, but not in noncarrier samples (wild-type (WT)). (**c**) 4C-seq experiments using the super-enhancer as the viewpoint verified the highly specific physical interaction with *IGF2* in TanDup carriers (but not in WT samples). Further control data for an additional WT sample are presented in **Supplementary Figure 11**. (**d**) New model for high-level gene overexpression at the *IGF2* locus in CRC that involves TanDup-mediated *de novo* contact domain formation resulting in the hijacking of a lineage-specific super-enhancer.

with the SNP6-based single-copy duplications in terms of size and position (**Fig. 4a**,**d**). These data indicate that the recurrent gain at the *IGF2* locus results from single-copy tandem duplications.

IGF2 activation through a super-enhancer mediated by *de novo* contact domain formation

We next performed ChIP-seq at the *IGF2* locus and detected an accumulation of the active chromatin mark H3K27ac at a previously identified *IGF2* enhancer[65] herein referred to as the *IGF2* cognate enhancer (**Fig. 5a** and **Supplementary Fig. 11**). An even more pronounced H3K27ac peak, however, intersected with an element previously inferred to represent a lineage-specific super-enhancer in CRC cell lines (VACO-400 and VACO-9M)[46]. To verify enhancer function, we performed luciferase assays, which revealed enhancer activity of cloned fragments of this previously inferred super-enhancer[46] in the HCT116 colon cancer line but not in a control (HeLa) cancer cell line (**Supplementary Fig. 12** and **Supplementary Table 6**). Notably, to

our knowledge this super-enhancer has not previously been reported to physically interact with or regulate *IGF2*, and it indeed may not normally have the capacity to do so, as it resides in an adjacent TAD (**Fig. 5a** and **Supplementary Fig. 11**).

The *IGF2* locus tandem duplications were observed to extend over the intervening TAD boundary and encompass this super-enhancer (**Fig. 5a** and **Supplementary Fig. 11**). We therefore used 4C-seq to investigate whether *IGF2* dysregulation could be driven by topological or contact domain reorganization. Indeed, these data revealed the lineage-specific super-enhancer as the strongest interaction partner of *IGF2* in CRCP5S and CRCP7S, with a complete absence of this interaction in control spheroids (**Fig. 5b** and **Supplementary Fig. 11**); we verified this interaction in a reciprocal 4C-seq experiment using the super-enhancer as the viewpoint (**Fig. 5c** and **Supplementary Fig. 11**). By comparison, 4C-seq reads connecting *IGF2* with its cognate enhancer were absent, indicating that *IGF2* is not activated by its cognate CRE in this context (**Fig. 5b** and **Supplementary Fig. 11**).

Our observations can be summarized in a model whereby a *de novo* 3D contact domain comprising a gene locus relevant to cancer (*IGF2*) and a super-enhancer forms in between preexisting TADs, resulting in oncogenic locus dysregulation (**Fig. 5d**). Indeed, we inferred that the tandem duplications resulted in copies of *IGF2* and the super-enhancer being positioned in a head-to-tail orientation, with each able to contact the other via chromatin looping (see our model in **Fig. 5d**). We further examined the potential of the tandem-duplication sequence to form a new contact domain by carrying out ChIP-seq of CTCF, a DNA-binding protein that resides at contact domain boundaries[36,66], and we observed increased CTCF binding consistent with boundary use (**Supplementary Fig. 11**). We also identified three larger somatic duplications of *IGF2* in the TCGA data that, on the basis of their size and location with respect to TAD boundaries, were inferred to not lead to the formation of a 3D contact domain comprising *IGF2* and this super-enhancer (**Supplementary Fig. 13**). Notably, none of these three *IGF2* duplication carriers exhibited appreciable levels of *IGF2* overexpression, and they showed significantly lower *IGF2* expression compared with tandem duplications with the potential to lead to 3D contact domain formation ($P = 0.01$, Wilcoxon rank-sum test), lending additional support to our new model. Taken together, our findings show that rather than a gene dosage increase, a hitherto undescribed mechanism—tandem-duplication-mediated *de novo* contact domain formation resulting in physical interaction between the *IGF2* promoter and a normally hidden super-enhancer—drives overexpression of *IGF2* in CRC (**Fig. 5d**).

DISCUSSION

We developed CESAM to enable systematic discovery of enhancer hijacking events in cancer genomes, and we inferred 18 candidate enhancer hijacking events in a pan-cancer analysis and 98 in tumor-type-specific analyses. Previous studies provided comprehensive views of recurrent SCNAs in cancer, and the GISTIC algorithm[23,27] has emerged as an important standard for identifying recurrent SCNAs in cancer. Our analyses using CESAM in pan-cancer and tumor-type-specific settings notably identified 16 cancer-related genes previously assigned to GISTIC peaks as CESAM hits (e.g., *IRS4* and *FAM135B*). Our data collectively suggest that activation of cancer genes by juxtaposition of CREs is a fairly common process that may be comparable to recurrent in-frame gene fusions leading to 3′ target overexpression in cancer (for example, recent work by Yoshihara *et al.*[67] uncovered 39 such events as recurrent in at least four cancer samples (a similar threshold as used in our study) in an analysis encompassing 4,300 TCGA donors).

