CTCF/cohesin binding sites are frequently mutated in cancer

Authors Riku Katainen[1,2,6], Kashyap Dave[3,6], Esa Pitkänen[1,2,6], Kimmo Palin[1,2,6], Teemu Kivioja[1], Niko Välimäki[1,2], Alexandra Gylfe[1,2], Heikki Ristolainen[1,2], Ulrika A. Hänninen[1,2], Tatiana Cajuso[1,2], Johanna Kondelin[1,2], Tomas Tanskanen[1,2], Jukka-Pekka Mecklin[4], Heikki Järvinen[5], Laura Renkonen-Sinisalo[1,5], Anna Lepistö[5], Eevi Kaasinen[1,2], Outi Kilpivaara[1,2], Sari Tuupanen[1,2], Martin Enge[3], Jussi Taipale[1,3], Lauri A. Aaltonen[1,2]

Affiliations

[1]Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Finland.

[2]Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Finland.

[3]SciLife Center, Department of Biosciences and Nutrition, Karolinska Institutet, Sweden.

[4]Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland.

[5]Department of Surgery, Helsinki University Central Hospital, Hospital District of Helsinki and Uusimaa, Helsinki, Finland.

[6]These authors contributed equally to this work.

Correspondence should be addressed to L.A.A. and J.T. (lauri.aaltonen@helsinki.fi and jussi.taipale@ki.se).

Abstract

Cohesin is present in almost all active enhancer regions, where it is associated with transcription factors[1, 2]. Cohesin colocalizes frequently with CTCF (CCCTC-binding factor), affecting genomic stability, expression, and epigenetic homeostasis[3, 4, 5, 6]. Cohesin subunits are mutated in cancer[7, 8], but CTCF/cohesin binding sites (CBSs) in DNA have not been examined for mutations. Here we report frequent mutations at CBSs in cancers displaying a mutational signature where mutations in adenine-thymine base pairs predominate. Integration of whole-genome sequencing (WGS) data from 213 colorectal cancer (CRC) samples and chromatin immunoprecipitation sequencing (ChIP-exo) data revealed frequent point mutations at CBSs. In contrast, CRCs showing ultramutator phenotype caused by defects in the exonuclease domain of DNA polymerase epsilon (*POLE*) displayed a significantly fewer mutations at and adjacent to CBSs. Analysis of public data revealed that multiple cancer types accumulate CBS mutations. CBSs are a major mutation hotspot in the noncoding cancer genome.

Main

The non-coding cancer genome is largely unexplored and may harbor undiscovered somatic changes that are important for understanding how tumors arise. Examples of previous discoveries include recurrent mutations of *TERT* promoter, which create a binding motif for ETS transcription factors (TF) and thus significantly increase *TERT* transcriptional activity, as well as somatic mutations in T-cell acute lymphoblastic leukemia that introduce binding motifs for the MYB transcription factor creating a super-enhancer upstream of the *TAL1* oncogene[9, 10].

To further characterize the somatic mutation landscape beyond protein coding regions, we sequenced whole genomes of 213 colorectal cancers (CRCs) and respective normal samples with ≥40x depth of coverage. The CRC sample set consisted of 198 microsatellite stable (MSS), twelve unstable (MSI) tumors and three tumors showing ultramutator phenotype caused by mutations in the proofreading domain of *POLE* (POLε-exo^{-}). The catalogue of somatic substitutions in the MSS samples was first studied using mutation signature analysis, which revealed three distinct signatures in the tumors. The two most prevalent signatures were similar to signatures 1 and 17 identified by Alexandrov, L. B. *et al.*[11]. Signature 1 is strongly correlated with age at diagnosis, whereas signature 17 is characterized by A:T→C:G and A:T→G:C substitutions[11] (**Supplementary Fig. 1, 2a; Supplementary Table 1**). The latter signature has previously been associated with hepatic, esophageal, and gastric cancers, and B-cell lymphoma[11]. The third signature identified was unrelated to previously described signatures.

We have recently found that cohesin is present at almost all genomic sites where multiple transcription factors are located, and suggested that cohesin could play a role in inheritance of accessible chromatin across the cell cycle. We also found that ChIP-seq signal for the leading strand DNA-polymerase ε was decreased at cohesin positions, suggesting that POLε does not replicate cohesin sites[2]. This prompted us to examine mutation patterns in the POLε-exo^{-} samples at sites that bound both the cohesin subunit RAD21 and CTCF, a DNA-binding protein that can load cohesin to DNA. For this purpose, we determined the exact genomic binding sites of these proteins, as well as the transcription factors KLF5, MYC, MAX, and REST, with chromatin immunoprecipitation followed by exonuclease treatment (ChIP-exo) high-resolution method in a CRC cell line LoVo. A total of 28,331 CTCF binding sites also occupied by cohesin (CBSs) were discovered genome-wide (**Supplementary Table 2**), spanning a total of 1,171,396 bp. Compatible with the decreased signal in our earlier POLε ChIP-seq data, mutation frequency at and adjacent to CBSs in the POLε-exo^{-} tumors was much lower than in flanking regions (**Fig. 1a**, **Supplementary Fig. 3**). This decrease was highly significant from 400 bp regions around CBSs to 800+800 bp flanking regions (Wilcoxon p=2.41x10^{-24}), and was not explained by differences in sequence context between CBSs and their flanking regions. This observation supported the notion that DNA at CBSs is often replicated by a polymerase other than POLε[2].

Next, we analyzed somatic mutations at CBSs in 198 MSS samples, which revealed a pattern dramatically different from that in the POLε-exo^{-} tumors. The MSS samples showed an inverse pattern, where the somatic mutation frequency was highly increased at CBSs compared to regions flanking the binding sites (1,000 bp both in 5’ and 3’ directions, Wilcoxon p=2.21x10^{-6}) (**Fig. 1b, d**). The overall mutation frequencies at CBSs exceeded genome-wide frequencies by an order of magnitude (Wilcoxon p=3.08x10^{-34}, median mutation frequency at CBSs 8.75x10^{-6} and genome-wide 8.49x10^{-7}). Moreover, CBS substitutions, consisting predominantly of A:T→C:G and A:T→G:C mutations, strongly associated with mutational signature 17 (Spearman ρ=0.54, p=2.47x10^{-16}, **Supplementary Fig. 2b**). To determine whether the observed signal was due to the sequence context at

CBSs, we computed context-specific mutation frequencies expected under signature 17. The observed rate of A:T→C:G and A:T→G:C substitutions was found to be significantly higher than the expected rate (**Supplementary Fig. 4** and **Supplementary Table 22**). While in POLε-exo^{-} tumors the reduced relative mutation frequency at or close to CBSs was observed in a window of ~400 bp, MSS tumors accumulated mutations specifically at or immediately adjacent to the 17 bp CTCF binding motif (**Fig. 1d**). A total of 1,966 somatic mutations were found in 1,553 out of 28,331 CBSs (**Fig. 1b, d** and **Supplementary Fig. 5**; median 7 mutations per sample, range: 0–128). To put this finding in context, the mutation frequency at CBSs was approximately 1.5 times higher than in the compiled set of known cancer mutation target genes (547 genes of the cancer gene census; mutation frequency in 198 CRCs at CBSs $8.98\text{x}10^{-6}$ and in census $6.10\text{x}10^{-6}$).

