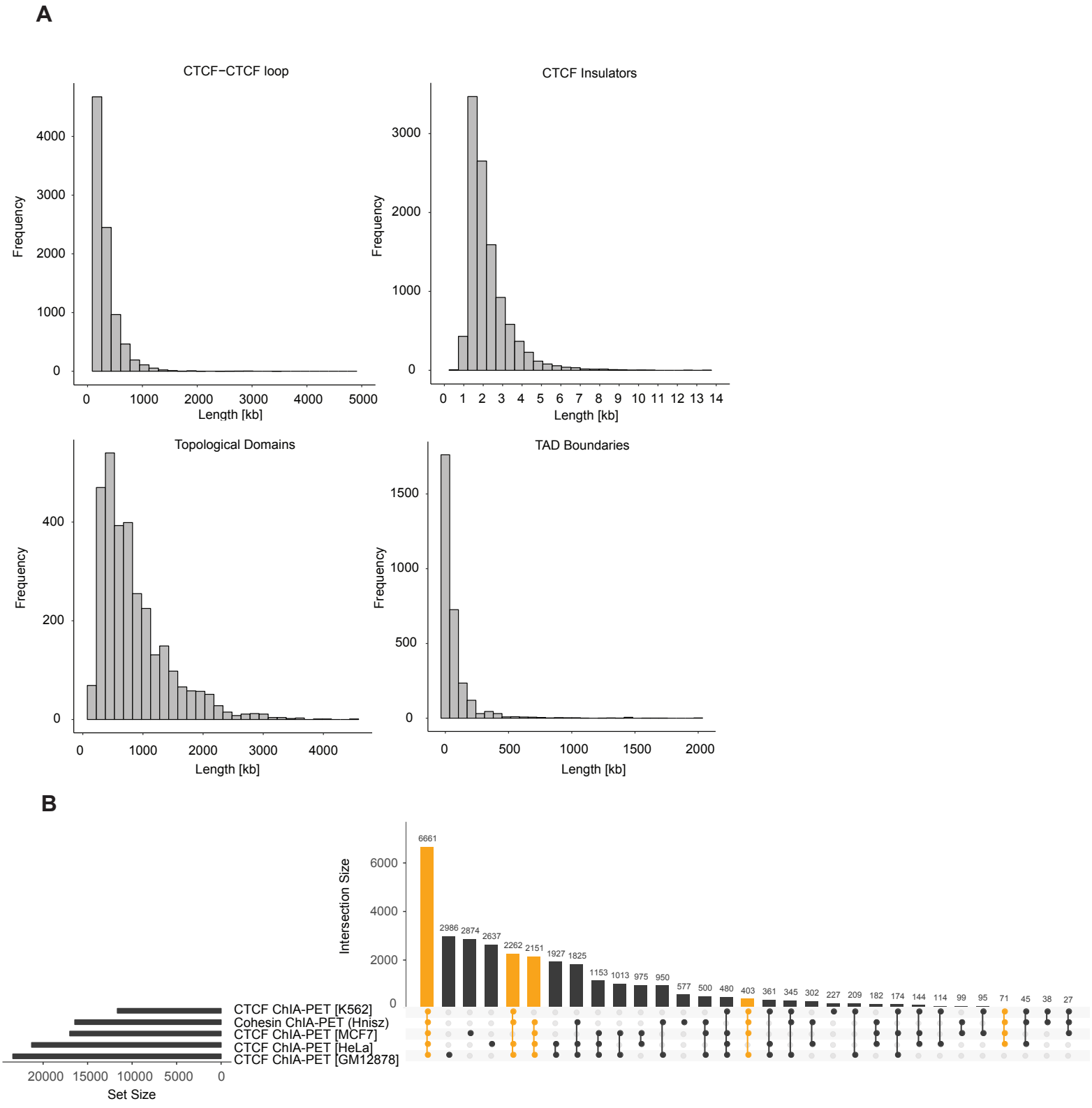
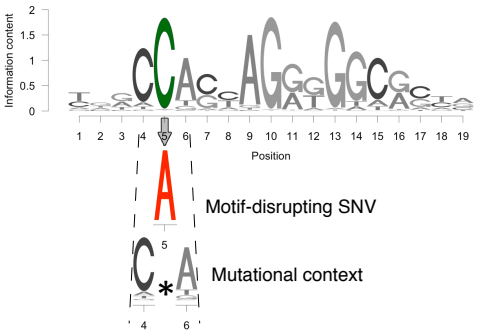


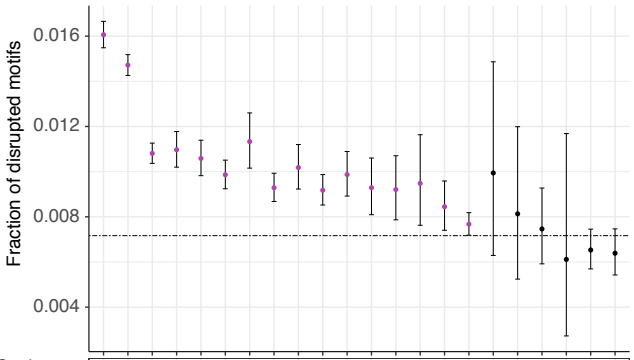
Figure S1. Details of CTCF/cohesin insulators, Related to Figure 1

(A) Length distributions shown for insulators, CTCF-CTCF loops, TADs and TAD boundaries. The median length of insulators is 2 kb. (B) Upset plot to represent the number of overlapping CTCF/cohesin insulators between CTCF ChIA-PET and cohesin ChIA-PET data. Yellow colors represent the conserved constitutive CTCF/cohesin insulators used in this study.

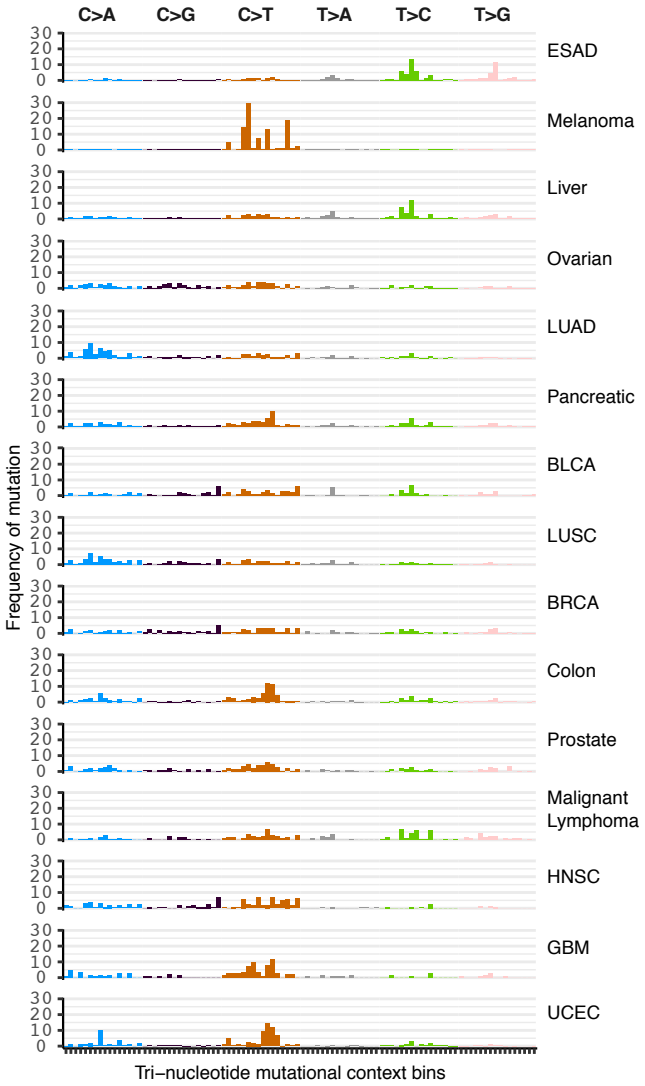
A Predicted CTCF motif-disruption due to SNVs