Hits uncovered by CESAM include, to our knowledge, the first validated cases of enhancer hijacking in adult solid cancers. We provide *in vivo* evidence for a tumor-growth-promoting role of *IRS4*, a gene dysregulated in conjunction with deletions in *cis* in several cancers. The identified upregulation of *IRS4* (~400-fold overexpression in deletion carriers) in LUSC is associated with a marked gain in active chromatin marks at the gene's promoter as well as at a candidate enhancer region. Notably, our observations of a stable promoter–enhancer chromatin looping state present in both active and silent contexts show similarity to observations of gene regulation during fruit fly development, where marked changes in expression typically do not involve alterations in enhancer–promoter contacts but arise among pre-existing chromatin loops[68]. Our data are compatible with disruptions of CTCF insulators at TAD boundaries through recurring deletions[34,42], the consequence of which seems to be the spreading of active chromatin marks in the context of *IRS4* (see our model in **Supplementary Fig. 9**). Consistent with our findings, CRISPR-mediated deletion of a CTCF insulator region at the *HOX* gene cluster has recently been shown to lead to spreading of active chromatin to neighboring gene regions in embryonic stem cells[69].

Furthermore, our tumor-type-specific analyses showed that enhancer hijacking mediates gene dysregulation at the *IGF2* locus in CRC. This involves a previously undescribed mechanism whereby tandem-duplication-mediated *de novo* formation of a contact domain accompanying a super-enhancer normally inaccessible to *IGF2* results in >250-fold gene upregulation. *IGF2*, an imprinted gene[65,70,71], is associated with aggressive and chemotherapy-resistant cancer (reviewed in ref. 63), and our findings unexpectedly revealed enhancer hijacking as the dominant mechanism of high-level overexpression at this well-studied locus.

Because CESAM does not consider recurrent focal amplicons leading to locus copy numbers of four or higher, our analysis did not include super-enhancer amplification events. These have recently been shown to lead to up to fourfold overexpression of super-enhancer target genes in epithelial cancers[72]—another remarkable mechanism by which tumors can exploit the regulatory genome. As some candidates uncovered by CESAM presented with expression fold changes of >100-fold, it is tempting to speculate that enhancer hijacking may result in comparably more pronounced expression changes, possibly by providing access to otherwise inaccessible regulatory regions (**Fig. 5d** and **Supplementary Fig. 9**). Finally, we note that previously described examples of enhancer hijacking have occasionally involved balanced translocations[28,73], which are incompletely captured by SCNA profiling. In the future, similarly sized sets of whole-genome-sequenced cancer genomes with matched expression data, including those that will be provided by the Pan-Cancer Analysis of Whole Genomes initiative[74], may enable such events to be incorporated into systematic CESAM searches. Given its potential to systematically uncover enhancer hijacking events, CESAM has repercussions for the design of analysis strategies to uncover genetic driver alterations in cancer genomes.

URLs. Multiplexion, http://www.multiplexion.de/en/; Active Motif, https://www.activemotif.com; ICGC, http://icgc.org; TCGA Research Network, http://cancergenome.nih.gov.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

*Note added in proof: After our paper's provisional acceptance, a research study addressing limb malformation as a phenotype elegantly demonstrated via Hi-C data that tandem duplications intersecting with TAD boundaries can lead to de novo formation of TADs, herein referred to as neo TADs (Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature **538**, 265–269 (2016)), in the context of developmental disease. Our paper demonstrates somatic de novo contact domain formation through TAD boundary intersecting tandem duplications in cancer, and we note that the observed IGF2 de novo contact domains bear largely the same features and thus are likely to underlie somatic neo TAD formation.*

ACKNOWLEDGMENTS

This research project was funded in part through Network of Excellence funding by the European Commission (260791 to J.O.K. and A.H.), grants from the German Ministry for Science and Education (BMBF) (01KU1505F to J.O.K. and S.M.P.; 01ZX1303A to R.T. and M.P.; 01ZX1406 to M.P.), a European Research Council (ERC) Starting Grant (336045 to J.O.K.), the Danish Medical Research Council (DFF-4183-00233 to J.W.), the DFG (KFO 227/BA 4806/1-2 to C.R.B. and KFO 227/GL286/1-1 to H.G.), the Baden–Württemberg Stiftung (P-LS-ASII/33 to C.R.B. and H.G.), iMed (Helmholtz Initiative on Personalized Oncology to H.G.), the EU framework program Horizon2020 (TRANSCAN-2 ERA-NET to H.G.), and the German Cancer Aid (Colon-Resist-Net to C.R.B. and H.G.). S.M.W. received funding through an SNSF Early Postdoc Mobility Fellowship (P2ELP3_155365) and an EMBO Long-Term Fellowship (ALTF 755-2014). T.D. was supported by a scholarship from the German Cancer Research Center. S.M.D. was supported by the Heidelberg School of Oncology. The results reported here are in part based upon data generated by the TCGA Research Network ("URLs"), and we acknowledge the specimen donors as well as the research groups involved in the sampling, sequencing and processing of these data. We are grateful to the NCT Tissue Bank for providing samples in accordance with the regulations of the tissue bank and the approval of the ethics committee of Heidelberg University. We are grateful to the GeneCore, IT, mouse facility and Flow Cytometry core facilities at EMBL for excellent assistance. We thank E. Furlong and B. Klaus for valuable discussions during early stages of this project, N. Sidiropoulos for assistance with Hi-C plots, and N. Habermann for assistance with manuscript formatting and proofreading. pMD2.G and psPAX2 were gifts from D. Trono (EPFL, Lausanne, France).