We next examined whether accumulation of mutations at CTCF binding sites required co-localization of cohesin complex, as determined by the ChIP-exo data of CTCF and RAD21. We found that CTCF binding sites lacking RAD21 co-localization did not display an increased mutation frequency (Wilcoxon p=0.29, **Supplementary Fig. 6**). We did not observe accumulation of mutations around RAD21 peaks in the absence of an adjacent CTCF signal in ChIP-exo data (Wilcoxon p=0.72, **Supplementary Fig. 6**), indicating that CTCF presence is required for the mutational process. Furthermore, higher CTCF motif affinity associated with CBS mutations (negative binomial regression $p<1.687\text{x}10^{-58}$), supporting this interpretation. In contrast to MSS samples, MSI samples displayed a slight decrease of mutation frequency at CBSs (Wilcoxon $p=2.36\text{x}10^{-3}$, mutation frequency $4.10\text{x}10^{-5}$ at CBSs and $4.57\text{x}10^{-5}$ in flanking sequence) (**Fig. 1c**). Finally, we detected no change in mutation frequency at the binding sites of the other TFs studied, KLF5, MYC, MAX and REST in the MSS CRCs (**Supplementary Fig. 7**).

The relative lack of CBS mutations in POL*ε*-exo^{-} CRCs suggests that the mechanisms affecting CBS mutation frequencies in CRCs might be associated with DNA replication. We next studied whether the emergence of signature 17 and CBS mutations in MSS CRCs was associated with replication timing[12]. Similarly to what was observed in HeLa cells[13], high mutation frequency in MSS samples associated with later replication timing (**Fig. 2a, b**). In particular, substitutions at A:T base pairs occurred preferably in genomic regions that are replicated later in the S-phase (**Fig. 2b**). Compatible with these observations, those CBSs which replicated later harbored more mutations than the sites which replicated early (negative binomial regression $p<4.492\text{x}10^{-49}$). Finally, a significant proportion of signature 17 mutations exhibited low allelic fractions suggesting an ongoing mutational process, resembling clonality pattern seen in MSI CRCs due to a persisting MMR defect (**Supplementary Fig. 8**).

To further investigate the significance of the observed mutations at CBSs, we searched for clusters of somatic point mutations in the MSS samples. We focused on the regulatory genome defined as the genomic regions where binding of at least three out of 380 TFs was identified in the CRC LoVo cell line using ChIP-seq (hereafter termed LoVo-regions), spanning ~7% of the genome (~206 Mbp)[2]. In addition, we examined regions predicted to be enhancers, promoters or CTCF binding sites in ENCODE, spanning ~10% of the genome (~300 Mbp)[14]. A total of 67 and 102 somatic mutation clusters (>5 mutations/100 bp, >4/198 mutated samples) were identified in LoVo- and ENCODE-regions respectively, containing 755 mutations in total. A major proportion of the mutation clusters in regulatory genome was explained by CBS mutations (34/67 (51%) in LoVo-regions and 34/102 (33%) in ENCODE-regions) (**Supplementary Table 3**). CBSs thus comprised by far the most frequently mutated compact feature of the regulatory genome.

We validated the highest ranking CBS mutation cluster (**Supplementary Table 3**, cluster id 6814) by Sanger sequencing, and screened the locus in a validation set of 913 CRC samples, resulting 30 additional mutations. The changes aggregated in three hotspots on A:T bases of the CTCF motif (**Fig. 3**).

To detect possible associations between mutations in key CRC genes as well as clinical features and the frequency of mutations seen at CBSs, we performed multiple regression analysis. We observed 1.4-fold increase in the fraction of substitutions occurring at CBSs associating with *TP53* mutation positivity (p=2.34x10^{-4}, 95% CI [1.18, 1.72]). An effect of similar magnitude but lower statistical significance was observed with *BRAF* mutation positivity (1.5 fold change, p=0.047, 95% CI [1.01, 2.11]). In contrast, mutations in *FBXW7* associated with reduced mutation frequency at CBSs (0.7, p=0.005, 95% CI [0.55, 0.90]). Higher age at diagnosis contributed a one percent increase per year to the number of observed CBS mutations (p=0.005, 95% CI [1.00..1.02x]).

In contrast to MSS CRCs, MSI CRCs mutated C:G bases more commonly both at CBSs and genome-wide, and did not exhibit signature 17 in the same degree. Similarly, germline variants of the 1000 Genomes Project showed an enrichment of C:G substitutions and a slight reduction in variants observed at CBSs (**Supplementary Fig. 9**), the latter probably due to evolutionary selection pressure against germline CBS mutations.

Finally, we scrutinized CBSs in the catalogue of somatic mutations of 16 whole-genome sequenced tumor types available through the International Cancer Genome Consortium (ICGC)[15]. Mutations at CBSs - similar to those observed in MSS CRCs - were frequently encountered in several cancer types (**Fig. 4**, **Supplementary Figs. 10** and **11**). In many tumor types, a smallish subset of tumors often contributed a major proportion of CBS mutations (**Supplementary Fig. 11**). This was particularly true for virus-associated hepatocellular carcinomas, where 7/208 tumors displayed a remarkably high mutation frequency with a uniform base change pattern at CBSs. Those gastrointestinal cancer samples contributing most to the CBS mutation load showed proportionally more A:T→C:G and A:T→G:C substitutions than the other four substitution types (**Fig. 4**, **Supplementary Figs. 10** and **11**). As in MSS CRC, these mutations were compatible with signature 17, appearing often in CpTpT contexts.

Here we report two novel mutation phenomena occurring at CTCF binding sites colocalized with the cohesin complex. In POL*ε*-exo^{-} samples we observed a relative reduction in mutations caused by the defective exonuclease function, indicating that these sites are not primarily replicated by POL*ε*. In contrast, in many cancers with functional POL*ε* we found a highly significant increase in mutation frequency at CBSs.

The unimodal distribution of CBS mutations across samples suggests a multifactorial and/or environmental basis for the mutations. Five of our findings provide some clues to the factors contributing to the CBS mutations. First, the distribution of the observed somatic CBS variants is very different from that seen in germline, and thus they are likely to arise under aberrant conditions rather than normal cellular homeostasis. Second, the CBS mutations tended to accumulate in late replicating regions. Third, the reduced CBS mutation frequency in POLε-exo^{-} samples suggests that the CBS sites represent a special region in DNA replication, such as the origin[16]. In such a case, aberrations in factors associated with replication initiation would be a candidate mechanism for CBS mutations. Fourth, we associated a mutation signature preferring A:T→C:G and A:T→G:C substitutions to occurrence of CBS mutations. All gastrointestinal tumor types studied here provided examples of cases with striking accumulation of CBS mutations at A:T base pairs. Thus, it is possible that environmental exposures contributing to gastrointestinal cancer underlie the mutations. While the underlying factors in some somatic mutation signatures are known[17], to our knowledge signature 17 has not yet been explained[18]. Mutagenic factors such as nucleotide pool imbalance as well as oxidative stress are credible candidates[19]. Fifth, mutations in *TP53* were associated to increased substitutions in A:T base pairs, at CBSs and genome-wide. TP53 is involved in maintenance of genomic stability, and defects in the protein may thus contribute to the mutation load at CBS sites.

