bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The presence of copy number variants in specific topologically associating domains has prognostic value in many cancer types

Lifei Li[+], Nicolai K. H. Barth[+], Christian Pilarsky[#] and Leila Taher[+*]

[+] Division of Bioinformatics, Department of Biology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

[#] Department of Surgery, Friedrich-Alexander-Universität Erlangen-Nürnberg and Universitätsklinikum Erlangen, Germany.

* To whom correspondence should be addressed. E-mail: leila.taher@fau.de

Running Title

TADs have prognostic value in cancer

Keywords

Topologically-associating domains (TADs), copy number variants (CNVs), Lasso Cox regression analysis, cancer, biomarkers

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The human genome is organized into topologically associating domains (TADs), which represent contiguous regions with a higher frequency of intra-interactions as opposed to inter-interactions. TADs contribute to gene expression regulation by restricting interactions between regulatory elements, and their disruption by genomic rearrangements can result in altered gene expression and, ultimately, in cancer. Here, we provide a proof-of-principle that mutations within TADs can be used to predict the survival of cancer patients. For this purpose, we first constructed a set of 1,467 TADs representing the three-dimensional organization of genome across 24 normal human tissues. We then used Cox regression analysis to assess the prognostic value of the TADs in different cancer types, and identified a total of 35 TADs that were prognostic for at least one of nine cancer types. Interestingly, only 46% of the prognostic TADs comprised one or more genes with a known causal association with cancer. Moreover, for those TADs encompassing such a gene, the prognostic effect of the TAD was only directed related to the presence/absence of mutations in the gene in 13% of the cases. These observations indicate that the predictive power of a large proportion of the prognostic TADs is independent of whether pan-cancer genes are mutated or not. Furthermore, 34% of the 35 prognostic TADs showed strong structural perturbations in the cancer genome, which might mediate cancer development and progression. This study has important implications for the interpretation of cancer-related non-coding mutations and offer insights to new strategies for personalizing cancer medicine.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The need for an efficient high-level organization of the genomic DNA, especially in a spatial sense, is obvious given the enormous amount of information stored in the eukaryotic genome. Thus, the human genome is three-dimensionally organized into topologically associated domains (TADs), which are hundreds of kilobases to megabases (Mb) in size and encompass multiple genes and regulatory elements (1). TADs were originally defined based on Hi-C interaction matrices as genomic blocks exhibiting preferential physical interactions within them in contrast to between them (2). Binding sites for the insulator protein CCCTC-biding factor (CTCF) and the protein complex cohesin were later found to be enriched in TAD boundary regions (TBRs) and to contribute to the confinement of chromatin loops within the TADs (3). This organization into domains and loops is thought to serve as foundation for the interaction of regulatory elements, which in turn mediates gene expression, e.g., between promoters and enhancers. Supporting this assumption, TADs have been shown to play a crucial role in gene regulation during evolution and development (4,5). Comparative analysis of Hi-C data from different species and cell types has demonstrated that TADs exhibit a high level of conservation across cell types and even across species (2,6-8). However, TADs can be subject to changes, especially in the context of disease and cancer (9,10).

Large-scale structural variants (SVs) such as (structural) copy number variants (CNVs) have been reported to drive tumorigenesis by changing the number of copies of entire genes, most commonly through dosage effects (11). Thus, many frequently mutated genes serve as effective biomarkers for cancer diagnosis, prognosis and clinical management (12,13). CNVs in non-coding regions can also be pathogenic, for example, by altering the number of gene regulatory elements or affecting their interactions (14,15). Furthermore, CNVs can disrupt TADs (16,17), causing disease. While the fusion of adjacent TADs may allow enhancers from neighboring TADs to activate oncogenes (“enhancer hijacking”), the fragmentation of a TAD into sub-domains may insulate promoters and enhancers and prevent their interactions, resulting in gene dysregulation. For instance, *TAL1* and *PDGFRA*, two oncogenes associated with lymphoblastic leukemia and gliomas, respectively, have been reported to be activated through the fusion of adjacent TADs (e.g., (18,19)). Also, in prostate cancer cells, SVs were shown to fragment a TAD containing the *TP53* tumor suppressor gene into two smaller TADs, which successively resulted in the dysregulation of a number of genes (20).

Currently, clinical genomics analyses focus on protein-coding regions of the genome. Despite the progress achieved by international efforts such as ENCODE and FANTOM in the last decade, the interpretation of non-coding variants – and hence, their clinical application – remains challenging. Novel paradigms are needed to realize the full potential of genomic information in healthcare. Here, we show how the presence or absence of CNVs in TADs can be used to predict and explain patient outcome. To this end, we utilized data from “The Cancer Genome Atlas” (TCGA, (21)) to identify TADs enriched for cancer-related CNVs. Based on these TADs, we successfully modelled survival in 19 out of 25 cancer types using LASSO Cox regression models. To obtain an insight into the mechanisms by which variants involving the TADs that were identified as prognostic may lead to cancer, we separated the prognostic TADs into two groups, depending on their conservation in the cancer genome. We found that a considerable fraction of prognostic

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TADs have been disrupted in cancer. Our results have important implications for the interpretation of cancer-related non-coding mutations and could motivate new strategies for personalizing cancer medicine.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Topologically associating domains (TADs) maps

TAD maps for the human genome (hg38) were downloaded from http://promoter.bx.psu.edu/hi-c/publications.html (accessed on August 8, 2018, Supplementary Table S1). The underlying Hi-C data were generated in the course of six different studies (2,7,22-24), but processed with the same pipeline (2). These TAD maps represent the three-dimensional organization of the genome in 24 human healthy ("normal") tissues/cell lines and 11 cancer cell lines.

TAD size comparison between normal and cancer states

We computed the median size of the TADs identified in each tissue/cell line, and then the median of the medians across i) normal tissues/cell lines, and ii) cancer cell lines.

Similarity between TAD maps of different tissues/cell lines

Given a pair of TAD maps *A* and *B*, we searched for TADs in *B* overlapping each TAD in *A*; if one or more overlaps were found in B for a TAD in *A*, we recorded the TAD covering the largest fraction of the TAD in *A*; if no overlaps were found, we recorded that as 0. The median across all TADs in *A* was used to represent the similarity between *A* and *B*. Note that this definition of similarity is not symmetric in *A* and *B*.

Construction of consensus TADs

Contribution of each tissue/cell line in total gene expression divergence

We obtained total RNA-seq read counts from the ENCODE data repository (https://www.encodepsroject.org/, accessed on August 28, 2018, (25)) for each of the tissues/cell lines corresponding to the TAD maps of interest. For those tissues/cell lines for which total RNA-seq data were not available, we used poly-A RNA-seq (see Supplementary Table S2). We combined the data for all protein-coding genes reported by the ENCODE project (https://www.encodeproject.org/search/?type=Experiment&status=released&assay_title=total+RNA-seq&award.project=ENCODE&assembly=GRCh38, accessed on August 29, 2019; (25)) and transformed the count data using the vst() function with parameter “blind=TRUE” (for a fully tissue/cell line unaware transformation) in the DESeq2 R/Bioconductor package (version 1.16.1, (26)). We next ranked the genes according to their maximum absolute deviation from the median expression value across all tissues/cell lines and selected the 2,000 genes with the highest ranks. Then, we computed the Pearson correlation coefficient between the expression profiles of these 2,000 genes of all pairs of tissues/cell lines and hierarchically clustered the tissues/cell lines based using average linkage. The resulting dendrogram represents the relationships among the tissues/cell lines based on their gene expression profiles (see Supplementary Figure S1 and S2). We then applied BranchManager (BM, (27)) to compute a *weight* that summarizes the contribution of each tissue/cell line to the total gene expression divergence according to the topology and branch length of the dendrogram. BM assumes that the differentiation of the tissues/cell lines is a Brownian process in which each tissue/cell line can be regarded as an endpoint. Based on this, it infers the differentiation trajectories of the tissues/cell lines and each tissue/cell line is assigned a weight equal to its

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

contribution to the inference. Specifically, BM computes the trajectories from the centroid to all the tissues/cell lines in the N-dimensional space in which the dendrogram can be embedded, where N is the number of tissues/cell lines in the analysis.