AUTHOR CONTRIBUTIONS

H.G. and J.O.K. share joint senior authorship. J.W. and J.O.K. developed the CESAM methodology; J.W., A.P.D., S.M.W., T.Z., S.E. and J.O.K. carried out computational pan-cancer analysis of the TCGA public resource data set; J.W., A.P.D., B.R.M., T.D., C.R.B., H.G. and J.O.K. designed experiments; A.P.D., B.R.M., A.M.S., T.D., B.R., T.E., G.B., R.T., M.P., A.R.H., A.H., C.R.B., J.W., H.G. and J.O.K. worked on experiments in primary cancer samples; A.P.D., B.R., A.M.S., T.E., B.R.M., J.W. and J.O.K. conducted experiments in cell lines; T.D., A.P.D., B.R.M., A.M.S., S.M.D., J.W., C.R.B., H.G. and J.O.K. conducted experiments in spheroid cultures; C.S., S.M.D., C.R.B. and H.G. carried out enhancer luciferase experiments; J.W., A.P.D., B.R.M., Y.C., T.D., C.R.B., H.G. and J.O.K. prepared manuscript display items; M.S., A.H., M.P., H.B., W.W., O.T.B., P.A.N., S.M.P., I.P., S.K.S. and E.T. provided clinical information and human cancer tissue; W.W. carried out pathology analysis of spheroids; Y.C., A.P.D. and S.M.W. performed mouse experiments; and R.S., M.J., A.P.D., Y.C., J.W. and J.O.K. designed mouse experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ley, T.J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
2. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
3. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
4. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. *Cell* **148**, 59–71 (2012).
5. Jones, D.T. *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105 (2012).
6. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
7. Jones, D.T. *et al.* Recurrent somatic alterations of *FGFR1* and *NTRK2* in pilocytic astrocytoma. *Nat. Genet.* **45**, 927–932 (2013).
8. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
9. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
10. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
11. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
12. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
13. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
14. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**, R754–R763 (2010).
15. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
16. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
17. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
18. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
19. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
20. Melton, C., Reuter, J.A., Spacek, D.V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
21. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
22. Mansour, M.R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
23. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA* **104**, 20007–20012 (2007).
24. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
25. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
26. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
27. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
28. Northcott, P.A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
29. Peifer, M. *et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).
30. Valentijn, L.J. *et al. TERT* rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat. Genet.* **47**, 1411–1414 (2015).
31. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
32. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
33. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant *EVI1* and *GATA2* deregulation in leukemia. *Cell* **157**, 369–381 (2014).
34. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
35. Chen, J. & Weiss, W.A. When deletions gain functions: commandeering epigenetic mechanisms. *Cancer Cell* **26**, 160–161 (2014).
36. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
37. Nora, E.P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
38. Dekker, J. & Heard, E. Structural and functional diversity of Topologically Associating Domains. *FEBS Lett.* **589**, 2877–2884 (2015).
39. Anderson, E., Devenney, P.S., Hill, R.E. & Lettice, L.A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).
40. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390–400 (2014).
41. Waszak, S.M. *et al.* Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**, 1039–1050 (2015).
42. Lupiáñez, D.G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
43. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* **107**, 9546–9551 (2010).
44. Song, Y. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91–95 (2014).
45. Roy, N. *et al.* Brg1 promotes both tumor-suppressive and oncogenic activities at distinct stages of pancreatic cancer formation. *Genes Dev.* **29**, 658–671 (2015).
46. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).

47. Davis, C.F. *et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
48. Qu, B.H., Karas, M., Koval, A. & LeRoith, D. Insulin receptor substrate-4 enhances insulin-like growth factor-I-induced cell proliferation. *J. Biol. Chem.* **274**, 31179–31184 (1999).
49. Xia, Z., Zhang, N. & Ding, D. Proliferation and migration of hepatoblastoma cells are mediated by IRS-4 via PI3K/Akt pathways. *Int. J. Clin. Exp. Med.* **7**, 3763–3769 (2014).
50. Hoxhaj, G., Dissanayake, K. & MacKintosh, C. Effect of IRS4 levels on PI 3-kinase signalling. *PLoS One* **8**, e73327 (2013).
51. Homma, Y. *et al.* Insulin receptor substrate-4 binds to Slingshot-1 phosphatase and promotes cofilin dephosphorylation. *J. Biol. Chem.* **289**, 26302–26313 (2014).
52. Shimwell, N.J. *et al.* Adenovirus 5 E1A is responsible for increased expression of insulin receptor substrate 4 in established adenovirus 5-transformed cell lines and interacts with IRS components activating the PI3 kinase/Akt signalling pathway. *Oncogene* **28**, 686–697 (2009).
53. Lingohr, M.K. *et al.* Decreasing IRS-2 expression in pancreatic beta-cells (INS-1) promotes apoptosis, which can be compensated for by introduction of IRS-4 expression. *Mol. Cell. Endocrinol.* **209**, 17–31 (2003).
54. Mehine, M., Mäkinen, N., Heinonen, H.R., Aaltonen, L.A. & Vahteristo, P. Genomics of uterine leiomyomas: insights from high-throughput sequencing. *Fertil. Steril.* **102**, 621–629 (2014).
55. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
56. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
57. Uchida, T., Myers, M.G. Jr. & White, M.F. IRS-4 mediates protein kinase B signaling during insulin stimulation without promoting antiapoptosis. *Mol. Cell. Biol.* **20**, 126–138 (2000).
58. Hinsby, A.M., Olsen, J.V. & Mann, M. Tyrosine phosphoproteomics of fibroblast growth factor signaling: a role for insulin receptor substrate-4. *J. Biol. Chem.* **279**, 46438–46447 (2004).
59. Ahmad, I., Iwata, T. & Leung, H.Y. Mechanisms of FGFR-mediated carcinogenesis. *Biochim. Biophys. Acta* **1823**, 850–860 (2012).
60. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
61. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
62. van de Werken, H.J. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol.* **513**, 89–112 (2012).
63. Brouwer-Visser, J. & Huang, G.S. IGF2 signaling and regulation in cancer. *Cytokine Growth Factor Rev.* **26**, 371–377 (2015).
64. Li, X. *et al.* Oncogenic transformation of diverse gastrointestinal tissues in primary organoid culture. *Nat. Med.* **20**, 769–777 (2014).
65. Leighton, P.A., Saam, J.R., Ingram, R.S., Stewart, C.L. & Tilghman, S.M. An enhancer deletion affects both H19 and Igf2 expression. *Genes Dev.* **9**, 2079–2089 (1995).
66. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
67. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).
68. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96–100 (2014).
69. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015).
70. Venkatraman, A. *et al.* Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence. *Nature* **500**, 345–349 (2013).
71. Hark, A.T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486–489 (2000).
72. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.* **48**, 176–182 (2016).
73. Nambiar, M., Kari, V. & Raghavan, S.C. Chromosomal translocations in cancer. *Biochim. Biophys. Acta* **1786**, 139–152 (2008).
74. Stein, L.D., Knoppers, B.M., Campbell, P., Getz, G. & Korbel, J.O. Data analysis: create a cloud commons. *Nature* **523**, 149–151 (2015).
75. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
76. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