We find here that CBSs display striking accumulation of mutations in multiple different cancer types. The mutations occur either at the CTCF site, or immediately adjacent to it, in a highly stereotypic pattern. Challenged DNA replication may underline the mutations.

While most of the CBS mutations are likely to be passengers, many are predicted to affect CTCF binding affinity (**Supplementary Fig. 12**). Some of these may drive tumorigenesis by causing cellular defects such as aberrant gene expression, epigenetic changes, as well as genetic instability[20]. Our work identifies a completely new and unexpected class of cancer mutations, and calls for vigorous efforts to elucidate its causes and consequences.

Accession codes

International Cancer Genome Consortium (ICGC) somatic mutation data was obtained from ICGC data repository (https://dcc.icgc.org/repository/release_16/; release 16, May 15, 2014).

Acknowledgements

We thank Sandra Augsten, Lijuan Hu, Heikki Metsola, Sini Nieminen, Alison Ollikainen, Marjo Rajalaakso, Sirpa Soisalo, Inga-Lill Svedberg, Mikko Turunen, Iina Vuoristo and Anna Zetterlund for technical assistance. Further, we thank Mervi Aavikko, Iikki Donner, Auli Karhu and Miika Mehine for additional help with the manuscript. We acknowledge the International Cancer Genome Consortium (ICGC) for the somatic mutation data. This work was supported by grants from the Academy of Finland (Finnish Center of Excellence Program 2012–2017, #250345, personal grants #137680 and #274474 for O.K.), the University of Helsinki (personal grant for S.T.), the Finnish Cancer Society, European Research Council (ERC, #268648), the Sigrid Juselius Foundation, SYSCOL (an EU FP7 Collaborative Project, #258236), the Finnish Cancer Society (Personal grant to K.P.), Nordic Information for Action eScience Center (NIASC); a Nordic Center of Excellence financed by NordForsk (Project number 62721, personal grant to K.P.) and State Research Funding of Kuopio University Hospital (B1401). We acknowledge the computational resources provided by the ELIXIR node hosted at CSC - IT Center for Science funded by the Academy of Finland (grants 271642, 263164), CSC and the Ministry of Education and Culture, Finland.

Author Contributions

R.K., K.P., E.P. performed the somatic mutation analyses; K.D., M.E. performed the ChIP-seq and ChIP-exo experiments; R.K., T.K., K.P., E.P. performed the sequence motif analyses; E.P., R.K., H.R. performed the analyses of ICGC data; U.H., T.C. J.K., T.T., S.T., O.K. prepared the colorectal cancer samples; J.-P.M., H.J., L.R.-S. A.-L. contributed the patient samples; E.K., N.V. contributed to the data analysis; E.P., R.K. performed the primary sequence analysis; L.A.A., J.T. designed and supervised the study. R.K., K.D., E.P. and K.P. contributed equally to the study. All authors have contributed to writing of the manuscript.

References

1. Kagey, M.H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).
2. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
3. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422–433 (2008).
4. Rubio, E.D. *et al.* CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8309–8314 (2008).
5. Wendt, K.S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).
6. 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).
7. Kon, A. *et al.* Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nature Genetics* **45**, 1232–1237 (2013).
8. Leiserson, M.D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, doi:10.1038 (2014).
9. Huang, F.W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
10. Mansour, M.R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346,** 1373–1377 (2014).

11. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

12. Chen C.L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research* **20**, 447–457 (2010).

13. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

14. Hoffman, M.M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research* **41**, 827–41 (2013).

15. Hudson, T.J. *et al.* International network of cancer genome projects. *Nature.* **464**, 993–998 (2010).

16. Guillou, E. *et al.* Cohesin organizes chromatin loops at DNA replication factories. *Genes & Development* **24**, 2812–2822 (2010).

17. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* **15**, 585–598 (2014).

18. Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genetics* **45**, 478–486 (2013).

19. Satou, K., Kawai, K., Kasai, H., Harashima, H. & Kamiya, H. Mutagenic effects of 8-hydroxy-dGTP in live mammalian cells. *Free Radical Biology and Medicine* **42**, 1552–1560 (2007).

20. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017-1021 (2015).

Figure 1: Total number of somatic substitutions occurring at 28,331 CBSs (39 bp). Somatic substitutions at CBSs with a flanking sequence of 1,000 bp both in 5’ and 3’ directions in **(a)** three POLε-exo^{-}, **(b)** 198 MSS and **(c)** 12 MSI CRC samples, respectively. **(d)** Total number of somatic substitutions in 198 MSS CRC samples at 28,331 CBSs and a flanking sequence of 30 bp both in 5’ and 3’ directions. The six substitution types are indicated by colors in the histogram. CTCF motif sequence logo is shown underneath the histogram (black lines highlight the 17 bp core motif).

Figure 2: Somatic substitution rates and replication timing in chromosome 1. Substitution rates in those 20 of 198 MSS CRCs that displayed **(a)** lowest and **(b)** highest signature 17 exposure. Substitution rates in both groups are similar, except for the high rates of A:T→C:G and A:T→G:C substitutions seen in (**b**) but not in (**a**). In both groups, substitution rates correlate strongly with replication timing. **(c)** Substitutions occurring in three POLε-exo^{-} tumors display a clear correlation with replication timing, while **(d)** MSI CRCs show a distinct pattern in C:G→T:A substitutions.

Figure 3: Somatic mutations at one of the most recurrently mutated CBSs (chr6:73,122,088–73,122,127 GRCh37). This site was mutated in 39 of 1,111 (3.5%) CRCs evaluated by Sanger sequencing. Mutations were clustered at adenine bases within the 17 bp CTCF core motif (black rectangle). Base-level conservation (GERP) shown underneath the sequence logo.

Figure 4: Sample-wise somatic substitutions across all 28,331 CBSs in MSS CRCs, virus-associated hepatocellular carcinomas (LIRI-JP, ICGC) and MSI CRCs.

(a) The 40 tumors that harbor the highest number of CBS substitutions are shown both for MSS CRC (198 tumors in total) and hepatocellular carcinoma (208 tumors). CBS substitutions for the other ICGC data are shown in **Supplementary Figs. 10** and **11**. **(b)** The relative fraction of all somatic substitutions hitting CBSs in each sample. Seven hepatocellular carcinomas show a very high fraction of total mutation load occurring at CBSs, exceeding 1% in two cases.

Online Methods

Sequencing and primary sequence analysis.