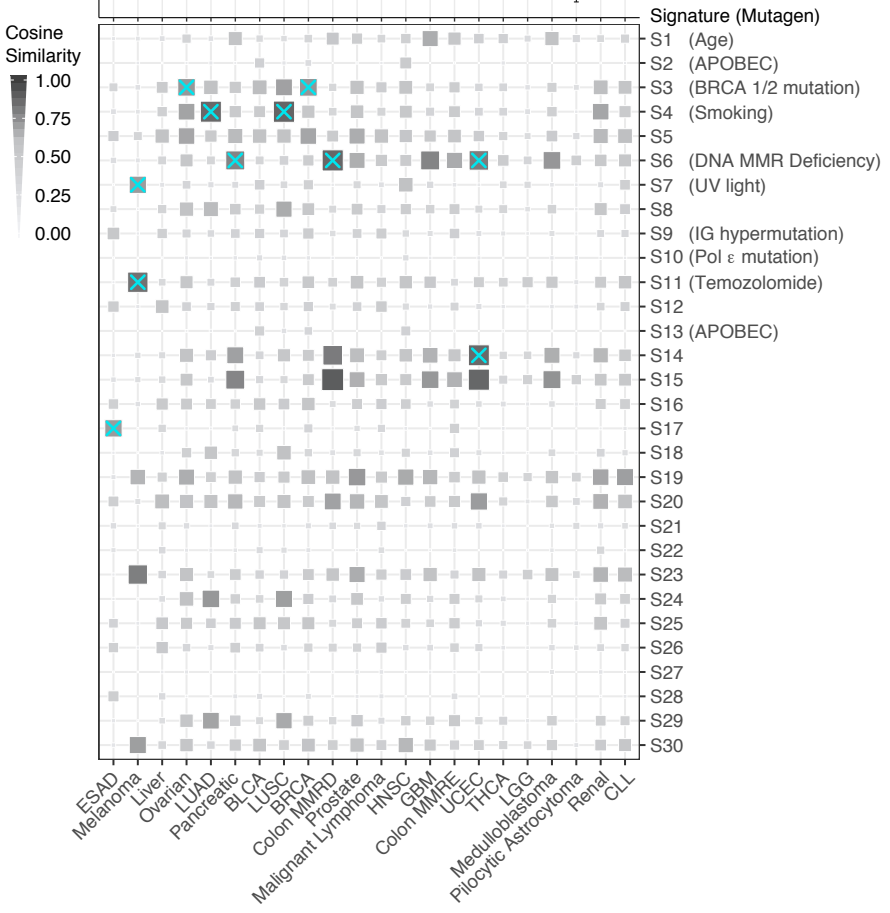
B Fraction of CTCF motif disrupted per cancer type



C Tri-nucleotide distribution of CTCF motif-disrupting mutations



D



Cosine similarity:
$\cos(\theta) = (A \cdot B)/(\|A\| \|B\|)$
A: Tri-nucleotide distribution of CTCF motif-disrupting mutation
B: Mutational signauture

× $\cos(\theta)$ >= 70% for known mutational signature in the corresponding cancer type

Figure S2. CTCF motif-disrupting mutations, Related to Figure 1

(A) Schematic representation of a SNV predicted to disrupt CTCF motifs and its tri-nucleotide context. (B) Fraction of disrupted CTCF motifs relative to total disrupted motifs for all transcription factors (TFs). Horizontal dashed line is the proportion of CTCF motifs among the total number of TF motifs within ENCODE ChIP-Seq or DHS peaks. Purple is for cancer types showing significant enrichment of disrupted CTCF motifs. Colon cancer samples are divided into mismatch-repair-deficient (MMRD) and mismatch-repair-efficient (MMRE). (C) Observed tri-nucleotide distribution of CTCF motif-disrupting mutations for each cancer type. (D) Matrix depicting cosine similarity between the CTCF motif-disrupting distributions per cancer type and the COSMIC mutational signatures. Bigger size and darker color of the grey square-points indicates higher similarity. Cosine similarity ≥ 70% with mutational signature reported for the corresponding cancer type is denoted by blue cross. Mutations at all CTCF binding sites were considered for this analysis. BLCA: Bladder urothelial carcinoma; GBM: Glioblastoma; LGG: Lower grade glioma; BRCA: Breast invasive carcinoma; CLL: Chronic lymphocytic leukemia; UCEC: Uterine corpus endometrial carcinoma; ESAD: Esophageal adenocarcinoma; HNSC: Head and Neck squamous cell carcinoma; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; THCA: Thyroid carcinoma.