Conservation scores

Each TAD map was assigned one of the aforementioned weights according to the tissue/cell line from which the underlying Hi-C data originated. Two different TAD maps were available for the cell lines GM12878 and K562; in both cases, each of two TAD maps was assigned half of the weight computed for the corresponding cell line. The TAD maps and their weights were used to compute a consensus TAD map representing the majority of the tissues/cell lines considered. Specifically, we computed a TAD "conservation" score and a TAD "boundary" score for 40 kb-long sliding windows across the entire genome. Given n TAD maps, we defined the conservation score of the i^{th} window of length L (here, $L = 40000$) as:

$$c_i = \frac{1}{L} \sum_{j=i}^{L+i-1} \sum_{k=1}^{n} w_k \cdot I_T(j,k)$$

where w_k is the weight for the k^{th} TAD map and

$$I_T(j,k) = \begin{cases} 1, & \text{if the } j^{th} \text{ nucleotide of the window is within a TAD in the } k^{th} \text{ TAD map} \\ 0, & \text{otherwise.} \end{cases},$$

and the "boundary" score as:

$$b_i = \sum_{k=1}^{n} w_k \cdot I_B(i,k)$$

where

$$I_B(i,k) = \begin{cases} 1, & \text{if the } i^{th} \text{ window includes a TAD boundary in the } k^{th} \text{ TAD map} \\ 0, & \text{otherwise.} \end{cases},$$

The windows were shifted by 1000 bp (i.e., 97.5% overlap between adjacent sliding windows). The TAD conservation and boundary scores ranged from zero to one.

Finally, we merged all adjacent windows as long as i) the nucleotide-wise average TAD conservation score was ≥ 0.5, and ii) none of the windows had a TAD boundary score ≥ 0.5. The resulting genomic regions were defined as consensus TADs if their size was ≥ 40 kb. A region between two adjacent consensus TADs was considered as a TBR if its length was ≤ 400 kb and as an "unorganized chromatin region" otherwise (as in (2)).

Constitutive and non-constitutive TADs

We divided the normal consensus TADs into three groups, according to both the fraction of their sequence overlapping with a cancer TAD and the fraction of the sequence of the corresponding cancer TAD overlapping with it. Specifically, for each normal TAD $\boldsymbol{N_i}$, we recorded the fraction of its sequence overlapping with a cancer TAD. If it overlapped with two or more cancer TADs, we recorded the highest

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fraction. In turn, we examined the corresponding cancer TAD C_i, and recorded the fraction of its sequence overlapping with a normal TAD. Analogously, if the cancer TAD overlapped with two or more normal TADs, we recorded the highest fraction. Let the corresponding normal TAD be N_j. If $N_i = N_j$ we called the relationship between N_i and C_i "reciprocal". Then, we defined a) (normal) TADs with reciprocal fractions ≥ 0.95 as "constitutive" TADs; b) (normal) TADs with fractions ≤ 0.7 and/or for which the corresponding cancer TAD had a fraction ≤ 0.7 as "non-constitutive" TADs; and c) (normal) TADs that did not satisfy any of the above criteria as "ambiguous" TADs.

Enrichment analysis of CTCF peaks and house-keeping genes (HK genes) in consensus TADs and TBRs

CTCF ChIP-seq datasets matching 17 of the normal and cancer tissues and cell lines used to construct the consensus TADs were obtained from the ENCODE data repository (https://www.encodeproject.org/, accessed on August 28, 2018, (25)). 6,290 house-keeping genes were downloaded from https://www.tau.ac.il/~elieis/HKG/, accessed on August 28, 2018, (28). CTCF and house-keeping gene densities were computed with the "reference-point" sub-command of the computeMatrix command from deepTools (version 3.1.2, (29)) with parameters --referencePoint center -a 200000 -b 200000 --binSize 5000 --missingDataAsZero. The plotProfile command was used for visualization.

Cancer-related copy number variants (CNVs)

We collected masked somatic copy number variation (CNV) for the 25 different types of cancer with at least 100 primary tumor samples in The Cancer Genome Atlas (TCGA) database (see Supplementary Table S3). We only considered the CNVs longer than 1 kb and up to 10 Mb with a segment mean larger than 0.1 or smaller than -0.1 and a number of probes of at least 10.

TADs enriched/depleted for CNVs

For each consensus TAD we selected a random region from the human genome with the same size. The procedure was repeated 1000 times. We then computed the number of patients exhibiting CNVs in that particular TAD and compared it to those computed for its random counterparts. A patient was considered to exhibit a CNV in a TAD if the CNV overlapped with the TAD by at least 1bp. A TAD was considered enriched for CNVs if at most 5% of its random counterparts displayed a number of patients greater than or equal to that observed for the TAD. A TAD was considered depleted for CNVs if at most 5% of its random counterparts displayed a number of patients smaller than or equal to that observed for the TAD.

Pan-cancer genes

717 pan-cancer genes were downloaded from the COSMIC database ("Cancer Gene Census" file; release v89, https://cancer.sanger.ac.uk/cosmic, last accessed on May 15, 2019, (30)). The TADs were defined as comprising pan-cancer genes if the pan-cancer genes overlapped with TADs with at least 1bp.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Functional and Pathway analysis

Functional and pathway analysis was performed with the Database for Annotation and Integrated Discovery (DAVID Bioinformatics Resources 6.8, http://david.abcc.ncifcrf.gov/, (31,32)) In particular, we focused on the ontologies KEGG_PATHWAY, Biological Processes (BP), Molecular Function (MF) and Cellular Component (CC). Terms associated with a false discovery rate (FDR) ≤ 0.05 were considered significantly enriched.

Survival analysis

Overall survival was analyzed using Cox regression, Kaplan–Meier curves and log-rank tests.

TAD- and pan-cancer gene-based overall survival Cox regression models

Least absolute shrinkage and selection operator (LASSO) Cox regression analysis (33) was used to construct TAD- or gene-based models for predicting overall patient survival and to identify prognostic TADs/pan-cancer genes. In particular, we trained and tested models for the 19 cancer types in the dataset that were represented by at least 100 patients and had at least 10% of lethal outcomes using TADs/pan-cancer genes, age and sex as predictors or *features*. Overall patient survival was defined as the time to death or to the last follow-up date of the patient. The analysis was conducted with functions of the glmnet R package (34).

The patient samples of each cancer type were randomly separated into a training (2/3) and a test (1/3) set. We then identified TADs significantly enriched for CNVs in the genomes of the patients in the training set and recorded the presence (encoded as 1) or absence (encoded as 0) of CNVs for each of those patients in the identified TADs. In addition, we encoded the age of the patient as 1 if it was larger than the median age in the training set, and 0 otherwise; and the sex of the patient as 1 for females, and 0 for males. Finally, we performed a LASSO Cox regression analysis on the TADs enriched for CNVs, age and sex. To choose the value of the LASSO regularization parameter (λ), we performed a 5-fold cross-validation using the cv.glmnet() function with parameters 'family="cox", alpha=1, nfold=5' and computed the value of λ leading to the minimum prediction error ("lambda.min"). We then extracted the coefficients of the model trained with the glmnet() function with parameters 'family="cox", alpha=1' for this particular value of λ, sorted them based on their absolute values in decreasing order, and selected the top ten. In cases in which cv.glmnet() failed to converge (the maximum number of iterations was the default, 100,000), we selected the top ten features based on the glmnet() ranking according to their appearances in the regularization path. The performance of the model implicitly defined by these coefficients was evaluated on the test set using the c-index (35). The entire procedure above was repeated 1,000 times. Thus, for each cancer type, we trained and tested 1,000 models.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Prognostic TADs and pan-cancer genes

To select the features to be included in the final model, which was trained on all available patients, we applied an approached similar to that of Laimighofer et al (36). First, for the i^{th} model we computed a weight:

$$w_i = \begin{cases} \frac{1}{K} exp\left(log2x \frac{devCI_i}{0.1}\right), & \text{if } CI_i \geq 0.5 \\ 0, & if CI_i < 0.5 \end{cases}$$

where K is the total number of models (here, $K = 1000$) and $devCI_i = CI_i - \frac{1}{K}\sum_{j=1}^{K} CI_j$. Next, we normalized the weights w_i to add up to one:

$$w_i' = \frac{w_i}{\sum_{j=1}^{K} w_j}.$$

Then, we defined the aggregated normalized weight of the j^{th} feature as:

$$P_j = \sum_{i=1}^{k} w'_i \cdot I\,(j, i)$$

where $I(j, i) = 1$ if the j^{th} feature was selected in the i^{th} model, and $I(j, i) = 0$ otherwise.

The ten features with the highest aggregated normalized weights were selected to train the final Cox regression model. In turn, the features of the final model were selected using backward stepwise elimination. Features with regression coefficients significantly (P-value ≤ 0.05) different from zero (Wald test) were considered prognostic.

Patient stratification

The hazard ratios derived from the final Cox regression model were used to calculate a “risk score” for each patient. The risk score was defined as the sum of the weighted values for the prognostic features, with the weights being the coefficients estimated by the Cox regression model. The patients were separated into high- and low-risk groups according to the median risk score across all patients.

Kaplan-Meier (KM) survival analysis was performed to compare the overall survival of different groups of patients. The log-rank test was used to determine differences between the groups; a P-value ≤ 0.05 was considered significant. These analyses were performed with the ggsurvplot() function in the survival R package (37), with default parameters.