ONLINE METHODS

***Cis* expression structural alteration mapping (CESAM).** CESAM integrates SCNA-derived breakpoints with RNA-seq data (RSEM, for RNA-seq by expectation maximization) to identify expression changes associated with breakpoints in *cis*. SCNA (n = 10,320, SNP6-derived) and RNA-seq (n = 9,999) data (representing 27 tumor types), embargo-free, were downloaded from the TCGA data portal (15.11.2015, hg19). In total, 7,416 donors having both SCNA and expression data, and associated with 26 tumor types, were considered in our analysis (this excludes breast cancer; see below).

SCNA-derived, TAD-bound breakpoint occurrence matrix. CESAM performs linear regression of expression (molecular phenotype) on SCNA-derived breakpoint (somatic genotype) data. To identify breakpoints associated with *cis* expression, we used recently published TAD data from the IMR90 cell line[36] (mean TAD size: 830 kb). We constructed a somatic genotype matrix based on 'TAD bins' using BEDTools (v2.24.0)[77] by annotating for every sample the presence or absence of breakpoints within a TAD. For the purpose of CESAM, we defined as 'TAD bins' annotated TAD boundaries[36] extended by 50 kb on either side, allowing for flexibility in boundary precision. We then (somatically) genotyped every TAD bin (row) in every donor (column) and excluded TAD bins with fewer than four donors based on our independent filtering criteria. In extended genomic regions where adjacent TAD bins exhibit similar somatic genotypes (i.e., where donors show similar patterns of presence/absence across neighboring TADs—for example, in the presence of recurrent SCNAs harboring their breakpoints in two neighboring TADs), CESAM performs neighbor 'TAD bin merging' combining adjacent TAD bins with similar somatic breakpoint genotypes into 'meta bins' based on PLINK[78]. We triggered TAD bin merging if two adjacent TADs had midpoints ≤ 1,000 kb apart and showed a somatic genotype similarity of $R^2 \geq 0.2$. In practice, this can prevent similar somatic genotypes from being tested repeatedly in adjacent TAD bins.

RNA-seq-derived gene expression matrix. RNA-seq-derived gene expression matrices comprising RSEM values (hg19) were scaled by $\log_2$-transformation, and independent filtering[43] was used to remove genes with low expression variance (i.e., genes with variance below the 20th percentile). To alleviate the effect of gene dosage, CESAM's regression analysis adjusts for SCNAs by dividing each gene's expression (before $\log_2$ transformation) by the tumor/normal gene copy-number ratio. It is known that the relationship between signal and copy number is not linear in SNP microarrays, which are subject to saturation effects[79] that especially affect regions with high copy-number status. As this may have affected our ability to reliably identify enhancer hijacking events in such regions with CESAM's dosage-adjusted regression analysis, our independent filtering criteria further conservatively removed genes recurrently deleted or amplified to a level of four or more copies in >0.4% of samples ('4-per-mille criterion'; in very small cohorts, we used a minimum of two amplicons to trigger filtering). The 4-per-mille criterion generally calls for rounding up to the next integer. In practice, whereas genes such as *KRAS* that are frequently highly amplified become excluded from CESAM analysis as a result of this criterion, *IGF2* and *IRS4* would not be filtered even if a more stringent '2-per-mille criterion' were used (i.e., when filtering genes with high-level amplicons in >0.2% of samples).

*Regression analysis***.** The regression analysis of CESAM involves a *cis*-eQTL search with the FastQTL (v2.1) algorithm[80], which conservatively uses a relatively large (2-Mb) *cis*-window centered on the TAD's midpoint to relate TAD-binned SCNA breakpoints with expression changes. Although in practice gene expression changes are nearly always most highly associated with SCNA breakpoints residing in the same TAD, the enlarged *cis* search window can facilitate the identification of genes with expression that correlates best with breakpoint occurrence in *cis.* We performed 1,000 permutations with FastQTL for statistical inference, using default parameters[80]. To minimize the effect of confounders, we used the following covariates in the regression: (i) the total number of SCNAs for each sample, to adjust for SCNA burden effects, and (ii) principal components (PCs), based on PC analysis[81] on the somatic SCNA-derived breakpoint matrix. An optimization step was executed whereby PCs were added sequentially until the genomic inflation factor λ (calculated using chi-squared statistics[82]) was in the desired range of <2. Because we failed to reach a genomic inflation factor of <2 for breast cancer samples, we excluded this cancer type from our CESAM analysis.