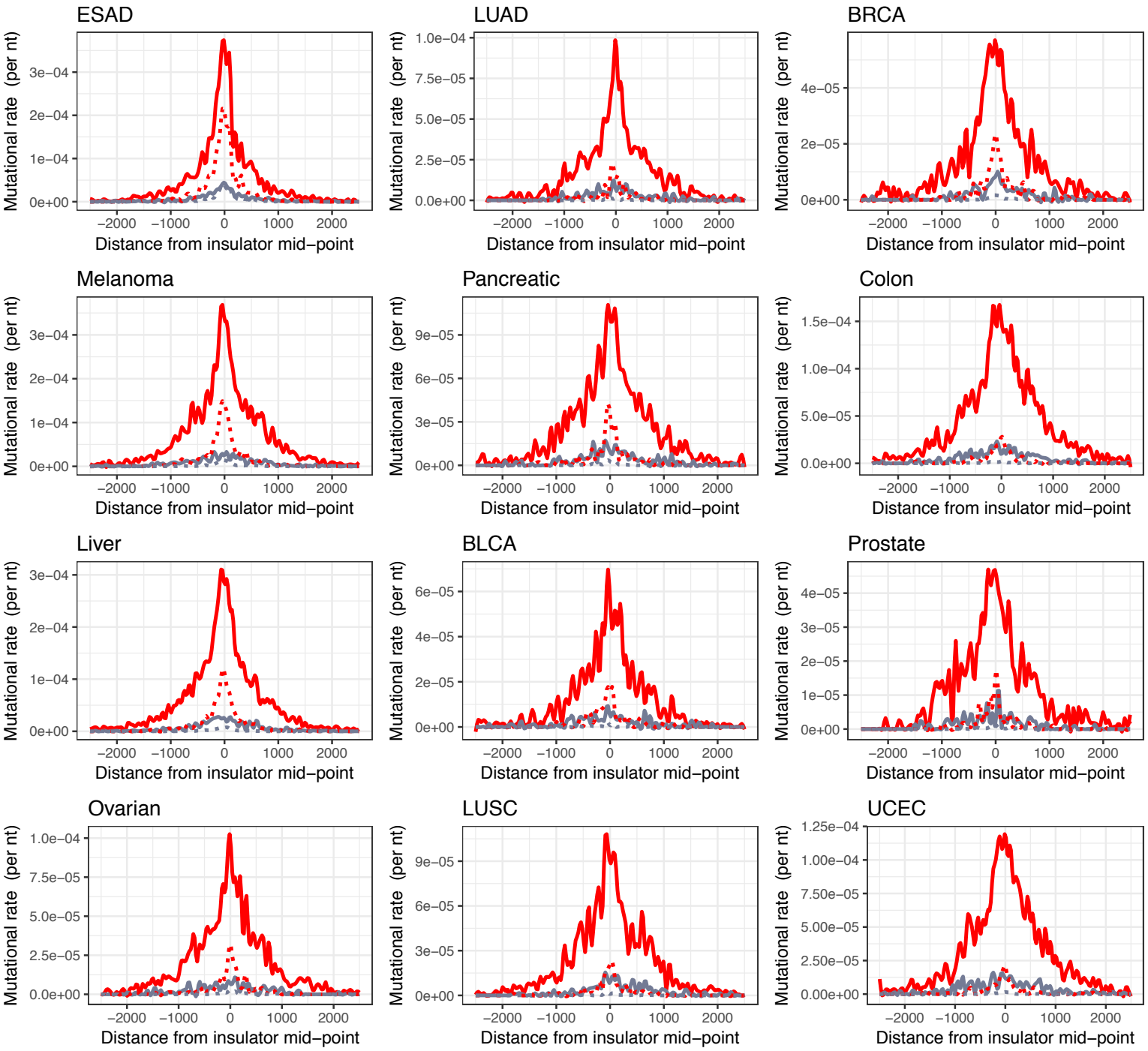


Figure S3. Aggregate mutation rate at CTCF motifs within insulators, Related to Figure 1
For each panel the solid red line represents the observed rate of mutations that disrupt CTCF motifs and dotted red line represents the expected rate. The observed rate is higher than expected for all cancer types. For comparison, the observed rate of mutations that are not predicted to disrupt CTCF motifs is shown by solid grey lines and the dotted grey corresponds to their expected rate. Only cancer types with more than 500 mutations predicted to cause CTCF motif-disruption within insulator regions are shown.

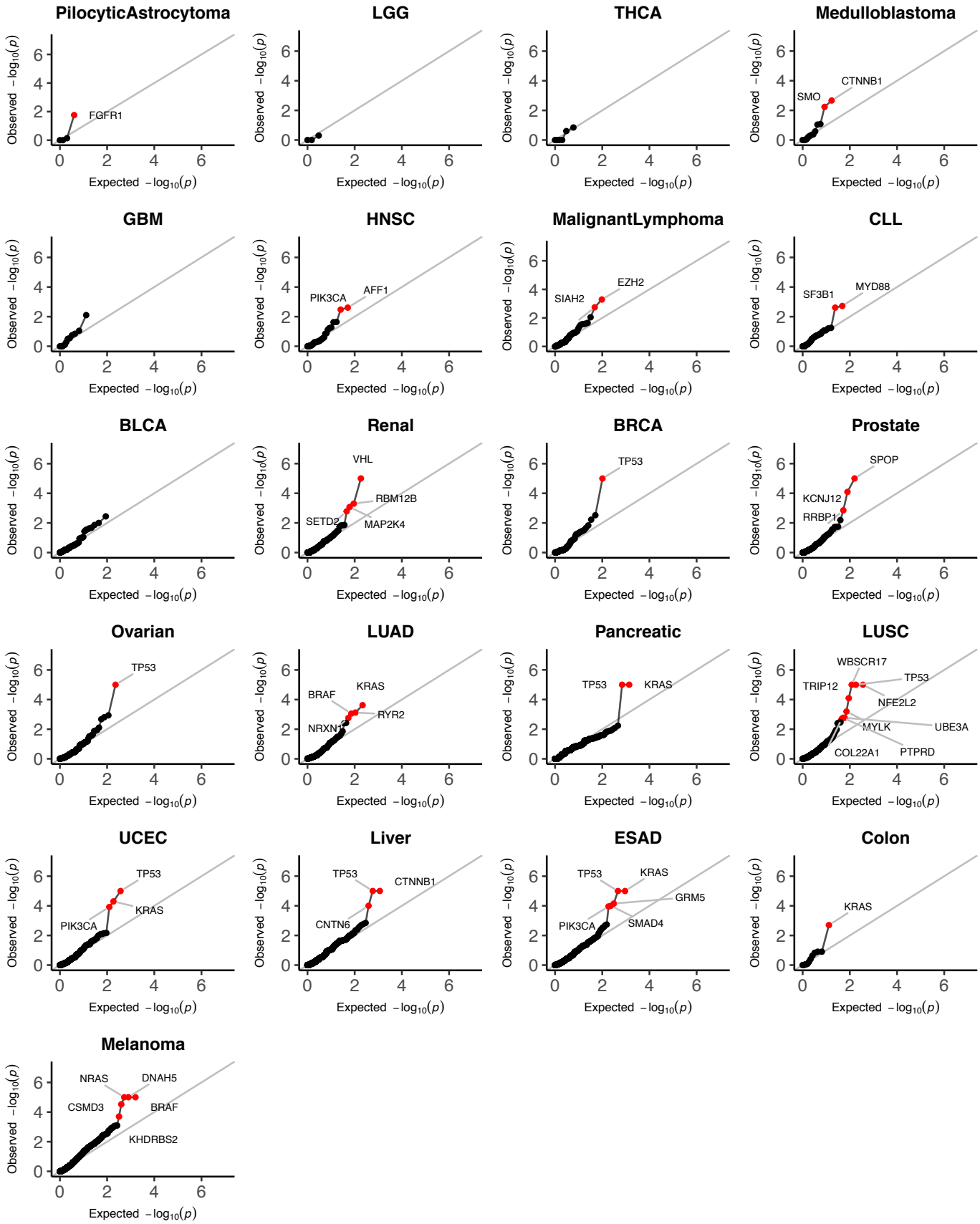


Figure S4. QQ plots of coding sequence (CDS) from CNCDriver, Related to Figure 1
QQ plots of CDS showing expected versus observed P values from CNCDriver. Candidates with Q value ≤ 0.1 are labeled and shown in red.