CNV densities in constitutive and non-constitutive TADs

CNV densities in all (normal) constitutive, all (normal) non-constitutive TADs and two subsets of (normal) non-constitutive TADs were computed with the “scale-regions” sub-command of the computeMatrix command from deepTools (version 3.1.2, (29)) with parameters “--averageTypeBins median -m 2000000 -a 1000000 -b 1000000 -bs 10000 –missingDataAsZero”. The plotProfile command was used for visualization. Specifically, we extracted i) (normal) non-constitutive TADs that are split into multiple TADs in the cancer genome such that two or more of the corresponding cancer TADs overlap with original (normal) TAD by at least 95% of their sequence; and ii) (normal) non-constitutive TADs that are fused together into one TAD in

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the cancer genome, with two or more of such TADs overlapping by at least 95% of their sequence with the corresponding cancer TAD.

Expression levels of *NOCR2* in SARC patients

Fragments Per Kilobase of transcript per Million mapped reads (FPKM) for *NOCR2* for all TCGA SARC patients were downloaded from OncoLnc (http://www.oncolnc.org/, accessed on September 18, 2019, (38)).

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

1,467 topologically associating domains (TADs) construct a consensus TAD map in human genome

TADs are known to be highly conserved in different tissues (2,6-8). To verify this, we examined 24 different maps of topologically associating domains (TADs) inferred for different human tissues from Hi-C data (see Materials and Methods). The TAD maps had a median of 1,676 TADs and a median TAD size of 1.12 Mb. In addition, we observed a generally high similarity between the TAD maps of most pairs of tissues (see Materials and Methods and Figure 1A).

To construct a single consensus TAD map representing the most prevalent features of the three-dimensional organization of the human genome, we examined the fraction of original TAD maps in which a given genomic region was comprised within a TAD. Specifically, we first calculated a “conservation score” and a “boundary score” within 40-kb windows across the entire genome (see Materials and Methods). While the former estimates the probability of the nucleotides in the window being comprised within a TAD, the latter estimates the probability that the window encompasses one or more TAD boundaries. Since the tissues considered cannot be considered independent and this is expected to be reflected in the three-dimensional organization of their genomes, we computed the conservation and boundary scores as weighted averages, with weights accounting for the relatedness between the tissues of interest and describing their fractional contribution to the total transcriptome diversity (see Materials and Methods). Finally, we merged adjacent or overlapping windows as long as i) their conservation score was greater than or equal to 0.5 and ii) their boundary score was lower than 0.5 (see Figure 1B and Materials and Methods). With this approach we inferred 1,467 consensus TADs.

The genomic regions between the TADs are traditionally separated into either topological boundary regions (TBRs) or disorganized chromatin regions, depending on their size, with the former being smaller than the latter ((2), see Materials and Methods). The consensus TADs had median size of 1.48 Mb and were separated by 1,368 TBRs and 76 disorganized chromatin regions (with a median size of ~80 kb and ~600 kb, respectively). The formation and stability of TBRs has been associated with the presence of binding sites for CCCTC-binding factors (CTCF) and transcription start sites (TSS) of house-keeping (HK) genes (2). Indeed, we observed that binding sites for CTCF and TSSs of HK genes were enriched at TBRs compared to TADs (see Figure 1C-D and Materials and Methods), supporting our definition of consensus TADs.

6% of consensus TADs are enriched or depleted for cancer-related CNVs

Genomic alterations such as copy-number variations (CNVs) are common in cancer. To investigate patterns of recurrent mutation inside the consensus TADs, we utilized CNV data from primary tumor samples of 32 different cancer types and 10,435 patients from the TCGA Data Portal (see Materials and Methods and Supplementary Table S3). For the purpose of reducing potential sources of bias, we excluded very small (< 1 kb) or large CNVs (> 10 Mb). Depending on the cancer type, the number of patient samples varied from 36 (CHOL, cholangiocarcinoma) to 1,097 (BRCA, breast invasive carcinoma). For comparability, we further restricted the analysis to the 25 cancer types for which CNV data for at least 100 patient samples were

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

available and randomly selected 100 patient samples for each cancer type. The median number of CNVs per patient in the resulting dataset ranged from 3 (THCA, thyroid carcinoma) to 280 (OV, ovarian serous cystadenocarcinoma, see Figure 2A).

Virtually all (99.6%) CNVs in the dataset (possibly partially) overlapped with a consensus TAD, to which for simplicity we will further refer as TADs. Moreover, due to their extent, on average, each CNV overlapped with four TADs, with 50% of the CNVs overlapping with at most three TADs (see Supplementary Figure S3). Conversely, almost all TADs (99.6%) overlapped with CNVs in the genome of at least one patient. Indeed, only five TADs were completely devoid from CNVs. Overall, on average, each TAD overlapped with CNVs in the genomes of 12% (298) of the patients, with 50% of the TADs overlapping with CNVs in the genomes of at most 12% (299) patients. Only 8% (115) of the TADs overlapped with CNVs in the genomes of 20% (500) or more patients and the largest number of patients associated with a TAD was 796 (32%). Importantly, the size of the TADs was only weakly correlated with the number of overlapping CNVs (Spearman's correlation coefficient = 0.32, see Supplementary Figure S4), suggesting that TADs overlapping with very few or very many CNVs might be involved in general cancer mechanisms.

To identify TADs with unusually large numbers of CNVs, we compared the number of patients with CNVs overlapping with each TAD with the expectation based on its size (see Materials and Methods). Independently of the cancer type, we found that 79 (6%) TADs were significantly enriched, with fold-differences of up to 3; in addition, the aforementioned five TADs overlapping with no CNVs were also significantly depleted, with their expectation ranging from 230 to 295 patients (see Materials and Methods and Figure 2B). Compared to other TADs, significantly enriched TADs were mainly associated with type I interferon response and natural killer cell activation (see Supplementary Table S4), which are known to be crucial for efficient tumor immune surveillance (e.g., (39)); depleted TADs were linked to transcriptional regulation. In addition, TADs enriched for CNV tended to be located towards the ends of the chromosomes, reflecting the genomic distribution of CNVs (40). As expected, different cancer types showed distinct patterns of enrichment, with a median of 64 TADs being significantly enriched for CNVs. In total, 487 (33%) TADs were significantly enriched for CNVs in the genome of the patients of at least one cancer type, including the 79 aforementioned TADs. We observed only one significantly depleted TAD, only for skin cutaneous melanoma (SKTM). Most (303) of the significantly enriched TADs were enriched for CNVs in the genome of the patients of two or more patient types (see Figure 2C). Moreover, although no TAD was significantly enriched for CNVs in the genome of the patients of all cancer types, thirteen TADs were for 15 or more cancer types, and might be considered the basis of a pan-cancer mutational signature. Intriguingly, genes with common alterations across different cancer types ("pan-cancer genes", see Materials and Methods), indicating that other mechanisms are likely to be involved.

TADs enriched for CNVs are valuable prognostic biomarkers in cancer

In order to assess the value of TADs enriched for CNVs in cancer prognosis, we trained LASSO Cox regression models for each of the 19 out of 25 cancer types for which at least 100 patients were available and for which at least 10% of the patients had a lethal outcome (see Materials and Methods). The models were based on age, sex and the presence/absence of CNVs in the TADs enriched for CNV in each particular patient cohort. The models for glioma tumors (GBM and LGG), clear cell and papillary renal carcinomas

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(KIRC and KIRP), gynecologic carcinomas (OV, BRCA and UCEC), rectum adenocarcinoma (READ) and sarcoma (SARC) displayed median c-indices between 0.55 and 0.8, and were therefore considered reliable. In order to establish a reference for evaluating the predictive power of our TAD-based models, we compared them to analogous nineteen gene-based models, trained and tested in the exact same manner (see Materials and Methods). We found that only six of the gene-based models were considered reliable and that all the corresponding TAD-based models had been considered reliable as well (see Figure 3A). Moreover, the TAD-based models performed better than the gene-based models in four cancer types (BRCA, OV, SARC, UCEC; P-values < 0.05, Wilcoxon's rank-sum test), worse in three (KIRP, KIRC and GBM; P-values < 0.05, Wilcoxon's rank-sum test) and similarly in two (LGG and READ).

The prognostic features of the TAD-based models can be used to stratify patients into high- and low-risk groups (all P-values ≤ 0.0001, log-rank test, see Materials and Methods and Figures 3B-E and Supplementary Figure S5). In addition to increasing age, which was generally associated with lower survival (6/9 models), we obtained a total of 35 prognostic TADs (see Materials and Methods and Figure 3F). Only three of these TADs were shared by at least two different cancer types (chr9:21240000-24400999 by LGG, GBM and KIRC; chr20:57399000-58159999 by BRCA and OV; and chr12:55679000-57720999 by GBM and KIRP), suggesting that the identity of the TADs harboring structural variants is mainly cancer-specific.