*Integrative analysis and filtering of CESAM hits***.** We used an FDR of 5% using the Benjamini–Hochberg procedure and required greater than twofold expression upregulation relative to controls for reporting CESAM candidate genes. Fold change was computed as the median expression in the group of SCNA carriers compared to the median of noncarrier control donors (median values were set to a minimum value of 1 RSEM in cases where a lower median expression level was seen). Candidate genes were then additionally filtered to adjust for gene fusion events as well as previously unaccountable 'residual' gene dosage effects. For fusion gene removal, CESAM identified candidate genes showing a predominance of SCNAs at the 5′ end of the gene, which were then compared with the TCGA fusion database[67] encompassing recurrent in-frame fusions with 3′ partners leading to gene overexpression (in practice this step readily identifies known fusions, for example, of *ERG* in prostate cancer). We also performed literature searches to remove previously described putative fusion genes. To recognize residual dosage effects, CESAM applies 'population-based dosage filtering' by evaluating for each CESAM candidate gene whether expression in SCNA carriers versus noncarriers is correlated linearly with the somatic gene copy-number status. Genes significantly correlated with somatic gene copy number (linear least-squares regression, $R^2 > 0.2$ and $P < 0.05$) are removed by this population-based dosage filtering module. In practice, although CESAM's regression analysis uses RNA-seq expression values that are already adjusted for copy number, we occasionally observe residual effects of gene copy number not properly accounted for that are attributable to array saturation effects[79], which are recognized by the population-based dosage filter. To identify SCNAs juxtaposing distal CREs[13,46] for any given SCNA with two breakpoints b_1 and b_2, with b_1 being closest to the candidate gene, CESAM identifies the closest CRE proximal to b_2.

*Code availability***.** The CESAM code is available upon request.

Primary lung squamous cell cancer samples. Primary squamous cell lung cancer samples were obtained from Oslo University Hospital and from Cologne University Hospital. Informed consent was obtained from each patient with appropriate approval by the relevant review boards.

Generation and culturing of tumor-initiating-cell-enriched primary CRC spheroid cultures. Primary human CRC samples or derived metastases were obtained from Heidelberg University Hospital in accordance with the Declaration of Helsinki. Informed consent for tissue collection was received from each patient, as approved by the University Ethics Review Board. The tumor tissue was minced and enzymatically digested using dispase (Stemcell Technologies). CRC tumor-initiating cells (TICs) were enriched in spheroid cultures from primary patient tumor tissue as previously described by Dieter *et al.*[83]. In detail, the digested tissue was filtered and the single-cell suspension was cultured under serum-free conditions in advanced DMEM/F-12 medium supplemented with glucose to 0.6% (Invitrogen), 2 mM L-glutamine (Invitrogen), 4 mg/ml BSA (Sigma-Aldrich), 5 mM HEPES (Sigma-Aldrich), 4 µg/ml heparin (Sigma-Aldrich), 1% penicillin–streptomycin (Invitrogen) in ultra-low-attachment flasks with the addition of cytokines: 10 ng/ml FGF basic and 20 ng/ml EGF (R&D Systems) as previously described[83]. Cytokines were added twice a week. Depending on the patient culture, spheroids were dissociated manually by pipetting up and down 15–20 times or by treatment with accutase (PAA Laboratories GmbH) for 10–60 min. All spheroid cultures were authenticated and checked by Multiplexion ("URLs") for contamination against various species of bacteria, viruses, contaminating cell lines and murine-cell contamination.

Isolation of nucleic acids for DNA sequencing and RNA expression analysis. After review by a pathologist, 30 µg of tumor tissue was used for the extraction of nucleic acids. In addition, patient-derived TIC-enriched spheroid cells were pelleted by centrifugation (800 r.p.m., 4 °C, 5 min) and washed two times with PBS to get rid of the residual media. DNA and RNA of primary patient tissue and spheroids were isolated using the DNeasy Blood & Tissue Kit (Qiagen) and AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's instructions. RNA extracted in Oslo was isolated using Standard TRIZOL methods (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The RNA was treated with an on-column DNase I digestion protocol based on the manufacturer's instructions (Qiagen) to get rid of any residual DNA. The DNA and RNA were quantified using Nanodrop and Qubit according to the manufacturer's instructions.

Chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-seq). 10^5 to 10^7 cells were expanded, fixed with freshly prepared formaldehyde solution (11% formaldehyde (Sigma-Aldrich), 0.1 M NaCl (Sigma-Aldrich), 1 mM EDTA (pH 8) (Sigma-Aldrich), 50 mM HEPES (pH 7.9) (Sigma-Aldrich)) and agitated for 15 min at room temperature. The reaction was stopped by the addition of 1/20 volume glycine solution (Sigma-Aldrich) and subsequently incubated for 5 min. The cells were washed with PBS to get rid of any media constituents and re-suspended in 10 ml of chilled PBS-Igepal (0.5%) (Sigma-Aldrich), and then centrifuged and re-suspended in PBS-Igepal (0.5%) along with 100 µl of PMSF (1 mM) (Sigma-Aldrich). The cells were then centrifuged, the supernatant was discarded and pellets were snap-frozen on dry ice.

Samples were submitted to Active Motif ("URLs") for ChIP-seq. Active Motif prepared chromatin and performed ChIP reactions. In brief, 3D cell cultures of pediatric tumors were fixed in PBS with 1% formaldehyde for 15 min and quenched with 0.125 M glycine. Chromatin was isolated using Active Motif's proprietary buffer for low-cell-number ChIP-seq. DNA was sheared to an average length of 300–500 bp with Active Motif's EpiShear probe sonicator (53051) and cooled sonication platform (53080). For preparation of genomic DNA (Input), aliquots of chromatin were treated with RNase, proteinase K and heat for de-cross-linking and were then subjected to ethanol precipitation. Pellets were re-suspended and the resulting DNA was quantified on a NanoDrop spectrophotometer. Extrapolation to the original chromatin volume allowed quantitation of the total chromatin yield.