The colorectal cancer (CRC) sample set consisted of a total of 213 matched normal-tumor pairs. It included 198 microsatellite-stable (MSS), 12 microsatellite-unstable (MSI) and three POLε-exo^{-} tumors. The study has been reviewed and approved by the Ethics Committee of the Hospital district of Helsinki and Uusimaa (HUS). Signed informed consent or authorization from the National Supervisory Authority for Welfare and Health has been obtained for all the sample materials used.

Whole-genome sequencing of the 213 CRCs and respective normals was carried out with (Illumina) HiSeq 2000 as Illumina service using a paired-end sequencing protocol. Read length was 100 bp. The median sequencing coverage at non-N reference was >40x.

Paired-end reads were mapped against the 1000 Genomes Phase 2 reference assembly hs37d5 using BWA[21] (version 0.6.2) with parameters -n 0.06 and -q 5. PCR duplicates were removed using samtools (version 0.1.18) rmdup. Local realignment around suspected indel sites and base score quality recalibration were performed using GATK IndelRealigner and BaseRecalibrator (GATK version 2.3.9)[22]. These steps resulted in final sets of analysis-ready mapped reads that were used in somatic substitution, indel and structural variant calling.

Somatic mutation analysis.

Somatic substitutions in each tumor were called using MuTect (version 1.1.4) with default parameters[23]. In order to improve calling specificity, MuTect was not run on regions covered by Duke excluded regions track or HiSeqDepth top 5% track (both tracks obtained from the UCSC Genome Browser). To call somatic indels, we used GATK SomaticIndelDetector (GATK version 2.3.9) on the same genomic intervals that were used in MuTect analysis[22]. We used raw MuTect substitution calls for mutation frequency calculation at CBSs, in replication timing and mutational signature analyses.

We performed mutation cluster analyses in regulatory regions using an in-house genome analysis tool (Katainen, R. *et al.*, manuscript under preparation). Two filtering phases were used to reduce the number of false positives in MuTect calls. First, only mutations occurring at strict accessibility regions (1000 Genomes Project) were retained. Second, we filtered somatic calls against a pooled set of whole-genome sequences from ten blood samples by excluding any somatic call which was found in three or more reads in the pooled data.

The regulatory genome was defined as genomic regions covered by at least three peaks in the ChIP-seq data from the LoVo cell line[2] and with the promoter, enhancer and CTCF regions predicted with six different cell lines in ENCODE data[15]. The total size of the regulatory genome defined with LoVo- and ENCODE-regions was ~206 Mbp (~7% of whole genome) and ~300 Mbp (~10%), respectively. The overlap between LoVo- and ENCODE-regions were ~150 Mbp. CBSs covered ~1.17 Mbp (**Supplementary Table 2**), which is (~0.5%) of LoVo- and (~0.4%) of ENCODE-regions.

Mutation signature analysis.

We performed signature analysis using non-negative matrix factorization of six substitution types in 5'-Xp[A/T]pY-3' context as described in [24] on a discovery set of 92 MSS CRCs. In addition to the method of Alexandrov, L. B. *et al.*, we used bootstrap sampling to improve modeling of population level mutation signatures and to discount the effect of sample data (the particular set of 92 individuals that have been sequenced). In particular, we calculated the Non Negative Matrix Factorization repeatedly for 92 CRCs sampled uniformly, with replacement, from the discovery set. The final signatures were extracted as in Alexandrov, L. B. *et al*. This process yielded three mutational signatures. The obtained signatures were compared to the published signatures of Alexandrov, L. B. *et al.* by mean Kullback-Leibler Divergence $[D_{KL}(p||q) + D_{KL}(q||p)]/2$.

We computed the exposure of each signature in 198 MSS, twelve MSI CRCs and three POLε-exo^{-} samples as projection of the mutation matrix to the signature weight matrix. These exposures were normalized by the number of somatic mutations in each tumor. This normalized quantity is used throughout the manuscript and supplement (**Supplementary Table 1**).

The expected mutation distribution, given the underlying sequence, was computed by summing over the signature weights for the particular nucleotide triplet. The mutation distribution was median scaled in a 10 kbp window to match the median mutation count in the region.

ChIP-exo.

LoVo (ATCC, cat. no. CCL229TM) cells were cultured in DMEM supplemented with 10% fetal bovine serum (FBS) and antibiotics. All antibodies used in ChIP-exo experiments were ChIP-grade. In each experiment a non-specific IgG was used as control. ChIP-exo was performed as described in [25] using antibodies for Rad21, CTCF, KLF5, HNF4A, REST, mouse and rabbit IgGs, MYC and MAX (Santa Cruz Biotechnology cat. no:s sc-98784, sc-15914X, sc-22797X, sc-8987X, sc-25398, sc-2025, sc-2027, 06-340- Millipore antibody, ab80336- Abcam antibody, respectively) with the following modifications: Cultured cells (~10^7 cells) were crosslinked by 1% formaldehyde for 10 minutes at room temperature. Cells were incubated in hypotonic buffer for 15 minutes and then DNA was sonicated to 200–500 bp fragments in lysis buffer (50 mM HEPES, pH 8.0; 2 mM EDTA, pH 8.0; 150 mM NaCl; 1% Triton X-100; 0.1% sodium deoxycholate; 0.2% SDS). After pre-clearing, lysate was subjected to immunoprecipitation (IP) overnight with the antibodies indicated (~2.5 x 10^6 cells in 1 ml lysate / IP). The antibodies were then precipitated using protein G sepharose beads (40 µl) for 3h at 4°C. Antibody concentration was used as per manufacturer's instructions. Immunoprecipitates were washed successively with 0.5 ml of IP buffer (100 mM NaCl, 5 mM EDTA, 0.33% (w/v) SDS and 1.5% Triton X-100 in 50 mM Tris-Cl, pH 8.0), 1 ml of Mixed micelle buffer (150 mM NaCl, 5 mM EDTA, 5.2% sucrose, 1.0% Triton X-100 and 0.2% SDS in 20 mM Tris-Cl, pH 8.0), Buffer 500 (250 mM NaCl, 25 mM HEPES, 0.5% Triton X-100, 0.05% sodium deoxycholate and 0.5 mM EDTA in 5 mM Tris-Cl, pH 8.0), LiCl/Detergent buffer (250 mM LiCl, 0.5% Igepal 630, 0.5% sodium deoxycholate and 10 mM EDTA in 10 mM Tris-Cl, pH 8.0), TE buffer (1 mM EDTA in 10 mM Tris-Cl, pH 8.0) and finally with 10 mM Tris-Cl (pH 7.5, 8.0 or 9.2 as per requirement of different enzymatic reaction steps).

Immunoprecipitates were subjected to consecutive enzymatic on-bead reactions as described below. After each on-bead reaction, the beads were washed successively with Mixed micelle buffer, Buffer 500, LiCl/Detergent buffer, TE buffer and Tris-Cl buffer.