Interestingly, most (19, 54%) prognostic TADs did not comprise any pan-cancer genes (see Materials and Methods and Figure 3F). Consequently, in most cases, the presence of pan-cancer genes does not explain the predictive power of the TADs. A remarkable case was SARC, because only the TAD-based models exhibited a reliable performance and because none of the five TADs that were prognostic for SARC comprised any pan-cancer genes (see Figures 4A and B and Supplementary Figure S6). Furthermore, among the 16 TADs that comprised at least one pan-cancer gene (four in KIRC, three in LGG, four in GBM, one in UCEC, three in KIRP, one in BRCA, two in OV and one in READ, with one of these prognostic TADs being shared by GBM, LGG and KIRC and another one by KIRP and GBM, see Figure 3F), the prognostic effect of the TAD was not necessarily directed related to the pan-cancer gene(s). In fact, seven of these TADs did not comprise any prognostic pan-cancer genes, and among the remaining nine, only two TADs showed lower prediction power than the prognostic pan-cancer genes that they comprised (chr7:54760000-58079999, a prognostic TAD in LGG, comprising the prognostic pan-cancer gene EGFR; and chr12:55679000-57720999, a prognostic TAD in GBM, comprising the prognostic pan-cancer gene DDIT3; see Materials and Methods and Figures 4C-D and Supplementary Figure S7). In other words, in most cases, the survival of the patients with CNVs affecting the pan-cancer gene(s) in a TAD did not differ to that of the patients with CNVs in the TAD that did not affect the pan-cancer gene (see Materials and Methods) and the presence of a CNV in the pan-cancer gene was as informative as the presence of a CNV within the rest of the TAD. This has enormous implications for patient stratification for prognosis. For example, the TAD chr9:21240000-24400999, which is prognostic for LGG, comprises the prognostic pan-cancer gene *CDKN2A*. LGG patients with CNVs in this TAD (n=96) exhibited lower survival than patients with no CNVs in this TAD (n=414; P-value < 0.0001, log-rank test), and at least part of the predictive power of the TADs was not associated with the gene: namely, the survival of the patients with CNVs affecting *CDKN2A* (n=87) did not differ from that of the patients with CNVs in the TAD that did not affect *CDKN2A* (n=9; log-rank test, see

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures 4E-F). This shows that the mutations involving the sequence of *CDKN2A* are as relevant to LGG prognosis as those that do not, and that the latter should not be disregarded.

These results show that a large proportion of prognostic TADs exhibit predictive power independent of pan-cancer genes and illustrate the potential of TAD-based models to complement traditional gene-based models.

34% of prognostic TADs tend to undergo large structural changes in cancer

For the purpose of gaining further biological insight into the prognostic features of the TAD-based models, we constructed a consensus TAD map representing the most prevalent features of the three-dimensional organization of the human cancer genome (see Materials and Methods). This cancer TAD map comprised 1,467 (consensus) TADs. When comparing the TAD map constructed for the "normal", healthy human genome to the cancer TAD map, we found that 44% (643, covering 37% of the genome) of the TADs in the former had a highly similar counterpart in the latter (i.e., with a reciprocal overlap ≥ 95%, see Materials and Methods) and can be considered "constitutive", 34% (505, covering 29% of the genome) showed marked differences (i.e., they overlapped with less than 70% of any cancer TAD and/or less than 70% of the normal TAD overlapped with any cancer TAD, see Materials and Methods and Supplementary Figure S8) and are "perturbed", and 22% (319) displayed intermediate similarities and are, thus, ambiguous with regards to their conservation between the "normal" and cancer states (See Supplementary Table S5). These results are in agreement with the expected high degree of conservation of the TADs across tissues/cell lines, but also bring to light differences that may be associated with altered regulatory interactions in cancer.

In general, constitutive and perturbed TADs did not display any differential enrichment for CNVs. Nevertheless, the distribution of CNVs along the TADs exhibited a clear trend, with CNVs being enriched at the center of the TADs as compared to their boundaries and immediately adjacent TBRs (see Figure 5A and Materials and Methods). These findings support an association between TAD boundaries and CNVs. We hypothesize that CNVs disrupt the TAD boundaries and that because TAD disruption can rewire entire gene expression programs, it is very likely deleterious. This would explain the pressure to preserve TAD boundaries. Moreover, we observed that two subsets of perturbed TADs outstandingly deviated from the rest: a set of 120 TADs that are split into two or more TADs in the cancer genome and a set of 111 TADs that are fused with one or more other TADs in the cancer genome. While the former showed a stronger enrichment towards the center, the latter featured generally high enrichment, independently of the relative position in the TAD (see Figure 5A and Materials and Methods). And indeed, under our assumption, TADs that are split in the cancer genome would be expected to be especially enriched for CNVs towards the body of the TAD, away from their boundaries, while TADs that are fused in the cancer genome would show enrichment for CNVs at the boundaries. Thus, whereas the CNVs would induce the formation of new boundaries in the former, they would lead to the elimination of existing boundaries in the latter.

While 43% of the 35 prognostic TADs were constitutive, 34% were perturbed, indicating that at least one third of the predictive power of the TADs could be directly associated with changes in the three-dimensional organization of the genome. The ratio of constitutive to perturbed prognostic TADs was similar for different cancer types. For instance, out of the eight TADs that are prognostic for LGG – the best

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

performing models –, three were constitutive and four perturbed. Also, out of the five TADs that are prognostic for SARC – for which only the TAD-based models performed reliably –, three were constitutive and two perturbed. Many of the perturbed prognostic TADs appear to have been transformed through multiple splitting and fusion events in the cancer genome and were associated with local changes in the number of patients with CNVs. Thus, chr9:21240000−24400999 and chr9:24431000−24564999, two of the perturbed prognostic TADs for LGG, showed a relatively high number of CNVs, with a peak at the location of the pan-cancer gene *CDKN2A* (see Figure 5B); in the region comprising the perturbed prognostic TAD for LGG chr1:8000000-10440999, the number of CNVs decreased towards the 3' end (see Figure 5C); and for the remaining perturbed prognostic TAD for LGG, chr3:196040000-198159999, we observed a general increase in the number of CNVs (see Figure 5D). Similarly, while the perturbed prognostic TAD for SARC chr17:70680000-73360999 was split into two smaller TADs in the cancer genome (see Figure 5E), the remaining perturbed prognostic TAD for SARC, chr12:125039000-128113999, was fused with a neighboring TAD (see Figure 5F); both TADs were linked to a local increase in the number of CNVs. Remarkably, chr12:125039000-128113999 only comprises two protein-coding genes: *TMEM132B* and *AACS,* which had not been reported to be associated with SARC. Its adjacent TAD chr12:123760000-124919999 – with which it is fused in the cancer genome – comprises, among other genes, the pan-cancer gene *NCOR2*. SARC patients with CNVs in *NCOR2* (n=84) exhibited significantly lower survival than patients without CNVs in *NCOR2* (n=174, P-value < 0.01, log-rank test); they also featured higher *NCOR2* expression levels (median FPKM values 6,122 vs. 5,035, P-value < 0.01, Wilcoxon rank-sum test, see method). Interestingly, the survival of the patients with CNVs in *NCOR2* did not differ from that of the patients with no CNVs in *NCOR2* but with CNVs in chr12:125039000-128113999 (n=30), nor did they show any differences in their *NCOR2* expression levels. In this context, it is reasonable to hypothesize that the fusion of chr12:125039000-128113999 and chr12:123760000-124919999 leads to enhancers in chr12:125039000-128113999 being hijacked by *NCOR2*. Aberrant expression of *NCOR2* has been associated with several cancers (41); in particular, the complex formed by NCOR1 and 2 in concert with HDAC3 epigenetically suppresses myogenic differentiation in Embryonal rhabdomyosarcoma (ERMS), which is required for tumor growth (42).

In summary, with the exception of *CDKN2A*, which is a clear outlier in the dataset under consideration, the changes in the local number of CNVs did not appear to be connected to the pan-cancer genes located in these regions (*CAMTA1* and *MTOR*, *MUC4* and *TFRC*, and *NCOR2*, see Figures 5C, D and F, respectively). Our findings exemplify how comparing between the three-dimensional structure of the healthy and cancer human genomes can provide insights into the mechanisms by which non-coding variants lead to gene expression dysregulation in cancer.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

TADs are known to be generally well conserved across cell types and species (2,6-8). To obtain a consensus TAD map representing the most prominent three-dimensional features of the human genome, we integrated the information from different TAD maps generated using Hi-C on 24 normal tissues. These tissues represent a wide coverage of the human body and were selected based on availability of Hi-C data and corresponding TAD maps. Importantly, the Hi-C data for this collection of samples were generated by a relatively small number of laboratories, using similar protocols, and processed with the exact same computational pipeline (2). Indeed, the Hi-C protocol is complex and variations at the numerous steps can affect the resulting data (43,44). Pairwise comparisons of the TAD maps from the 24 normal tissues indicated only minor batch effects associated with the laboratory that generated the data. Integrating further TAD maps would imply being able to distinguish batch effects from biological variability, which is not possible with the current understanding of the factors that contribute to the success of a Hi-C experiment. Moreover, the definition of our consensus TAD maps depends on several parameters. Specifically, we used a sliding window approach to decide whether each 40 kb-long genomic window is part of a consensus TAD or not based on the number and type of tissues in which it was found to be part of one of the original TADs. This is controlled by two parameters: the conservation score and the boundary score. We empirically chose to set both parameters to 0.5; effectively, this means that, on average, the genomic window of interest is part of a TAD in 50% of the TAD maps. Due to the high similarity observed between the TAD maps of different tissues, our consensus TADs are not expected to change dramatically upon moderately increasing this threshold. Finally, the approach takes into account that closely related tissues – as assessed from their gene expression profiles – are expected to show similar TAD maps, accounting for possible biases in the collection of samples. It will be interesting to revisit these analyses as broader Hi-C datasets become available.