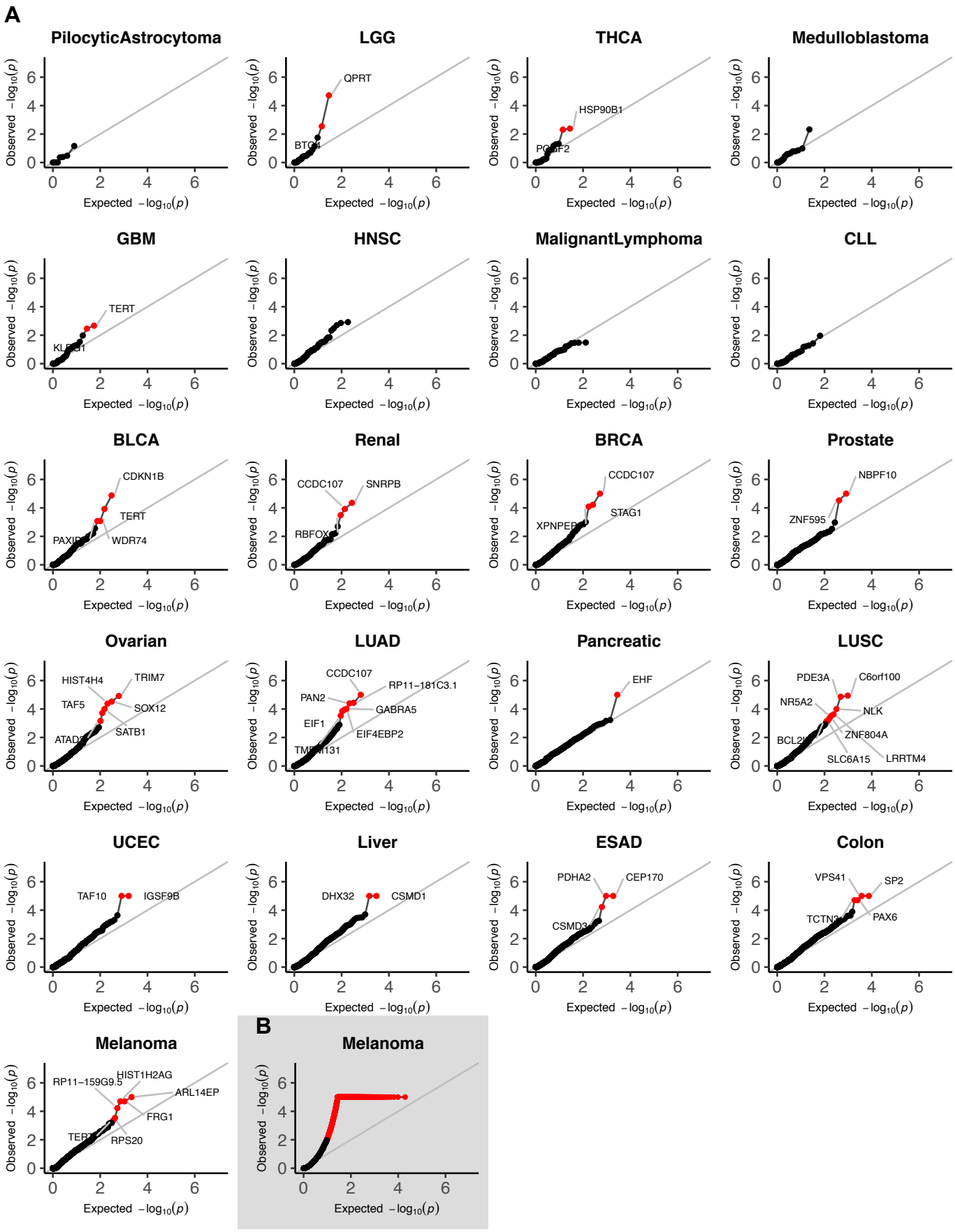

A
PilocyticAstrocytoma
LGG
QPRT
BTG1
THCA
HSP90B1
PCGF2
Medulloblastoma
GBM
TERT
KLF9
HNSC
MalignantLymphoma
CLL
BLCA
CDKN1B
TERT
PAXIP1
WDR74
Renal
CCDC107
SNRPB
RBFOX2
BRCA
CCDC107
STAG1
XPNPEP1
Prostate
NBPF10
ZNF595
Ovarian
HIST4H4
TRIM7
TAF5
SOX12
SATB1
ATAD2
LUAD
CCDC107
RP11−181C3.1
PAN2
GABRA5
EIF1
EIF4EBP2
TMEM131
Pancreatic
EHF
LUSC
PDE3A
C6orf100
NR5A2
NLK
ZNF804A
BCL2L1
SLC6A15
LRRTM4
UCEC
TAF10
IGSF9B
Liver
DHX32
CSMD1
ESAD
PDHA2
CEP170
CSMD3
Colon
VPS41
SP2
TCTN3
PAX6
Melanoma
RP11−159G9.5
HIST1H2AG
ARL14EP
FRG1
TERT
RPS20
B
Melanoma
Observed −log10(p)
Expected −log10(p)

Figure S5. QQ plots of promoters, Related to Figure 1

(A) CNCDriver results: QQ plots of promoters showing expected versus observed P values from CNCDriver. Candidates with Q value ≤ 0.1 are labeled and shown in red. In melanoma, KDM5A, RPL13A, GDAP1, RPL34 were further filtered due to extended TTCCG mutational signature (Fredriksson et al., 2017). (B) OncoDriveFML result: QQ plot of OncodriveFML promoter result in melanoma. Candidates with Q value ≤ 0.1 are shown in red. OncodriveFML predicted 2,057 significant candidates with Q value ≤ 0.1. Lambda value is 2.771 vs. the value of 1.313 for melanoma QQ plot from CNCDriver.

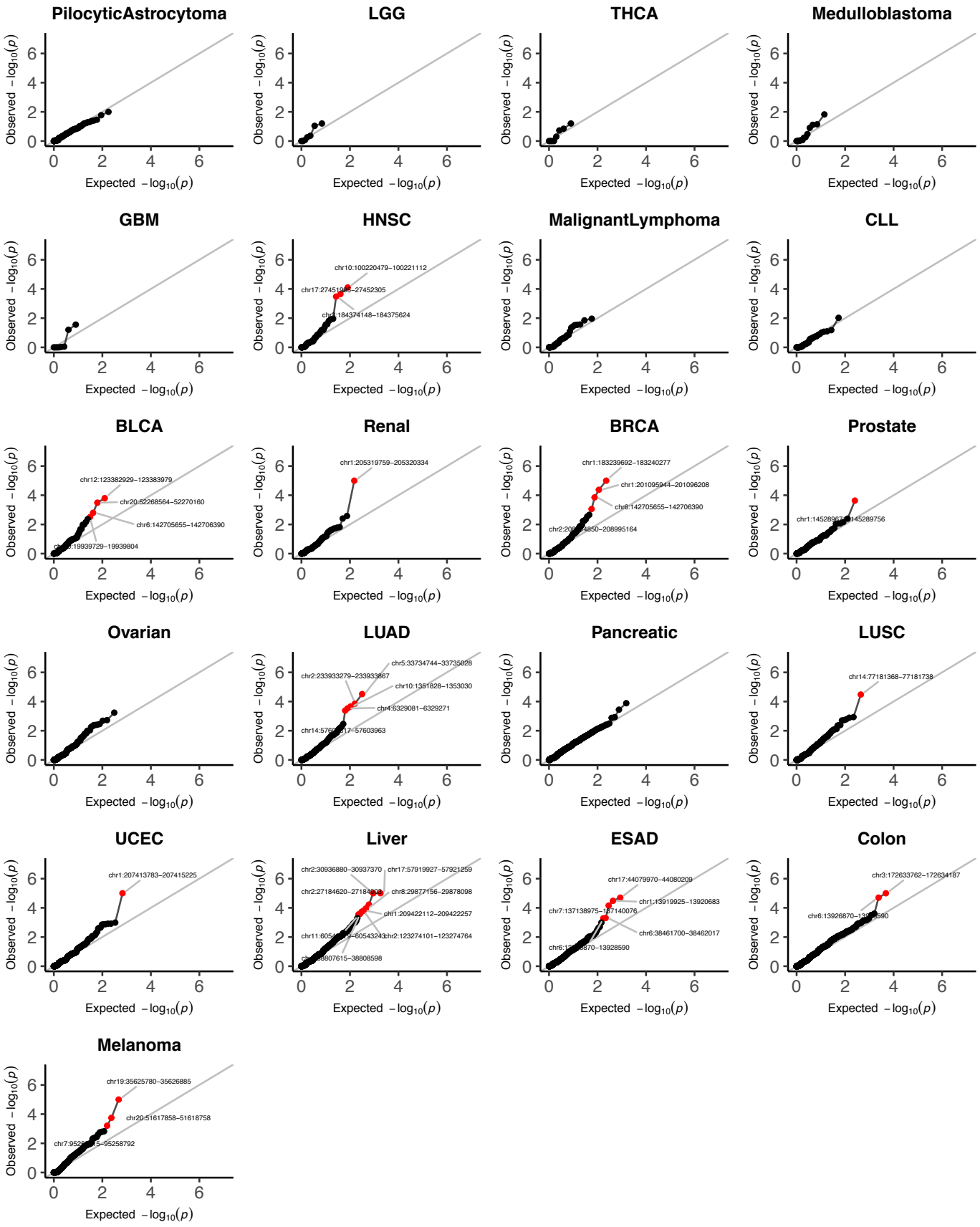


Figure S6. QQ plots of enhancers from CNCDriver, Related to Figure 1

QQ plots of enhancers showing expected versus observed P values from CNCDriver. Candidates with Q value ≤ 0.1 are labeled and shown in red.