Chromatin was pre-cleared with protein A agarose beads (Life Technologies). Genomic DNA regions of interest were isolated using 4 µg of antibody to CTCF (Active Motif, 61311, lot 2) and H3K27me3 (Millipore, 07-449, lot 2475696). Complexes were washed, eluted from the beads with SDS buffer, and subjected to RNase and proteinase K treatment. Cross-links were reversed by incubation overnight at 65 °C, and ChIP DNA was purified by phenol–chloroform extraction and ethanol precipitation.

Illumina sequencing libraries were prepared from the ChIP and input DNAs using the standard consecutive enzymatic steps of end-polishing, dA addition, and adaptor ligation. After the final 15-cycle PCR amplification step, the resulting DNA libraries were quantified and sequenced (Illumina platform). Sequences (75 bp, single end) were aligned to the human genome (hg19) using BWA-mem (0.7.4)[84]. Duplicate reads were removed, and only uniquely mapped reads (mapping quality ≥ 25) were used for further analysis. Alignments were extended *in silico* at their 3′ ends to a length of 200 bp, which is the average genomic fragment length in the size-selected library, and assigned to 32-nt bins along the genome. Filtering and peak calling were performed using HOMER (v4.7.2)[85] with standard settings.

4C-seq library preparation and sequencing. 4C-seq libraries were prepared according to the protocol in ref. 86, with some modifications. Briefly, 10 million cells from each spheroid culture were dissociated and fixed with 2% formaldehyde. The fixed genomic DNA was digested using the *NlaIII* enzyme and subsequently self-ligated. A second digestion reaction was performed with *DpnII* and was followed by ligation. After purification of the circularized DNA, inverse PCR was performed to obtain 4C-seq libraries. 1.6 µg of template DNA was used for the amplification of the final libraries. For primary LUSC samples, cells were dissociated with 0.0125% collagenase, and nuclei were isolated and subsequently fixed with 1% formaldehyde. Because of the low amount of tissue material, the 4C-seq protocol was modified to use 1/3 of the volumes stated in the original protocol. For these libraries, 800 ng of template DNA was used for final library amplification. The reading primers (**Supplementary Table 7**) had 4–6 nt of barcode sequences to allow for de-multiplexing of pooled libraries. PCR products were purified, mixed together and sequenced on an Illumina HiSeq 2000 and an Illumina NextSeq platform in 100-bp and 75-bp paired-end read length modes, respectively. Alignment was performed using BWA-mem in single-end mode (v 0.7.4) to reference genome hg19. 4C interactions were identified using FourCSeq[87].

RT-qPCR-based expression measurements. RT-qPCR was performed to identify samples with strong overexpression of *IGF2* and *IRS4*. 35 CRC tumor sample RNAs for *IGF2* were obtained from University Hospital Heidelberg (extracted with the AllPrep DNA/RNA Mini Kit (Qiagen)), and 94 squamous cell carcinoma tumor sample RNAs for *IRS4* were obtained from Oslo University (extracted with TRIzol (Invitrogen)). Only RNA samples with RIN values > 3 and with tumor content > 30% were used. Single-stranded cDNA was synthesized from 500 ng of total RNA using the SuperScript III First-Strand Synthesis SuperMix for qRT-PCR (Invitrogen) according to the manufacturer's protocol. qPCR primers were designed using the online Primer3 Plus program[88] with the qPCR settings activated. Primer sequences are available in **Supplementary Table 8**. We tested all primers by running a standard curve and requiring the primer efficiency to be between 90% and 100% and as close as possible to that of the housekeeping primer pair. The primer efficiency was 91.3% for globulin, 91.6% for *IGF2*, and 95.6% for *IRS4*. In addition, a single and discrete peak was detected in the melt curve analysis for all primers tested. The qPCR experiments were performed on a StepOnePlus 96 Fast machine (Applied Biosystems) in 20 µl using a 96-well plate. The mastermix contained 10 µl of 2× SYBR Green PCR Master Mix (Applied Biosystems), 0.4 µl of each primer (10 µM), 2.5–5 ng of sample cDNA in 5 µl, and 4.2 µl of nuclease-free H_2O. The reaction program was run in default ramping speed mode, and cycling conditions were 10 min at 95 °C, 40 cycles of 95 °C for 15 s and 60 °C for 1 min, followed by a melting curve stage. Non-template controls were included in all experiments, replacing cDNA with H_2O, and typically resulted in no detection at all. The results were analyzed using the StepOne analysis software v2.3 (Applied Biosystems). Relative expression levels for *IGF2* and *IRS4* were calculated relative to the housekeeping gene globulin using the C_t method. Each sample was measured in technical duplicates, and the relative fold expression difference was compared to the median expression value of all samples for *IGF2* or the median of seven representative samples with expression near the technical background for *IRS4*.