Final reaction volume for each reaction was adjusted to 60 µl. Different on bead reactions are as follows:

1) End polishing with 3U of T4 DNA polymerase (New England Biolabs, cat. no. M0203L), 1× NEBuffer 2, 150 µM dNTPs (Thermo Fisher Scientific, cat. no. R0193), 100 µg/ml BSA. Incubated at 12°C for 30 min. Final wash with Tris pH 7.5;

2) Kinase reaction with 10U of T4 polynucleotide kinase (New England Biolabs, cat. no. M0201L), 1× T4 DNA ligase buffer. Incubated 30 mins at 37°C. Final wash with Tris pH 8.0;

3) 'A' addition reaction with 5U of Klenow fragment exo^{-} (New England Biolabs, cat. no.M0212L), 1× NEBuffer 2, 100 µM dATP (GE Healthcare, cat. no. 28406503). Incubated at 37°C for 30 min. Final wash with Tris pH 7.5;

4) P2 adapter ligation with 1.25 µM of P2 Adapters (**Supplementary Table 4**, rows 1 and 2; Eurofins MWG Operon), 500U T4 DNA ligase (New England Biolabs, cat. no. M0202L), 1× T4 DNA Ligase buffer. Incubated at 25°C for 30 min and then at 16°C overnight. Final wash with Tris pH 7.5;

5) Filling-in reaction with 10U of phi29 DNA polymerase (New England Biolabs, cat. no. M0269L), 1× phi29 DNA polymerase buffer, 200 µg/ml of BSA, 150 µM of dNTPs. Incubated at 30°C for 20 min. Final wash with Tris pH 9.2;

6) λ exonuclease reaction with 10U of lambda exonuclease (New England Biolabs, cat. no. M0262L), 1× lambda exonuclease buffer. Incubated at 37°C for 30 min. Final wash with Tris pH 8.0;

7) RecJf exonuclease reaction with 30U of RecJf exonuclease (New England Biolabs, cat. no. M0264L), 1× NEBuffer 2. Incubated at 37°C for 30 min. Final wash with TE, pH 8.0. Immunoprecipitates were eluted in 400 µl of elution buffer (10 mM Tris; 1 mM EDTA, pH 8.0; 400 mM NaCl; 1% SDS and 70 µg/ml of RNaseA) and crosslinks reversed by addition of 20 µg of Proteinase K (Thermo Fisher Scientific, cat. no. EO0491) and incubation at 65°C overnight.

Samples were then extracted with Phenol:Chloroform:Isoamyl alcohol (25:24:1 v/v), precipitated with ethanol and then processed for library preparation. Samples resuspended in 20 µl of Tris-Cl, pH 8.0. Second strand synthesis was performed by primer extension by addition of 1µM of Primer P2 (**Supplementary Table 4,** row 3; Eurofins MWG Operon) to samples, after which the samples were denatured at 95°C for 5 min, followed by incubation at 58°C for 5 min and cooling to room temperature. Primer extension reaction was performed by addition of 10U of phi29 polymerase in phi29 DNA polymerase buffer (to final concentration 1×), BSA (100 µg/ml) and dNTPs (75 µM each) followed by incubation at 30°C for 20 min. Subsequently, the enzyme was heat inactivated at 65°C for 10 min. Double stranded DNA was purified using Agencourt AMPure magnetic beads (Beckman Coulter, cat. no. A63881) and eluted into 40 µl of 10 mM Tris-Cl, pH 8.0. To increase ligation efficiency an ‘A’ addition reaction was performed using 5U of Klenow fragment exo^{-} in 1× Klenow buffer with 100 µM of dATP. Samples were then incubated at

37°C for 30 min and DNA purified using Agencourt AMPure magnetic beads, followed by elution into 40 µl of 10 mM Tris-Cl, pH 8.0. Second adapter ligation reaction was then performed using 500 U of T4 DNA Ligase, 1× T4 DNA Ligase buffer and 0.4 µM of P1 Adapter (**Supplementary Table 4**, rows 4 and 5; Eurofins MWG Operon) followed by incubation at 25°C for 30 min and then at 16°C for overnight. DNA was purified using Agencourt AMPure magnetic beads and eluted into 30 µl of 10 mM Tris-Cl, pH 8.0. Library was PCR amplified with PCR primer sequences provided by Illumina (PE primers **Supplementary Table 4**, rows 6 and 7; Eurofins MWG Operon). PCR mix contained 2U of Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, cat. no. F-530S), 1× of HF phusion polymerase buffer, 0.5 µM of each of the primers and 250 µM dNTPs in final volume of 50 µl. PCR was carried out for 18–20 cycles. For size selection, 200–600 bp PCR products were gel purified from 2% agarose by QIAquick gel purification columns (Qiagen). Purified product was quantified using NanoDrop 2000 (Thermo Fisher Scientific Inc.) and 2100 Bioanalyzer (Agilent) and sequenced at Karolinska High Throughput Center using Illumina HiSeq 2000, according to manufacturer's instructions.

Sequence reads were mapped to the human reference genome (hg18), using bwa with default parameters. For peak-calling, we used GEM[26] with default parameters, and the genome size set to 2,700,000,000. The resulting peak summits were lifted over to GRCh37 (hg19) reference genome.

Sequence motif analysis.

ChIP-exo data from CTCF and RAD21 proteins were used to determine CBSs in the reference genome. CBSs were selected from CTCF binding coordinates where RAD21 was colocalized within 100 bp. Also, CTCF motif hit was required alongside ChIP-exo peak. A total of 28,331 such CBSs were identified genome wide (overlapping sites were counted as one). The 39 bp CTCF motif (17 bp core and 22 bp flanks) was generated from 40-bp-long sequences centered at CTCF ChIP-exo peak summits using the multinomial method[27] with 17 bp seed sequence CCACAAGATGGCAGCAG from reference genome locus chr6:73,122,088–73,122,127 (GRCh37) (**Fig. 3**) and maximum Hamming distance 3. Position frequency matrix for CTCF motif provided in **Supplementary Table 3**. For other TFs (HNF4A, KLF5, MAX and REST) sequence motif hits in the human genome was queried with SELEX matrix[28]. ChIP-exo peak was required alongside the motif hit, except for HNF4A, where ChIP-seq peaks were used. As expected, most of the somatic mutations in CRCs that occurred at conserved bases of CTCF motif decreased binding affinity (**Supplementary Fig. 12**).

Replication timing analysis.

We used unfiltered MuTect somatic substitution calls in global replication timing analysis. Replication timing for all substitutions was determined with the data produced by Chen, C. L. *et al.*[12]. The data consisted of estimated replication timing for autosomes at resolution of 100 kbp.

Structural variant data analysis.

Structural variants (deletions, inversions, tandem duplications and translocations) were called in the final set of mapped reads in each normal and tumor sample independently using DELLY (version 0.0.9)[29]. Structural variants were called only in regions not covered by Duke excluded regions and the HiSeqDepth top 5% track. Further, SV calling was performed only in regions that were uniquely alignable in the CRG Alignability 100-mers track (all data obtained from the UCSC Genome Browser). To identify somatic structural aberrations, SV calls in each tumor were filtered against the calls of the same variant type (e.g., deletion) in the respective normal sample. A tumor SV call was called somatic only if its breakpoints did not occur within 1,000 bp of a breakpoint of a normal SV of the same type.