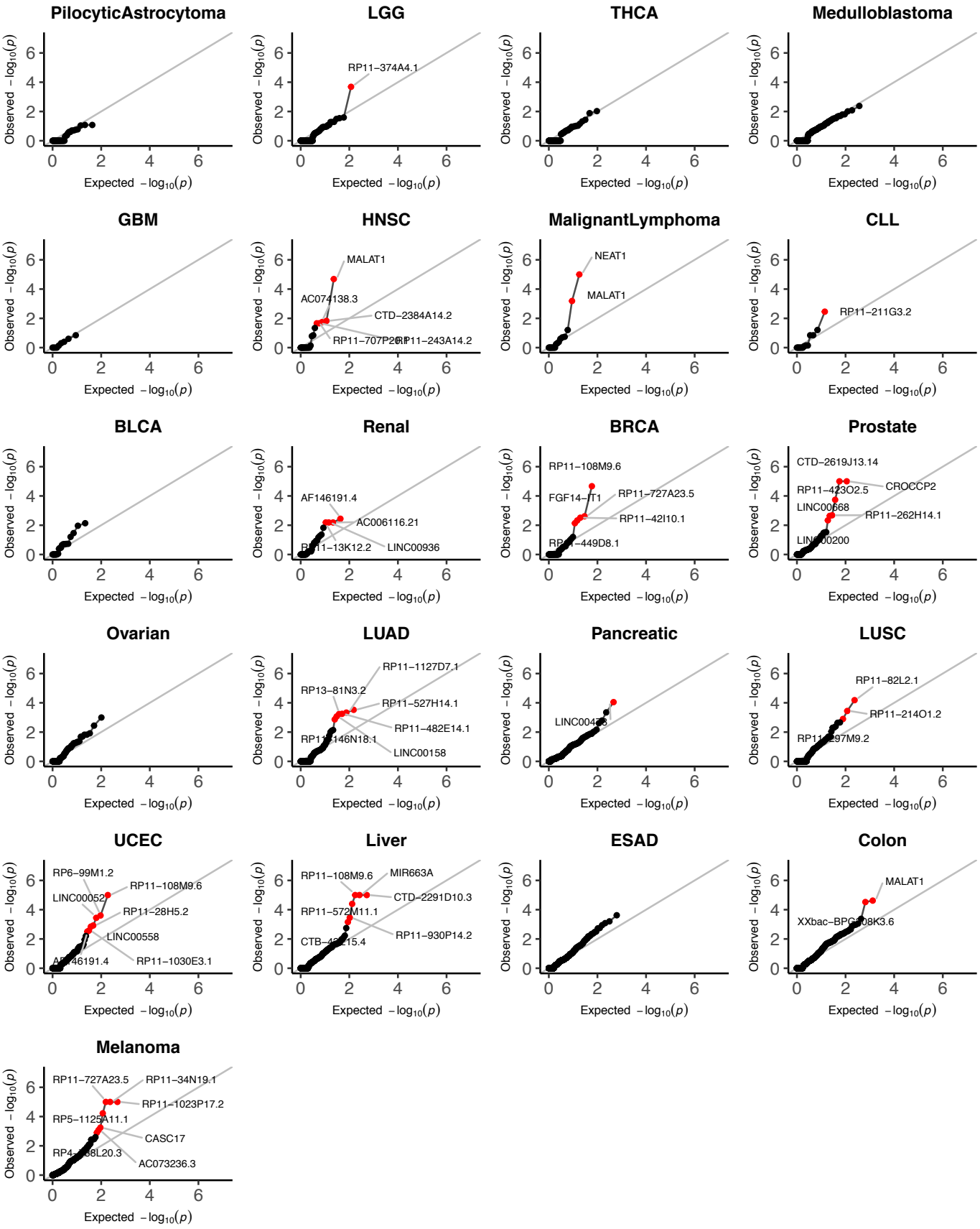


Figure S7. QQ plots of lincRNAs from CNCDriver, Related to Figure 1

QQ plots showing expected versus observed P values from lincRNA results of CNCDriver. Candidates with Q value ≤ 0.1 are labeled and shown in red. We note that NEAT1 and MALAT1 show enrichment of mutations due to AID somatic hypermutation in malignant lymphoma and their mutations are likely not drivers.