Many different types and locations of CNVs have been linked to cancer. To look for associations between the presence/absence of CNVs in specific TADs and cancer prognosis, we utilized CNV data from TCGA. TCGA is a rich resource comprising different genomic and transcriptomic data types for hundreds of patients from 32 different cancer types. There are certainly other similar resources, such as ICGC (45), which we could have employed in the analysis. Although the integration of such datasets would have resulted in larger cohorts and higher statistical power, their integration is not trivial and would have increased the number of unwanted artifacts. Using the TCGA dataset, we identified a group of 79 TADs that were susceptible to undergo copy number changes compared to random genomic sequences of the same size, and found that the specific TADs varied with the cancer type. In addition, using LASSO Cox regression modeling we showed that in 47% (9/19) of the assessed cancer types, the TADs enriched for CNVs were informative for patient prognosis. Furthermore, the TAD-based models performed similarly to models trained on 717 pan-cancer genes. The TAD-based models for sarcoma (SARC), which achieved a median c-index of 0.58, is an notable case. SARC patients usually harbor a relatively large number of variants in non-coding regions and no very good gene markers are known (46-48). Indeed, our gene-based models were unable to reliably predict survival (median c-index = 0.52). In contrast, using our TAD-based models identified five prognostic TADs, none of which comprised any pan-cancer genes. This observation illustrates the value of the TADs as prognostic biomarkers and their potential to improve upon gene-based models.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Remarkably, our TAD-based models provide insights into the mechanisms by which non-coding variants may contribute to cancer progression. In total, we identified 35 significantly prognostic TADs for nine different cancer types. Although functional assays are warranted to validate our findings, a relatively large fraction of these TADs appear to be associated with cancer-related changes in the three-dimensional organization of the genome. Indeed, among the 35 prognostic TADs, 34% were perturbed and, hence, had undergone changes in the majority of the cancer types examined. In addition, some of the constitutive TADs may be actually perturbed in a cancer type-specific manner. It is worth noticing that the distinction between constitutive and perturbed TADs depended on arbitrary thresholds on the overlaps between the normal and cancer TAD maps. The chosen thresholds were relatively strict; in particular, the threshold set for the constitutive TADs (95% reciprocal overlap) was chosen to reflect the expected high similarity, but to also allow for smaller rearrangements or variations in the protocols used for generating the underlying Hi-C data; and the perturbed TADs may display small structural alteration considering the high stability of TADs; analogously, the threshold set for perturbed TADs (at most 70% for one or both of the two overlaps considered, see Materials and Methods) was set to enable the identification of TADs with relatively large differences. Thus, the 22% of ambiguous TADs can be viewed as a safety buffer, since these represent TADs that cannot be unequivocally classified as either constitutive or perturbed. Changing the thresholds has a foreseeable effect on the classification; for example, while the number of constitutive TADs increases from 643 to 887 if we require a lower reciprocal overlap (80%), the number of perturbed TADs decreases from 505 to 274 if we enforce a lower maximum overlap (50%).

The comparison of both DNA-seq and Hi-C data for the same patient would have enabled a more direct examination of the effects of CNVs on the TADs. However, to date, TCGA does not contain Hi-C data. In fact, only a small number of patient Hi-C datasets have been generated. For instance, Kloetgen et al have produced and studied Hi-C data for six primary T-cell acute lymphoblastic leukemia (T-ALL) patients (49), and Díaz et al have done so for one large B-cell lymphoma patient (50). To conduct an independent validation of some of our results, we compared a list of 37 TADs that have been reported to undergo structural changes in the B-cells of lymphoma patients (50) to our normal consensus TADs; we observed that 78% (29/37) of these 37 TADs overlapped by at least 90% with our consensus TADs, but only 8% (3/37) overlapped with constitutive TADs, while the expectation based on their sizes is 27% (10/37). This observation supports our consensus TAD map and its utility to represent the normal, healthy three-dimensional structure of the genome.

Finally, we would like to stress that the performances of both the TAD- and gene-based models could be optimized, for example, by integrating other kinds of data – like RNA-seq – into the prediction or by opting for another feature section method. Our aim was simply to carry out a proof-of-principle study to demonstrate the value of TADs as prognostic markers for cancer. To our knowledge, this is the first time this has been investigated in a systematic manner. Our TAD-based models capture effects of CNVs in non-coding regions and are, thus, complementary to traditional gene-based models. In particular, our data hint towards a substantial fraction of prognostic features being linked to changes in the three-dimensional organization of the human genome. More generally, this study provides a framework for prioritizing non-coding variants for the development of personalized cancer therapies. The rapid increase in the number of

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

available patient Hi-C datasets promises to improve the efficacy of TAD-based models in cancer diagnosis and prognosis in the near future.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. Molecular cell **2016**;62:668-80
2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y*, et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **2012**;485:376-80
3. Nichols MH, Corces VG. A CTCF Code for 3D Genome Architecture. Cell **2015**;162:703-5
4. Acemel RD, Maeso I, Gomez-Skarmeta JL. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. Wiley Interdiscip Rev Dev Biol **2017**;6
5. Galupa R, Heard E. Topologically Associating Domains in Chromosome Architecture and Gene Regulatory Landscapes during Development, Disease, and Evolution. Cold Spring Harbor symposia on quantitative biology **2017**;82:267-78
6. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N*, et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature **2012**;485:381-5
7. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT*, et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **2014**;159:1665-80
8. Battulin N, Fishman VS, Mazur AM, Pomaznoy M, Khabarova AA, Afonnikov DA*, et al.* Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. Genome biology **2015**;16:77
9. Valton AL, Dekker J. TAD disruption as oncogenic driver. Current opinion in genetics & development **2016**;36:34-40
10. Kaiser VB, Semple CA. When TADs go bad: chromatin structure and nuclear organisation in human disease. F1000Research **2017**;6
11. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature reviews Genetics **2009**;10:551-64
12. Li BQ, You J, Huang T, Cai YD. Classification of non-small cell lung cancer based on copy number alterations. PLoS One **2014**;9:e88300
13. Zhang N, Wang M, Zhang P, Huang T. Classification of cancers based on copy number variation landscapes. Biochim Biophys Acta **2016**;1860:2750-5
14. Klopocki E, Mundlos S. Copy-number variations, noncoding sequences, and human phenotypes. Annual review of genomics and human genetics **2011**;12:53-72
15. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. Human molecular genetics **2016**;25:R157-r65
16. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E*, et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell **2015**;161:1012-25
17. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schopflin R*, et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature **2016**;538:265-9
18. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO*, et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature **2016**;529:110-4
19. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH*, et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science (New York, NY) **2016**;351:1454-8
20. Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM*, et al.* Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome research **2016**;26:719-31
21. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K*, et al.* The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics **2013**;45:1113-20
22. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A*, et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY) **2009**;326:289-93
23. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL*, et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell reports **2016**;17:2042-59
24. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY*, et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature **2015**;518:350-4
25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature **2012**;489:57-74
26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology **2014**;15:550
27. Stone EA, Sidow A. Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC bioinformatics **2007**;8:222