No change in the number of somatic DNA breakpoints in 10 kb and 100 kb flanks around mutated CBSs was observed for the CRC samples in whole-genome sequencing (Mann-Whitney p=0.92 and p=0.99, respectively) nor SNP chip data (p=0.99 and p=0.99).

Scrutiny of CBS mutations in ICGC genomes.

We counted somatic mutations hitting CBSs in the cancer mutation data available through the ICGC[15]. A total of 28,331 CBSs of length 39 bp were inspected in a targeted analysis. Comprehensive somatic mutation data from a total of 11 tumor types and 1,073 tumors were available containing 5,642 substitutions at the CBSs. The Data Release 16 (15th May 2014) from the following cancers were used in the study (ICGC project identifier given in parentheses): early onset prostate cancer (EOPC-DE), esophageal adenocarcinoma (ESAD-UK), hepatic cancer (LICA-FR, LINC-JP, LIRI-JP), malignant lymphoma (MALY-DE), ovarian cancer (OV-AU), pancreatic cancer (PACA-AU, PACA-CA), renal cancer (RECA-EU) and thyroid cancer (THCA-SA). Tumors with less than ten CBS substitutions were excluded from the analysis.

Sanger validation of CBS mutations.

All nine mutations identified in the WGS data at one CBS in chromosome 6 (chr6:73,122,088–73,122,127, GRCh37; **Supplementary Table 3**, cluster id 6814) were confirmed by Sanger sequencing in tumor and normal DNA of the WGS samples, and further validated in an extended set of 913 additional tumors. A total of 30 additional mutations (in 30 of 913 cancers, 3.3%) were detected in the extended set (**Fig. 3**). Sequencing reactions were carried out using the Big Dye Terminator v.3.1 kit (Applied Biosystems, Foster City, CA, USA) and electrophoresis was run on 3730xl DNA Analyzer (Applied Biosystems) at FIMM Technology Center, Finland. The sequence traces were analyzed with the Mutation Surveyor software (version v4.0.6, Softgenetics, State College, PA, USA).

Statistical analyses.

Mutation counts at CBSs and genome-wide were modeled using negative binomial regression with mutation status of key CRC genes (*APC, ARID1A*, *BRAF*, *FBXW7*, *KRAS*, *NRAS*, *PIK3CA*, *SMAD4*, *TCF7L2* and *TP53*), clinical features including gender, Dukes stage, age of diagnosis, survival in months, tumor gradus and tumor location in the colon, and the relative exposures of the three mutational signatures as covariates (Supplementary Note). Moreover, substitutions occurring at specific CBSs were similarly modelled using a negative binomial model with strand orientation, CTCF motif affinity and replication timing as covariates.

We tested whether there was a difference in the number of breakpoints around mutated and non-mutated CBSs in both WGS and SNP chip data using Mann-Whitney U-test. In both

WGS and SNP chip data, 10 kb and 100 kb flanks were tested (WGS: 10 kb flanks, Mann-Whitney U=2.9x10^8, two-tailed *p*=0.92; 100 kb flanks, U=2.9x10^8, *p*=0.95; SNP chip: 10 kb flanks, U=2.9x10^8, *p*=0.99; 100 kb flanks, U=2.9x10^8, *p*=0.99).