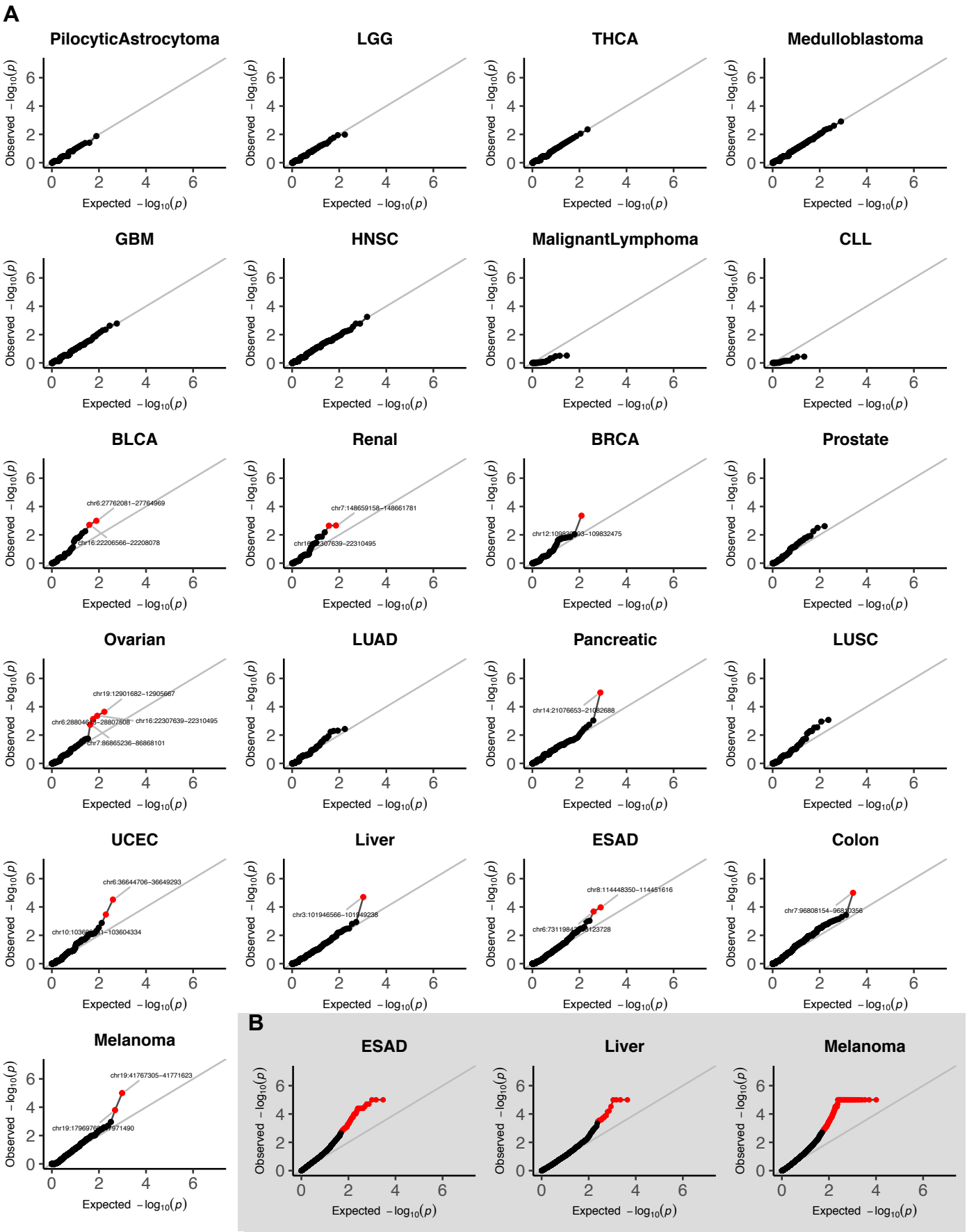

A
PilocyticAstrocytoma
LGG
THCA
Medulloblastoma
GBM
HNSC
MalignantLymphoma
CLL
BLCA
chr6:27762081–27764969
chr16:22206566–22208078
Renal
chr7:148659158–148661781
BRCA
chr12:109830693–109832475
Prostate
Ovarian
chr19:12901682–12905667
chr16:22307639–22310495
chr7:86865236–86868101
LUAD
Pancreatic
chr14:21076653–21082688
LUSC
UCEC
chr6:36644706–36649293
Liver
chr3:101946566–101949238
ESAD
chr8:114448350–114451616
Colon
chr7:96808154–96810358
Melanoma
chr19:41767305–41771623
B
ESAD
Liver
Melanoma
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

Figure S8. QQ plots of CTCF/cohesin insulators, Related to Figure 1
(A) CNCDriver results: QQ plots show p-values follow a uniform distribution. Candidates with Q value ≤ 0.1 are labeled and shown in red. 10 out of 21 cancer types show at least one significantly mutated insulator. (B) OncoDriveFML results: QQ plots shown for the three cancer types with highest enrichment of disrupted CTCF motifs (see Figure S2B). Candidates with Q value ≤ 0.1 are shown in red. Inflated QQ plots observed for esophageal cancer (ESAD) with 61 significant candidates, melanoma with 201 significant candidates and liver cancer with 18 significant candidates.

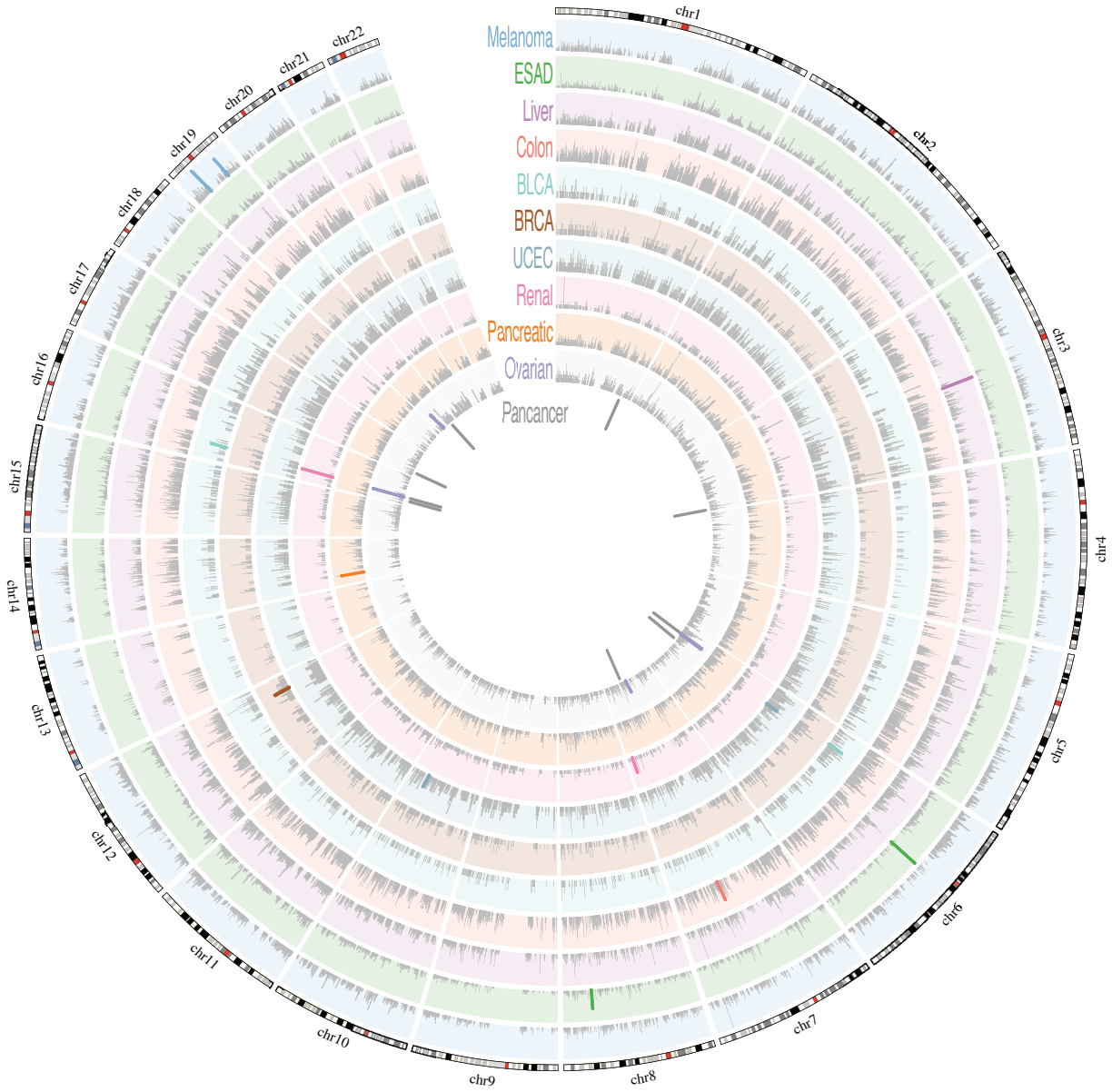


Figure S9. Landscape of mutation rate in insulators, Related to Figure 1
Complete landscape view for the fraction of mutated samples per insulator along the genome shown in the circos plot (light grey bars). The scale corresponds to the maximum fraction of mutated samples for each cancer type (up to 36% in Melanoma, 24% in BLCA, 21% in Colon, 18% in ESAD, 15% in UCEC, 13% in LUAD, 10% in Ovarian, 9% in Renal, 4% in BRCA, 7% in Liver and 5% in Pancreatic). Concentric circles correspond to different cancer types (see Figure S2 legend for cancer type abbreviations). CNCDriver identifies significantly mutated insulators that are highlighted as thicker bars in the corresponding cancer's color and in dark grey for pan-cancer analysis. In total, there are 21 significantly mutated insulators.