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends in genetics : TIG **2013**;29:569-74
29. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research **2014**;42:W187-91
30. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A*, et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British journal of cancer **2004**;91:355-8
31. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research **2009**;37:1-13
32. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols **2009**;4:44-57
33. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine **1997**;16:385-95
34. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw **2010**;33:1-22
35. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama **1982**;247:2543-6
36. Laimighofer M, Krumsiek J, Buettner F, Theis FJ. Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. Journal of computational biology : a journal of computational molecular cell biology **2016**;23:279-90
37. Therneau T. A Package for Survival Analysis in S. 2.38 2015.
38. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. PeerJ Computer Science **2016**:e67
39. Muller L, Aigner P, Stoiber D. Type I Interferons and Natural Killer Cell Regulation in Cancer. Frontiers in immunology **2017**;8:304
40. Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. PLoS Genet **2006**;2:e20
41. Wong MM, Guo C, Zhang J. Nuclear receptor corepressor complexes in cancer: mechanism, function and regulation. Am J Clin Exp Urol **2014**;2:169-87
42. Phelps MP, Bailey JN, Vleeshouwer-Neumann T, Chen EY. CRISPR screen identifies the NCOR/HDAC3 complex as a major suppressor of differentiation in rhabdomyosarcoma. Proceedings of the National Academy of Sciences of the United States of America **2016**;113:15090-5
43. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V*, et al.* Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nature communications **2018**;9:1028
44. Golloshi R, Sanders JT, McCord RP. Iteratively improving Hi-C experiments one step at a time. Methods (San Diego, Calif) **2018**;142:47-58
45. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J*, et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database : the journal of biological databases and curation **2011**;2011:bar026
46. D'Angelo E, Prat J. Uterine sarcomas: a review. Gynecologic oncology **2010**;116:131-9
47. Tsuyoshi H, Yoshida Y. Molecular biomarkers for uterine leiomyosarcoma and endometrial stromal sarcoma. Cancer science **2018**;109:1743-52
48. Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell **2017**;171:950-65.e28
49. Kloetgen A, Thandapani P, Ntziachristos P, Ghebrechristos Y, Nomikou S, Lazaris C*, et al.* Dynamic 3D chromosomal landscapes in acute leukemia. bioRxiv; 2019.
50. Diaz N, Kruse K, Erdmann T, Staiger AM, Ott G, Lenz G*, et al.* Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nature communications **2018**;9:4938

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures and Legends

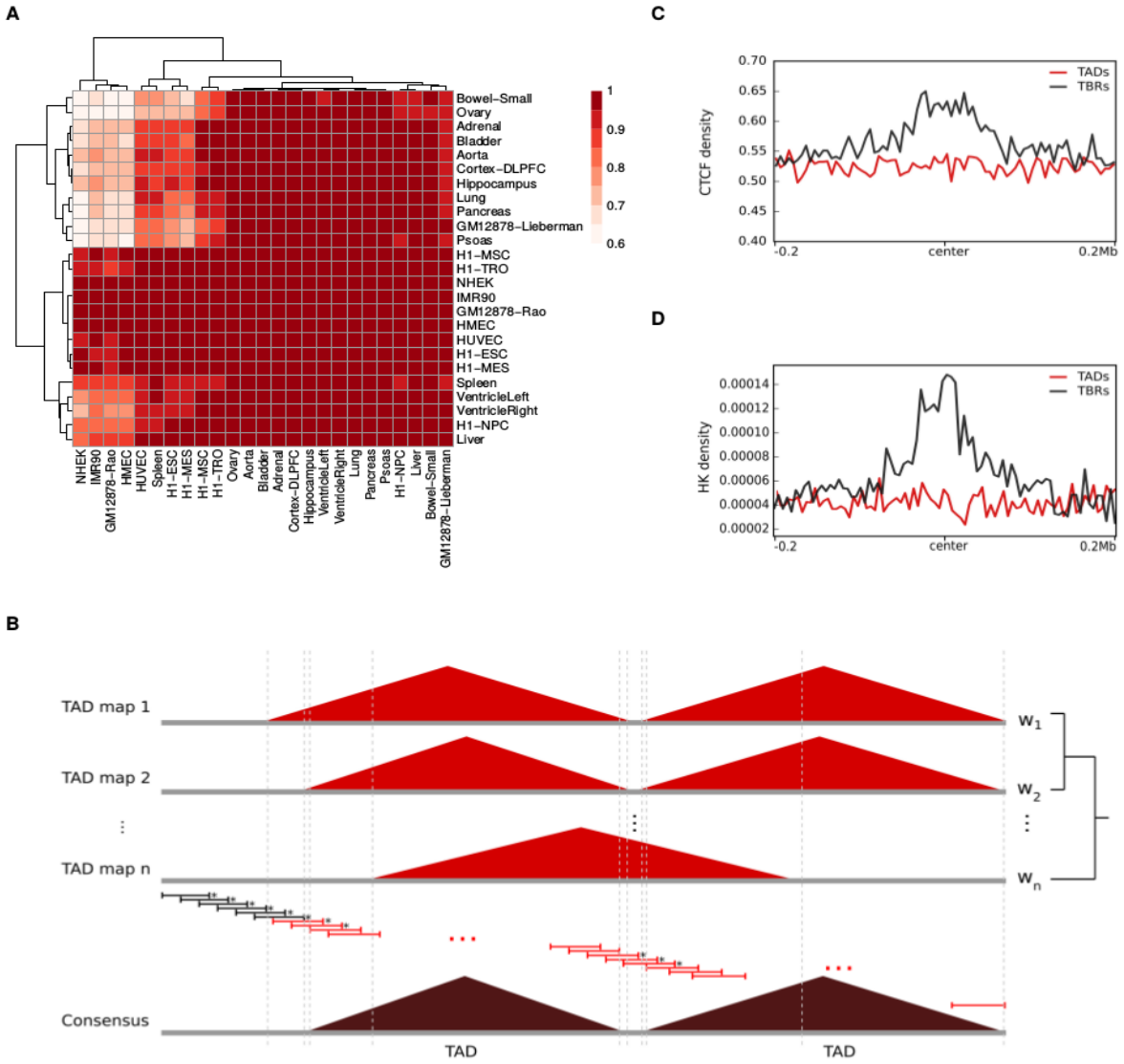

Figure 1. TAD maps across various human tissues can be combined into a consensus TAD map. A) TAD maps of different human tissues generally showed a high similarity. The heatmap summarizes the pairwise similarities of every pair of samples; specifically, we computed the maximum fraction of a TAD in the *j*th sample (columns) that overlapped with a TAD in the *i*th sample (rows). The color of the cells represents the median of the fractions for all TADs in the *j*th sample. Note that this is not symmetric. Rows and columns are clustered based on Euclidean distance using complete linkage. **B)** The consensus TADs were derived by computing TAD conservation and boundary scores in sliding windows across the genome. Adjacent windows were merged as long as they had a nucleotide-wise TAD conservation score ≥ 0.5 none of the windows had a TAD boundary score ≥ 0.5, and defined as consensus TADs if their size was ≥ 40 kb. **C and D)** TADs and TBRs showed distinct coverage of CTCF and housekeeping (HK) genes. Density of CTCF (C) and housekeeping genes (D) calculated for 5kb bp-long bins covering the center of the TADs (blue) /TBRs (green) ± 200 kb.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