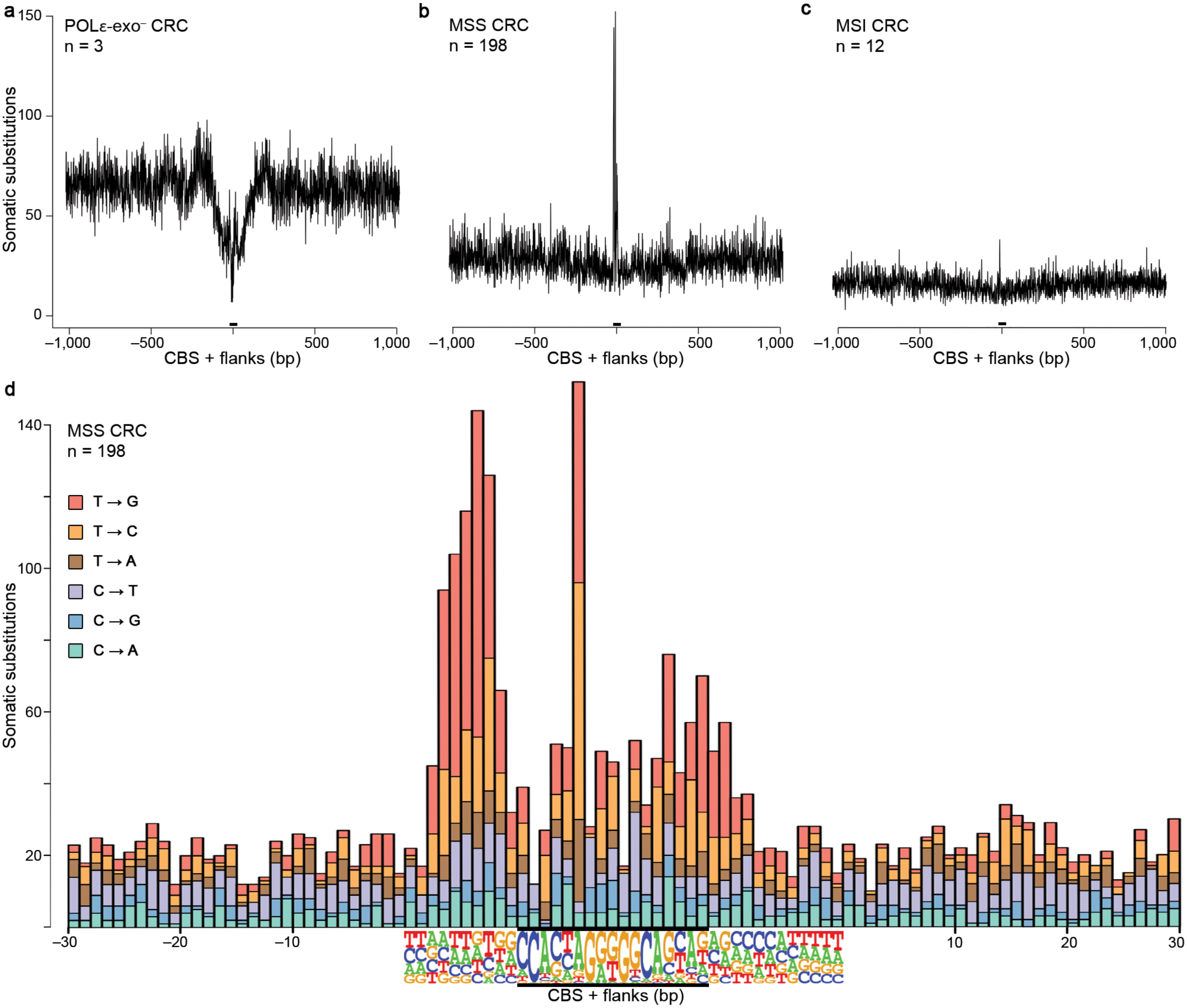
Method-only references

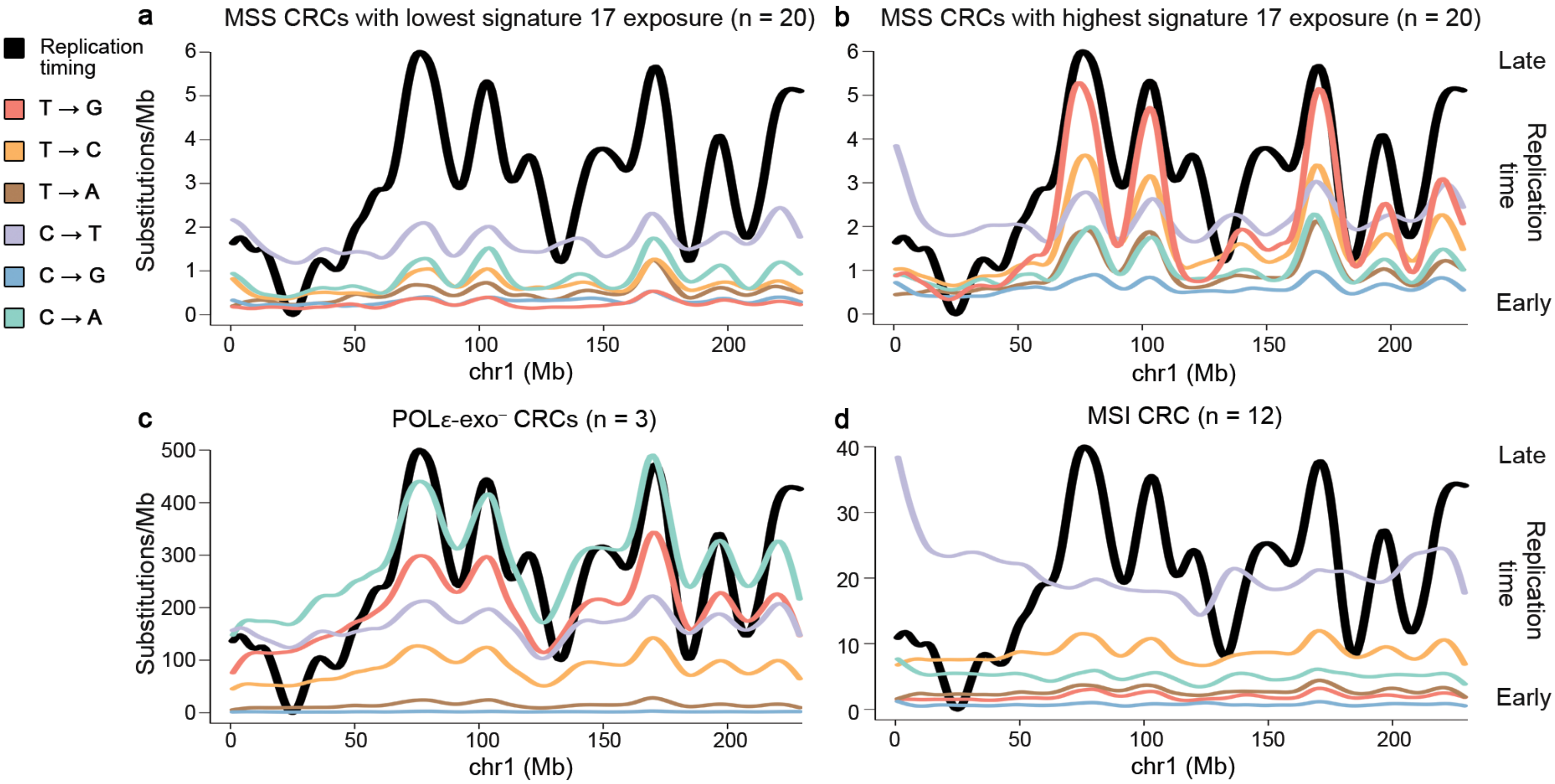
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
22. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).
23. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).
24. Alexandrov, L.B., Nik-Zainal, S., Wedge D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).
25. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408 (2011).
26. Guo, Y. *et al.* Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26**, 3028 (2010).
27. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861–873 (2010).
28. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).

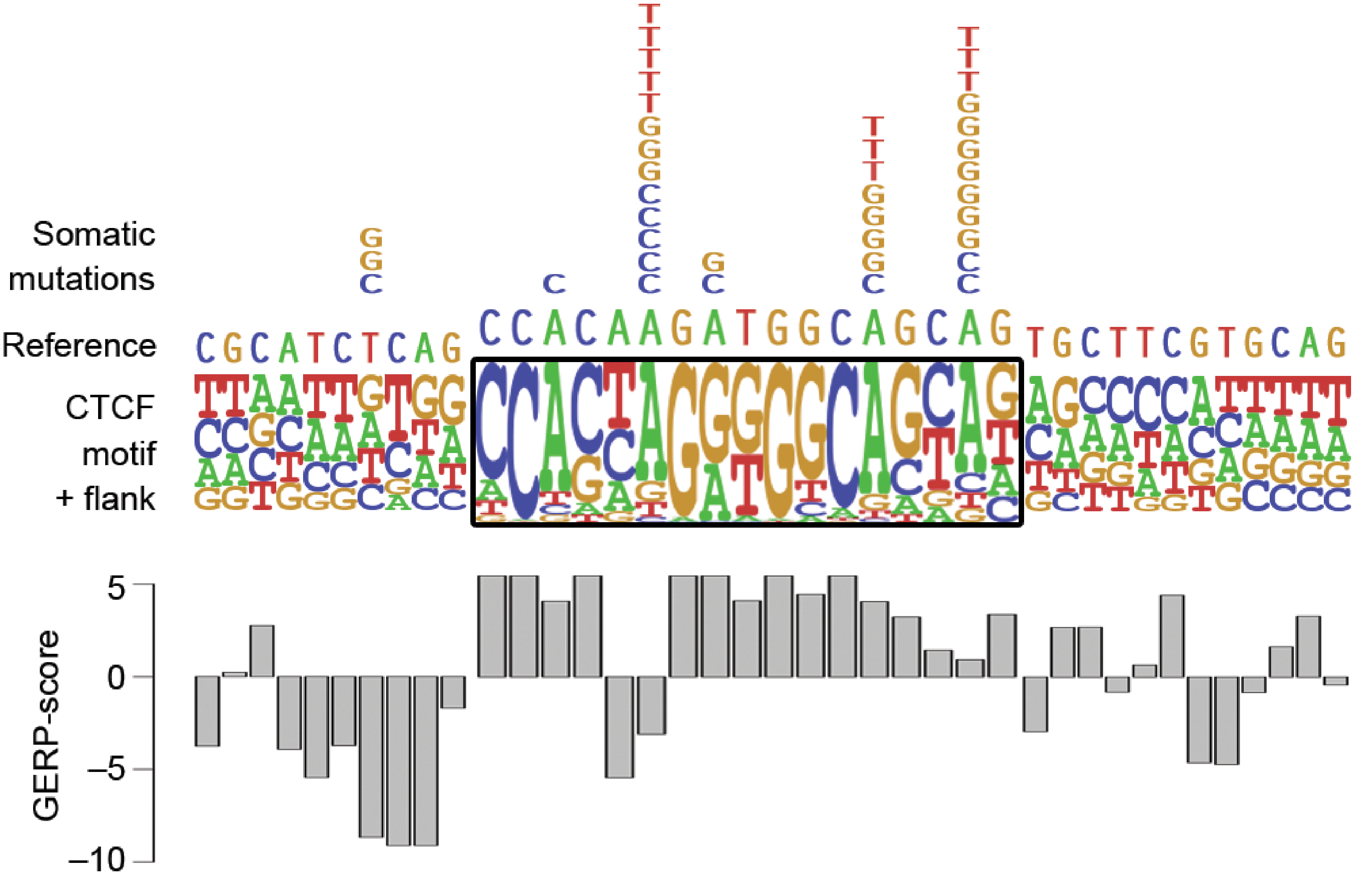
29. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