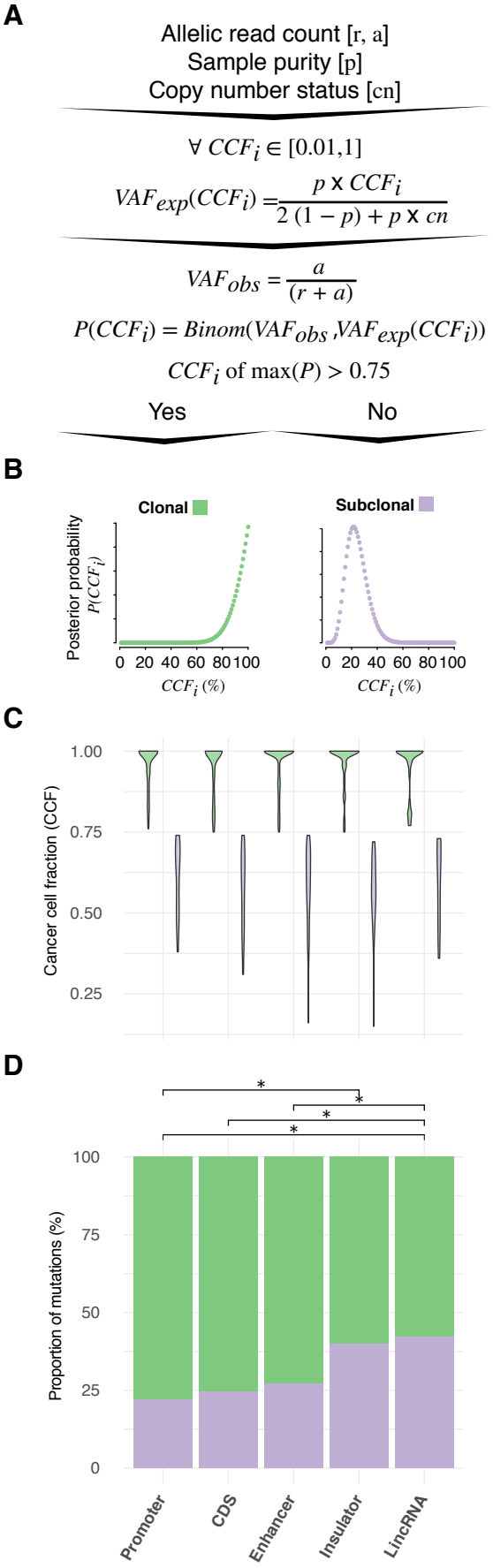


Figure S10. Clonality status of the mutations in the candidates predicted by CNCDriver, Related to Figures 1
(A) Workflow to assess the clonality status. Reference allelic read counts [r], alternative allelic read counts [a], cancer cell fraction [CCFi] within the uniform vector of 100 i values [0.01, 1], expected variant allelic fraction [VAFexp] given purity [p] and copy number [cn], [VAFobs] observed variant allele fraction given allelic read count. (B) Examples of posterior probability curve for clonal and subclonal mutations. For clonal mutations, the posterior probability curve tends to support higher CCF for a given mutation VAFobs. For sub-clonal mutations, for a given mutation VAFobs, lower CCF has higher probability in the posterior probability curve. (C) CCF distribution for mutations in candidates for each element: promoter, coding sequence (CDS), enhancer, insulator and lincRNA. (D) Proportion of clonal and sub-clonal mutations in driver candidates for each element (Marked with * p-value < 0.05, Fisher's exact test).