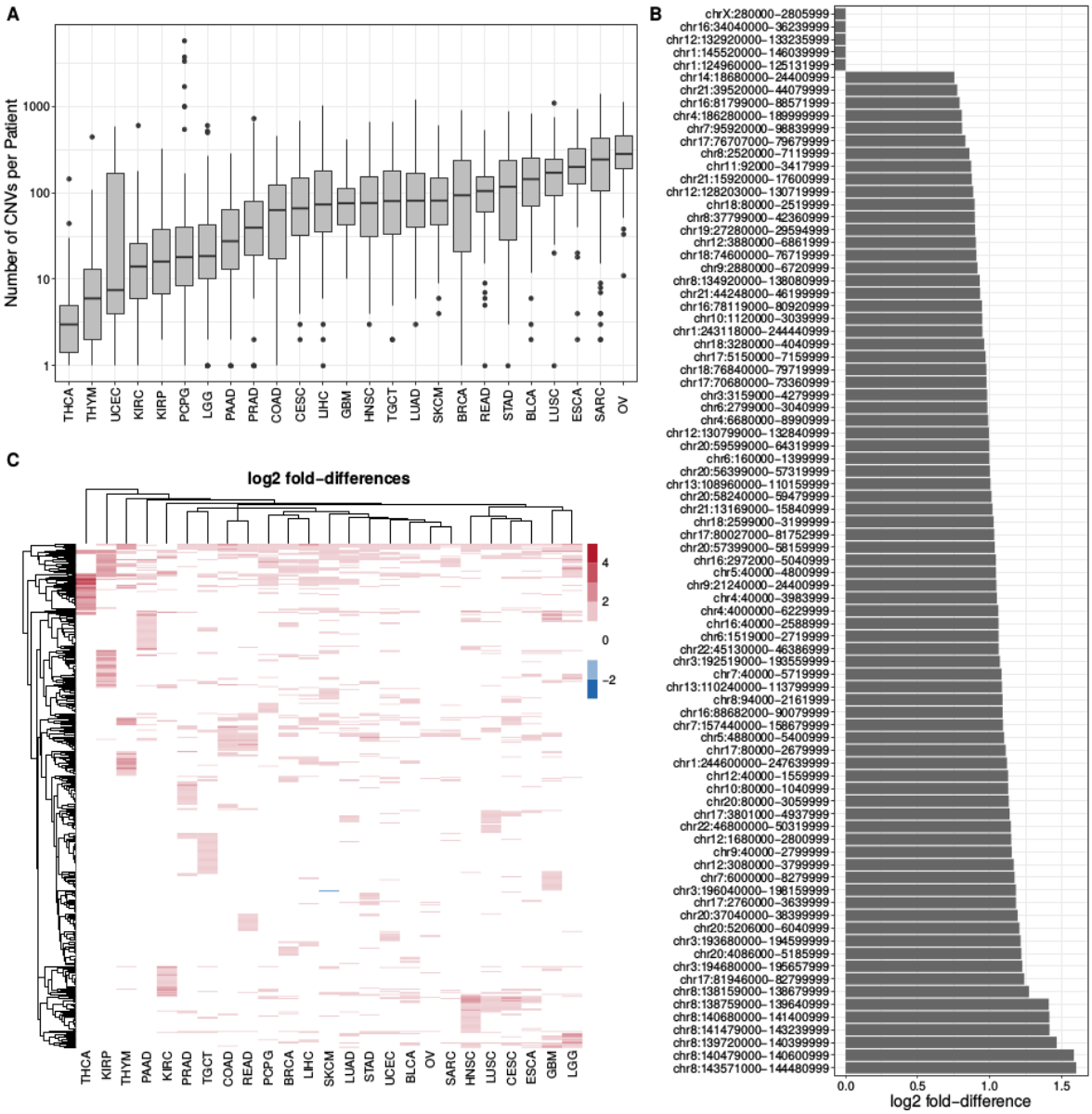

Figure 2. Some of the consensus TADs are enriched or depleted for cancer-related CNVs. A) The number of CNVs per patient differed among the 25 considered cancer types. The number of CNVs per patient (y-axis) is represented on a $\log_{10}$ scale. **B)** 79 TADs were enriched and 5 TADs were depleted for CNVs across all the patients of all 25 cancer types. The coordinates of the TAD are indicated on the y-axis; the bars show $\log_2$ fold-differences to the genomic expectation. **C)** In total, 487 TADs were significantly enriched or depleted for CNVs in patients of at least one of the 25 cancer types, including the 79 aforementioned TADs. The heatmap displays the $\log_2$ fold-differences to the genomic expectation for those TADs. Rows and columns are clustered based on Euclidean distance using complete linkage.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

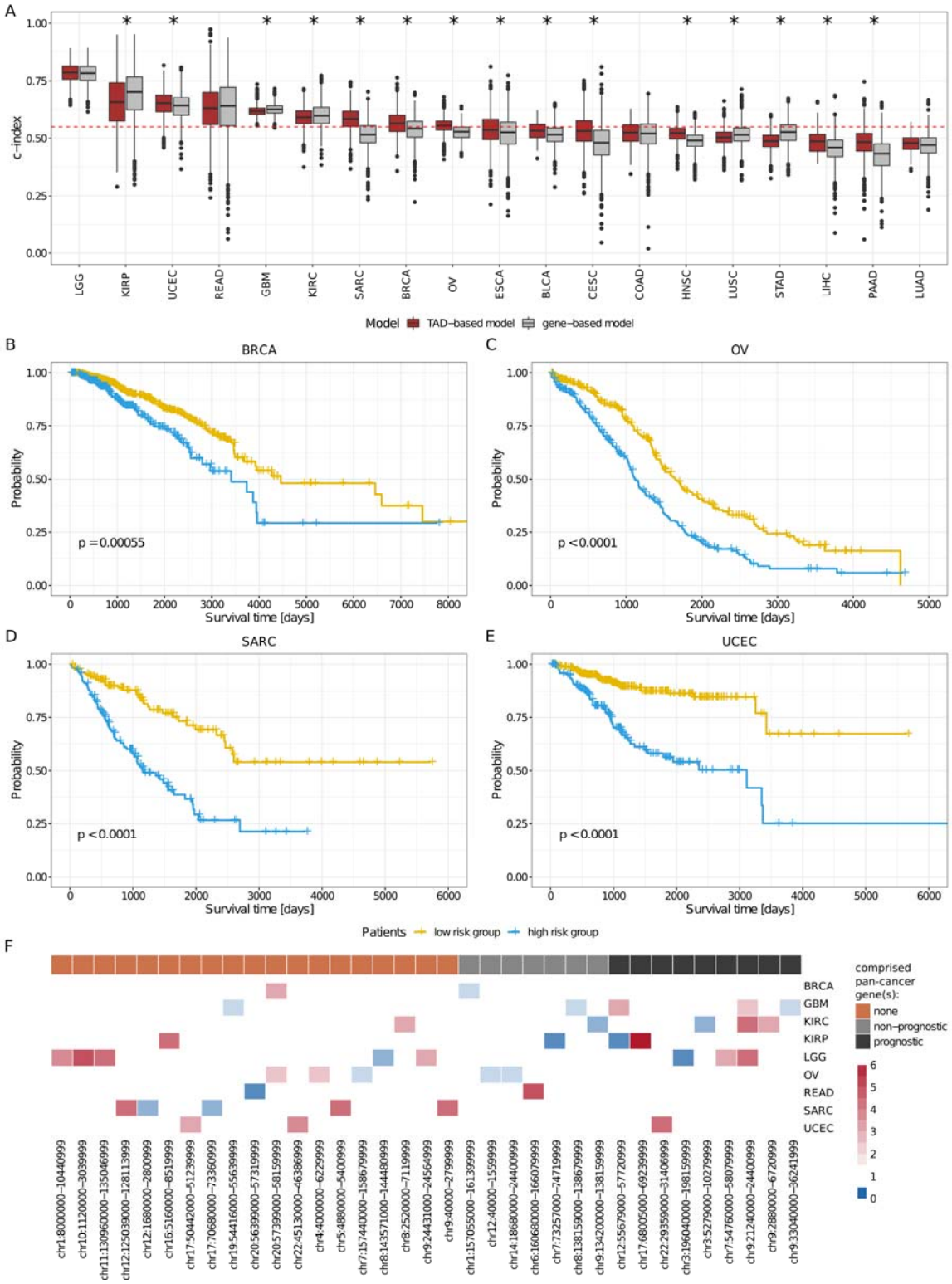

Figure 3. The presence/absence of CNVs in certain TADs is prognostic for overall survival in cancer patients. **A)** In nine out of 19 cancer types the TAD-based models worked and in four out of those nine they performed significantly better than the gene-based models, whereas there was no difference in two. The boxplots summarize the c-indices (CI) computed on 1,000 random test sets of patients, across 19 cancer

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

types, for the TAD- (red) and the gene-based (gray) models. The dotted red line indicates CI = 0.55. The asterisk indicates significant differences between the TAD- and gene-based models (Wilcoxon rank-sum test). A c-index = 1 indicates a perfect prediction, while a c-index = 0.5, a random prediction. **B-E)** Patients were separated into high- (blue) and a low-risk (yellow) groups according to the prognostic features of the final LASSO Cox regression model and subjected to Kaplan-Meier analysis. B) BRCA; C) OV; D) SRAC; E) UCEC. **F)** 35 TADs enriched for CNVs were associated with higher/lower overall survival and, thus, prognostic. The heatmap shows the hazard ratios (HZ) derived from the final LASSO Cox regression model for each prognostic TAD and cancer type; red indicates HZ > 1 (lower survival) and blue HZ < 1 (higher survival). Prognostic TADs were further categorized into three groups: i) TADs that did not comprise any pan-cancer genes (orange); ii) TADs that comprised only non-prognostic pan-cancer genes (gray); and iii) TADs that comprised prognostic pan-cancer genes (black).