Massively parallel DNA sequencing. Two types of Illumina next-generation sequencing libraries, long-insert-size paired-end mapping (mate-pair sequencing) and (regular) Illumina paired-end sequencing to 1× (low) coverage, were used to analyze somatic structural rearrangements in a locus-specific manner. In more detail, mate-pair DNA library preparation was performed using the Nextera Mate Pair Sample Preparation Kit (Illumina). In brief, 4 µg of high-molecular-weight genomic DNA was fragmented by the tagmentation reaction in 400 µl and then subjected to strand displacement. Samples were size-selected to 4–5 kb following the Gel-Plus path of the protocol. A total of 300–550 ng of size-selected DNA was circularized in 300 µl for 16 h at 30 °C. After an exonuclease digestion step to get rid of remaining linear DNA, fragmentation to 300–700 bp with a Covaris S2 instrument (LGC Genomics), and binding to streptavidin beads, the libraries were completed via end repair, A-tailing, and Illumina Truseq adaptor ligation. The final sequencing library was obtained after PCR for 1 min at 98 °C followed by nine cycles of 30 s at 98 °C, 30 s at 60 °C, 1 min at 72 °C, and a final elongation step of 5 min at 72 °C. Sequencing was carried out with an Illumina HiSeq2000 (2 × 101-bp reads) instrument using v3 or v4 chemistry to reach an average spanning coverage of 20–30×. Short-insert-size library preparation was performed using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs). Briefly, 250 ng of genomic DNA was fragmented with a Covaris S2 instrument (LGC Genomics) to 700–800 bp and then processed according to the manufacturer's protocol and sequenced in 2 × 125-bp mode[4,89]. We used an in-house Illumina HiSeq 2000 platform to sequence each library to average physical depths (spanning coverage) of 35× for mate-pair sequencing and <2× for low-coverage/short-insert-size sequencing, using 100-bp paired-end reads.

Structural variant calling was performed using the procedure described in ref. 90 by aligning reads to the hg19 reference genome assembly with BWA-mem (v0.7.4) and using DELLY2 (v0.6.8)[75] for structural variant discovery.

Identification of H3K27ac peaks with differential H3K27ac signal in candidate regions. Differential H3K27ac occupancy analysis was performed using Bioconductor, in particular the DiffBind[91] package. Briefly, LUSC (*cis* deletion carriers, $n = 3$; noncarrier controls, $n = 2$) H3K27ac peaks, as well as CRC (tandem-duplication carriers, $n = 2$; noncarrier controls, $n = 4$) H3K27ac peaks called by Homer[85] and corresponding H3K27ac ChIP-seq BAM files, were used as the input data for the analysis. Differentially bound peaks were identified with the modules dba.count, dba.contrast, dba.analyze and dba.report of the package, consecutively.

doi:10.1038/ng.3722

Specifically, we first performed an unbiased differential H3K27ac occupancy analysis comparing LUSC deletion carriers to noncarrier controls at the *IRS4* locus and in its vicinity and controlled the FDR at 5%. This analysis revealed only four peaks with differential H3K27ac signal within the relevant region shown in **Figure 3** (a nearly 1-Mb-long region that includes a TAD as well as inter-TAD space). The two peaks exhibiting the most significant differential H3K27ac signal corresponded to the *IRS4* gene itself and to the inferred novel *IRS4* enhancer, respectively (see asterisks in **Fig. 3**). In both cases H3K27ac signal was significantly higher in *cis* SCNA (deletion) carriers. The third most significant differential peak, which again exhibited more H3K27ac in SCNA carriers, localized ~20 kb upstream of *VSIG1*. However, in contrast to *IRS4*, *VSIG1* was barely expressed and its expression showed only slight increases in deletion carriers (2.7-fold for *VSIG1* versus 400-fold for *IRS4*; **Fig. 3** and **Supplementary Fig. 3**), which strongly implicates *IRS4* (rather than *VSIG1*) as the target of these recurrent *cis* SCNAs. The fourth peak, localizing at the bidirectional promoter of *COL4A5*/*COL4A6*, showed significantly less H3K27ac signal in deletion carriers, in line with promoter deletion in SCNA carriers and with the lower expression of *COL4A5* and *COL4A6* in LUSC deletion carriers (**Fig. 3**). Together with the observation that occasional locus amplifications and duplications clearly drive *IRS4* expression (**Supplementary Fig. 6**), these data nominate *IRS4* as the most plausible candidate gene becoming aberrantly activated as a consequence of recurrent SCNAs in this genomic region.

Differential H3K27ac occupancy analysis for CRC samples showing *IGF2* tandem duplication versus noncarrier controls did not reveal a single peak with differential H3K27ac signal on chromosome 11 when the FDR was controlled at 5%. By comparison, when controlling the FDR at 20%, we identified only one large peak covering *IGF2* itself and the respective TAD boundary (see the H3K27ac occupied region in **Fig. 5a**) as differentially marked with H3K27ac on chromosome 11 (which is consistent with the massive activation of *IGF2* as a consequence of recurrent locus rearrangements).

Cell line, vectors and virus preparation. The HCC-15 cell line was purchased from DSMZ and cultured in RPMI 1640 medium (Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific) and antibiotic–antimycotic (Thermo Fisher Scientific). An *IRS4*-overexpressing vector, pLenti-IRS4-Myc-DDK, was purchased from OriGene. An IRES-eGFP sequence was cloned from a pIRES2-AcGFP1 vector (Takara-Clonetech) into the pLenti-IRS4-Myc-DDK vector using the In-Fusion HD cloning kit (Takara-Clonetech) and is referred to here as pLenti-IRS4. Primers used for this purpose are shown in **Supplementary Table 8**. We created the control vector by removing IRS4-Myc-DDK by restriction enzyme digestion with EcoRI, and it is referred to here as pLenti-empty. Plasmids used for lentivirus production were pMD2.G (VSV-G envelope) and psPAX2 (second-generation lentiviral packaging plasmid), both gifts from Didier Trono (Addgene plasmids 12259 and 12260). Lentivirus production was conducted by transfection with Lipofectamine 3000 reagent (Thermo Fisher Scientific) of equal amounts of pMD2.G, psPAX2 and pLenti-IRS4-Myc-DDK-IRES-GFP/pLenti-IRES-GFP in 293FT cells (Thermo Fisher Scientific) according to the manufacturer's protocol. Cells were transduced with produced virus with the addition of 8 µg/mL polybrene (Sigma-Aldrich) by spinfection (centrifuged at 2,000 r.p.m. for 2 h) with the produced virus and were enriched by sorting according to eGFP intensity (see "Flow cytometry"). All cell lines were regularly checked for mycoplasma contamination.