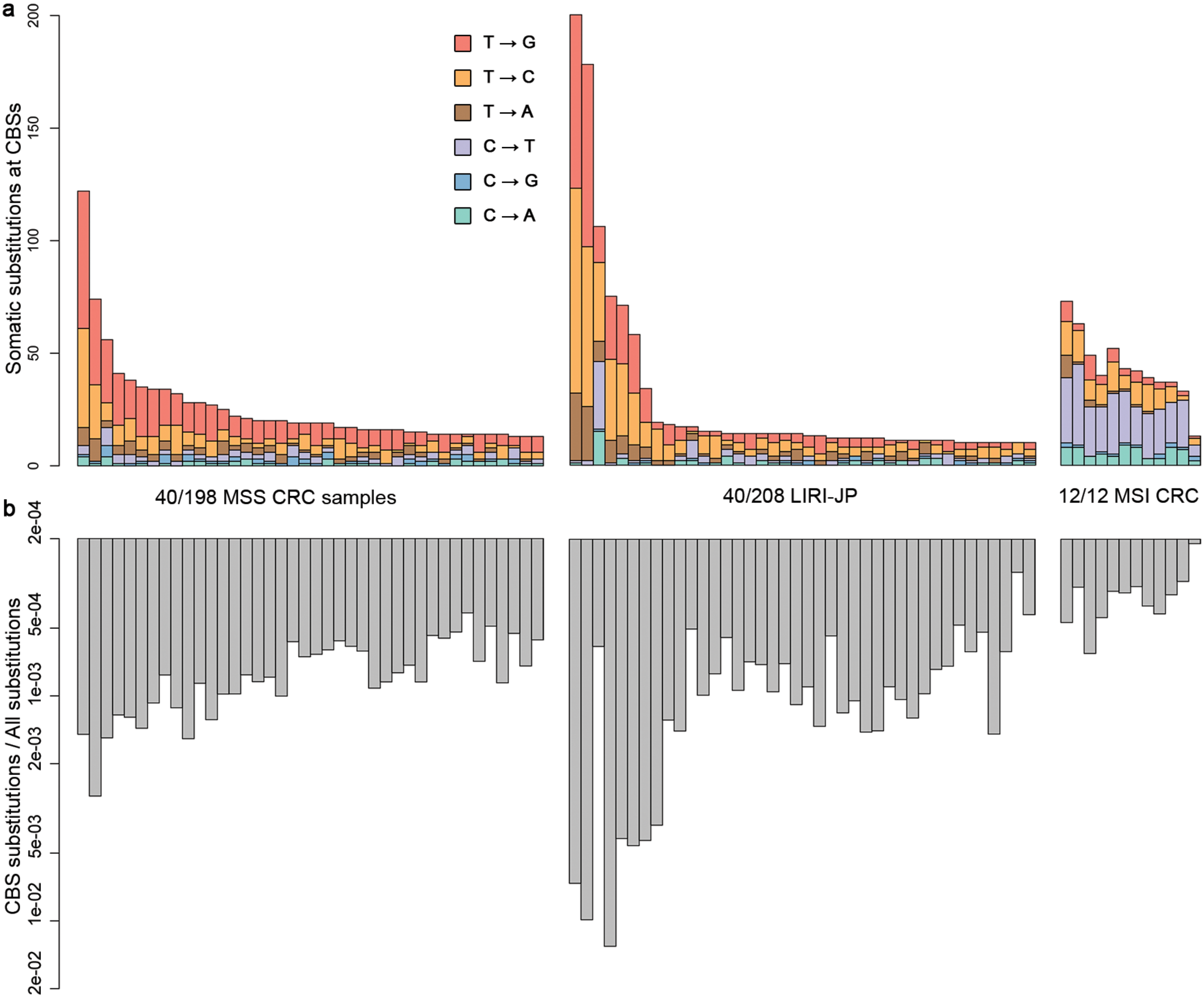
Competing financial interests

The authors declare no competing financial interests.


a
POLε-exo⁻ CRC
n = 3
Somatic substitutions
150
100
50
0
−1,000
−500
500
1,000
CBS + flanks (bp)
b
MSS CRC
n = 198
−1,000
−500
500
1,000
CBS + flanks (bp)
c
MSI CRC
n = 12
−1,000
−500
500
1,000
CBS + flanks (bp)
d
MSS CRC
n = 198
T → G
T → C
T → A
C → T
C → G
C → A
Somatic substitutions
140
100
60
20
−30
−20
−10
10
20
30
CBS + flanks (bp)


■ Replication timing
🟥 T → G
🟧 T → C
🟫 T → A
🟪 C → T
🟦 C → G
🟩 C → A
a
MSS CRCs with lowest signature 17 exposure (n = 20)
b
MSS CRCs with highest signature 17 exposure (n = 20)
c
POLε-exo⁻ CRCs (n = 3)
d
MSI CRC (n = 12)
Substitutions/Mb
0
1
2
3
4
5
6
0
100
200
300
400
500
0
10
20
30
40
0
50
100
150
200
chr1 (Mb)
Late
Replication time
Early


Somatic mutations
Reference
CGCATCTCAG
CCACAAGATGGCAGCAG
TGCTTCGTGCAG
CTCF motif + flank
GERP-score
5
0
−5
−10


a
b
Somatic substitutions at CBSs
0
50
100
150
200
T → G
T → C
T → A
C → T
C → G
C → A
40/198 MSS CRC samples
40/208 LIRI-JP
12/12 MSI CRC
CBS substitutions / All substitutions
2e-04
5e-04
1e-03
2e-03
5e-03
1e-02
2e-02