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

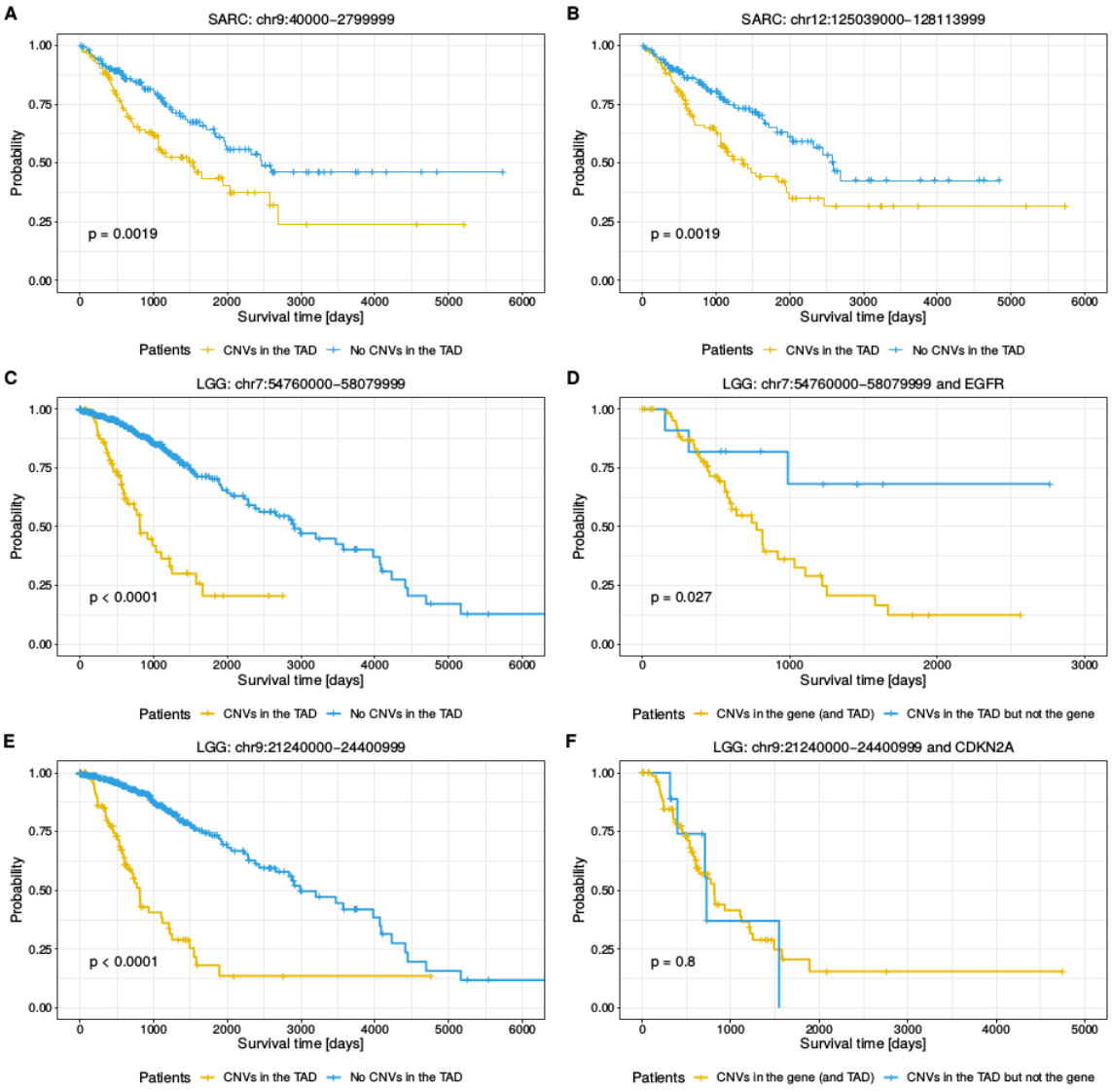

Figure 4. Kaplan-Meier survival analysis for SARC and LGG patients. Time is indicated in days. **A-B)** LASSO Cox regression models identified five prognostic TADs for SARC. Patients were separated into two groups, according to the presence (yellow) or absence (blue) of CNVs in two of these TADs: A) chr9:40000-2799999; and B) chr12:125039000-128113999. **C-D)** LGG patients were separated into two groups, according to the presence or absence of CNVs in the prognostic TAD chr7:54760000-58079999 and in the prognostic pan-cancer gene *EGFR*, comprised by this TAD. C) Patients with CNVs in chr7:54760000-58079999 (yellow) compared to those without CNVs in this TAD (blue). D) Patients with CNVs in *EGFR* (yellow) compared to those with CNVs in chr7:54760000-58079999 but not affecting *EGFR* (blue). Patients with CNVs in *EGFR* exhibited lower survival compared to those with CNVs in the TAD that did not affect the gene. **E-F)** LGG patients were separated into two groups, according to the presence or absence of CNVs in the prognostic TAD chr9:21240000-24400999 and in the prognostic pan-cancer gene *CDKN2A*, comprised by this TAD. E) Patients with CNVs in chr9:21240000-24400999 (yellow) compared to those without CNVs in this TAD (blue). F) Patients with CNVs in *CDKN2A* (yellow) compared to those with CNVs in chr9:21240000-24400999 in the TAD that did not affect *CDKN2A* (blue). The survival of patients with CNVs in *CDKN2A* did

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

not differ from that of patients with CNVs in chr9:21240000-24400999 that did not affect *CDKN2A*; and both groups exhibited lower survival than the patients without CNVs in this TAD. Hence, CNVs in the prognostic TAD, independently of whether they affect the gene sequence or not, were associated with lower survival of LGG patients.

bioRxiv preprint doi: https://doi.org/10.1101/777573; this version posted September 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

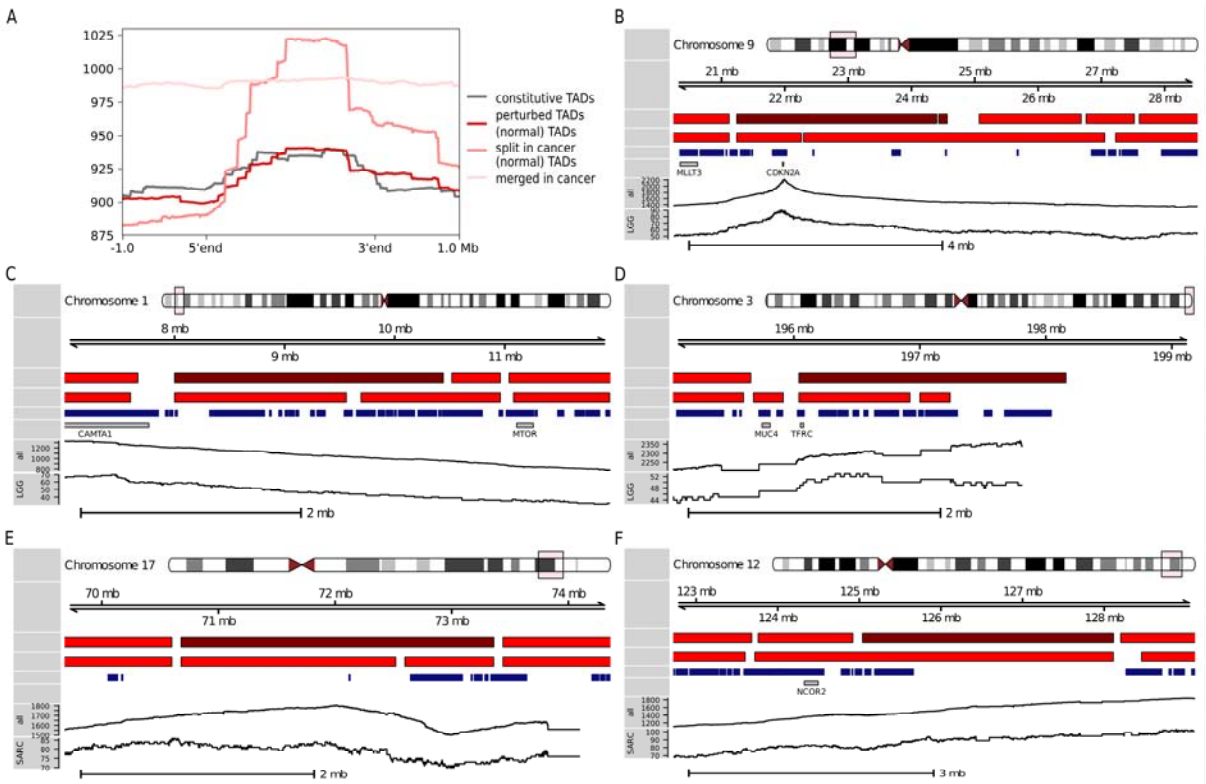

Figure 5. The presence of CNVs tends to be associated with TAD perturbations. A) Density of CNVs calculated for 10 kb-long bins along TADs and 1 Mb-long upstream and downstream genomic regions. TADs were scaled to 2 Mb. The 505 perturbed TADs (dark red) show no difference compared to the 643 constitutive TADs. However, the two subsets of perturbed TADs deviate from the general trend. **B-F)** Genomic context for 35 prognostic TADs. The first two tracks indicate the physical location on the chromosome and the coordinates of the displayed region. The third and fourth tracks represent the TADs in the normal and cancer genomes, respectively; the TAD in the genomic region under consideration are shown in red, with the TAD(s) of interest highlighted in dark red. The fifth and sixth tracks indicate protein-coding genes and pan-cancer genes, respectively. The sixth track visualizes the number of cancer patients with CNVs at each nucleotide of the genomic region under consideration; the last track restricts patients to those of the cancer type for which the TAD of interest was prognostic. **B)** Prognostic TADs for LGG chr9:21240000-24400999 and chr9:24431000-24564999 (displayed region: chr9:20240000-28519999); **C)** Prognostic TAD for LGG chr1:8000000-10440999 (displayed region: chr1:7000000-11959999); **D)** Prognostic TAD for LGG chr3:196040000-198159999 (displayed region: chr3:195040000-199159999). **E)** Prognostic TAD for SARC chr17:70680000-73360999 (displayed region: chr17:69680000-74360999) and **F)** Prognostic TAD for SARC chr12:125039000-128113999 (displayed region: chr12:122719000-129113999).