Flow cytometry. Transduced HCC-15 cells were sorted for eGFP expression on a MoFloXDP cell sorter (Beckman Coulter Inc.) equipped with a Coherent Innova 90C argon ion laser (Coherent Inc.) tuned to 488 nm at 200 mW. We sorted cells using a 100-ìm nozzle while running BD FACSFlow as sheath at 20 p.s.i. and at room temperature. Forward and side scatter height and area signals were used for gating of live cells and singlets. eGFP fluorescence was detected using a 530/40-nm bandpass filter combined with a 488-nm notch filter. eGFP-positive cells were sorted in purity mode (1 drop envelope) into 6-well or 96-well dishes with culture media. To measure eGFP intensity, we ran HCC-15 cells through an LSR-Fortessa SORP instrument (BD Biosciences) with a 488-nm laser (530/30 BP). All post-acquisition analysis was done with FlowJo 10.0.8 (Tree Star, Inc.).

Immunohistochemistry. Immunohistochemistry was performed according to ref. 92. Anti-IRS4 used was purchase from Abcam (clone EP907Y, product code ab52622, 1DegreeBio ID 1DB-001-0001145254).

Mouse experiments. One million transduced HCC-15 cells were suspended in DMEM mixed 1:1 (vol/vol) with Matrigel (BD Biosciences) and subcutaneously implanted into both flanks of nude mice (Charles River Laboratories, NMRI-*Foxn1*^{nu}/*Foxn1*^{nu} (homozygous) male mice; 8 weeks old at the time of injection). The total number of tumors was $n = 8$ for each group in the first experiment (i.e., two cell line injections in each of four mice, where we performed experiments in both flanks in each mouse), $n = 9$ for control (5 mice) and $n = 12$ (6 mice) for *IRS4*-overexpressing sample in the second experiment. Although at this sample size effect sizes were not robustly estimated, differences in tumor growth became readily evident. Mice were randomly assigned into two groups, and tumor sizes were measured twice weekly in two dimensions (length and width). Tumor volumes (V) were calculated as V (cm^3) = 0.5 × (length × width^2). Mice were euthanized once the biggest tumor volume was ~2 cm^3. Mice were housed and maintained according to animal use guidelines at EMBL Heidelberg. Both mouse grouping and tumor volume measurements were blinded.

Tissue preparation for flow cytometry. Small parts of fresh tumors grown in nude mice were cut and digested in DMEM F-12 media (Lonza) with 25 mM HEPES (Gibco), 100 I.U./ml penicillin–streptomycin, 150 U/ml collagenase (Worthington Biochemical), and 20 µg/ml Liberase (Roche) at 37 °C for 3 h. Supernatants were carefully removed after the addition of D-PBS (Gibco) and centrifuged at 1,000 r.p.m. for 5 min at room temperature. Cell pellets were subsequently digested by 0.25% trypsin (Gibco) for 45 min at 37 °C and deactivated by DMEM F-12 with 25 mM HEPES, 10% FBS (Biowest) and DNase I. After centrifuging and digestion with red blood cell lysis buffer (Sigma), cells were washed twice with D-PBS containing 2% FBS and filtered by 40-µm mesh.

Luciferase enhancer assays. Enhancer regions were amplified by PCR and cloned into the luciferase reporter vector pGL4.24[luc2P/minP] (Promega) containing a multiple cloning site followed by a minimal promoter and the luciferase reporter gene. Primer sets used for amplification of several stretches of the CRC super-enhancer region are shown in **Supplementary Table 8**.

For testing enhancer region activity, HCT116 (CRC) and HeLa (cervical cancer) cell lines were plated in 96-well plates in triplicate and transfected with 50 ng of enhancer region DNA using pGL4.24 reporter vectors and 10 ng of pRL-TK renilla luciferase control plasmid. 48 h after transfection, cells were lysed and luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega). The firefly luciferase signal of pGL4.24 vectors was normalized to the renilla luciferase signal of the pRL-TK vector and displayed as fold activity normalized to the pGL4.24 empty vector control. Experiments were performed in triplicate for E5. For constructs E2, E4 and E6, triplicate experiments were performed for each of the two independent experiments.

Data availability statement. Sequence data for ChIP, whole-genome mate-pair and 4C (BAM files and processed BED files) have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002066. Publically available TCGA Research Network data are available at http://cancergenome.nih.gov/.

77. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
78. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
79. Attiyeh, E.F. *et al.* Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res.* **19**, 276–283 (2009).
80. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
81. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
82. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).

83. Dieter, S.M. *et al.* Distinct types of tumor-initiating cells form human colon cancer tumors and metastases. *Cell Stem Cell* **9**, 357–365 (2011).
84. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
85. 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**, 576–589 (2010).
86. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
87. Klein, F.A. *et al.* FourCSeq: analysis of 4C sequencing data. *Bioinformatics* **31**, 3085–3091 (2015).
88. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–W74 (2007).
89. Mardin, B.R. *et al.* A cell-based model system links chromothripsis with hyperploidy. *Mol. Syst. Biol.* **11**, 828 (2015).
90. Weischenfeldt, J. *et al.* Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159–170 (2013).
91. Stark, R. & Brown, G. *DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data* (Univ. of Cambridge/Cancer Research UK–Cambridge Institute, 2011).
92. Sotillo, R. *et al.* Mad2 overexpression promotes aneuploidy and tumorigenesis in mice. *Cancer Cell* **11**, 9–23 (2007).

doi:10.1038/ng.3722