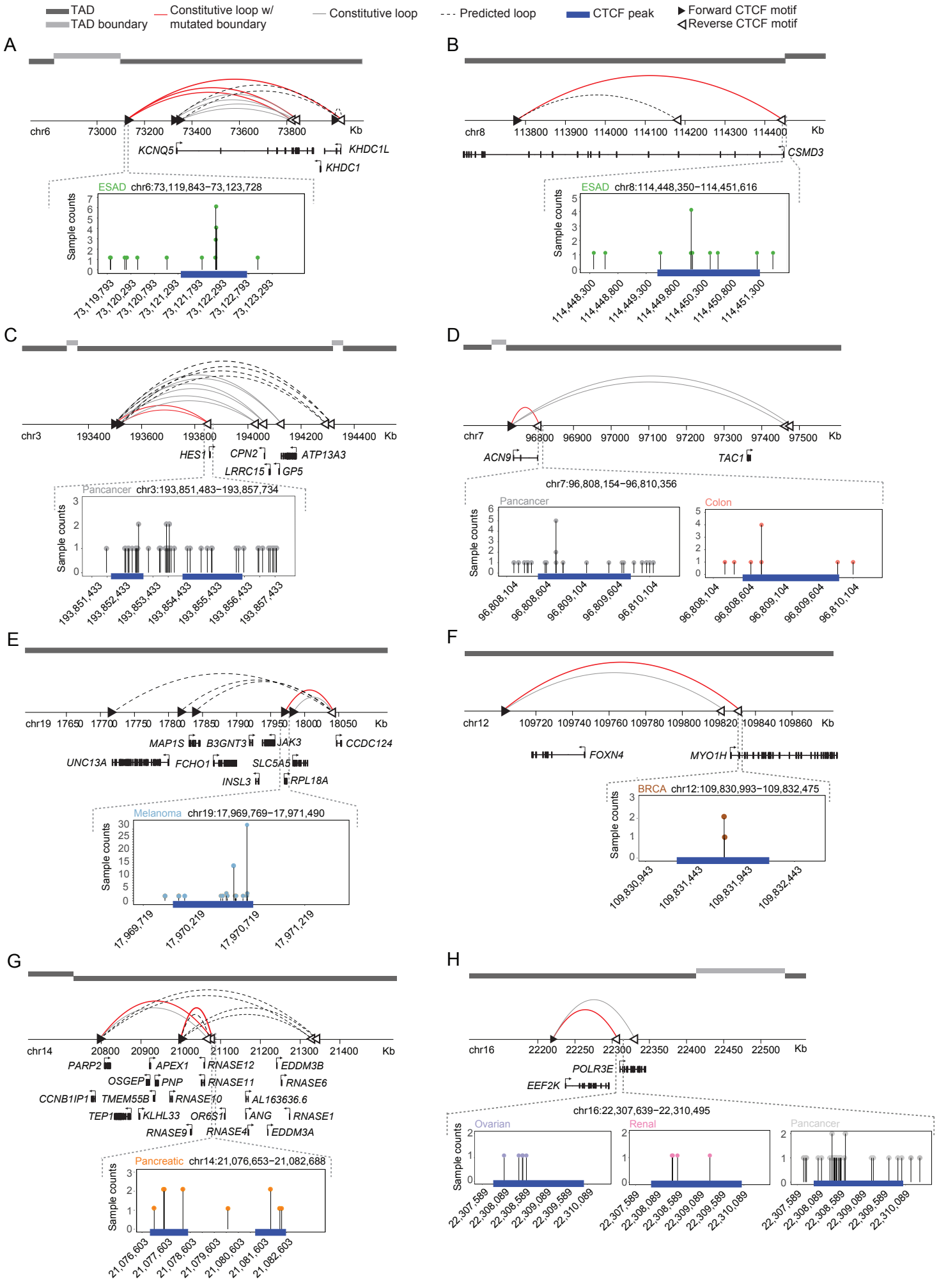
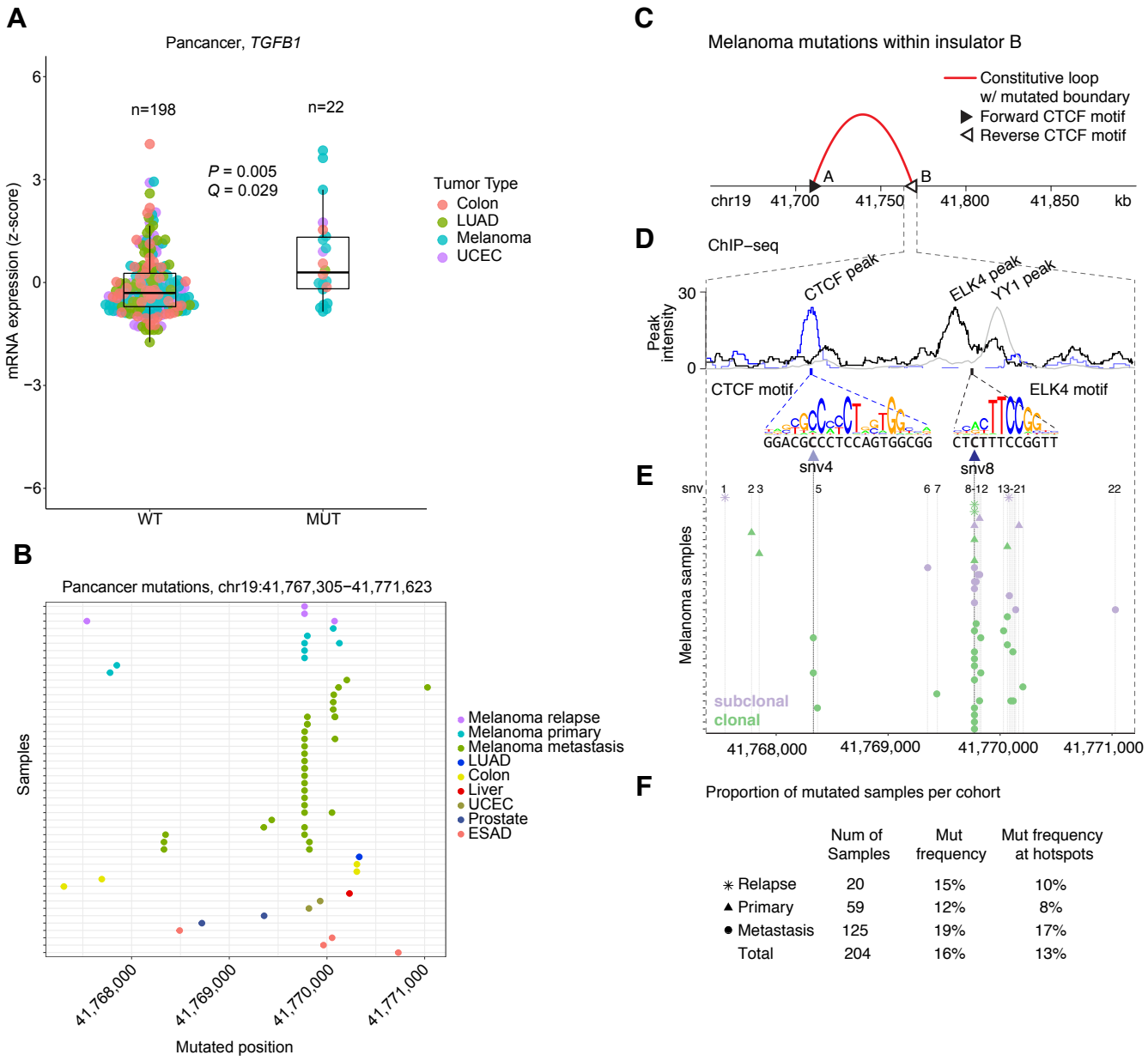

TAD
TAD boundary
Constitutive loop w/ mutated boundary
Constitutive loop
Predicted loop
CTCF peak
Forward CTCF motif
Reverse CTCF motif
A
chr6
73000
73200
73400
73600
73800
Kb
KCNQ5
KHDC1L
KHDC1
ESAD chr6:73,119,843−73,123,728
Sample counts
73,119,793
73,120,293
73,120,793
73,121,293
73,121,793
73,122,293
73,122,793
73,123,293
B
chr8
113800
113900
114000
114100
114200
114300
114400
Kb
CSMD3
ESAD chr8:114,448,350−114,451,616
Sample counts
114,448,300
114,448,800
114,449,300
114,449,800
114,450,300
114,450,800
114,451,300
C
chr3
193400
193600
193800
194000
194200
194400
Kb
HES1
CPN2
ATP13A3
LRRC15
GP5
Pancancer chr3:193,851,483−193,857,734
Sample counts
193,851,433
193,852,433
193,853,433
193,854,433
193,855,433
193,856,433
193,857,433
D
chr7
96800
96900
97000
97100
97200
97300
97400
97500
Kb
ACN9
TAC1
chr7:96,808,154−96,810,356
Pancancer
Colon
Sample counts
96,808,104
96,808,604
96,809,104
96,809,604
96,810,104
E
chr19
17650
17700
17750
17800
17850
17900
17950
18000
18050
Kb
MAP1S
B3GNT3
JAK3
CCDC124
UNC13A
FCHO1
SLC5A5
INSL3
RPL18A
Melanoma chr19:17,969,769−17,971,490
Sample counts
17,969,719
17,970,219
17,970,719
17,971,219
F
chr12
109720
109740
109760
109780
109800
109820
109840
109860
Kb
FOXN4
MYO1H
BRCA chr12:109,830,993−109,832,475
Sample counts
109,830,943
109,831,443
109,831,943
109,832,443
G
chr14
20800
20900
21000
21100
21200
21300
21400
Kb
PARP2
APEX1
RNASE12
EDDM3B
OSGEP
PNP
RNASE11
RNASE6
CCNB1IP1
TMEM55B
RNASE10
AL163636.6
TEP1
KLHL33
OR6S1
ANG
RNASE1
RNASE9
RNASE4
EDDM3A
Pancreatic chr14:21,076,653−21,082,688
Sample counts
21,076,603
21,077,603
21,078,603
21,079,603
21,080,603
21,081,603
21,082,603
H
chr16
22200
22250
22300
22350
22400
22450
22500
Kb
POLR3E
EEF2K
chr16:22,307,639−22,310,495
Ovarian
Renal
Pancancer
Sample counts
22,307,589
22,308,089
22,308,589
22,309,089
22,309,589
22,310,089

Figure S11. Mutations in predicted insulator drivers, Related to Figure 1
(A-H) Top: Graphical representation of the possible new loops formation (gray: alternate loops, dotted: predicted loops) associated with significantly mutated insulators. Original CTCF-CTCF loops predicted to be disrupted are shown in red. Black solid triangle represents CTCF motif in the forward orientation and white triangle represents CTCF motif in the reverse orientation. Dark grey bars represent a topological associated domain (TAD) and light grey represent a TAD boundary. Bottom: Zoom-in needle plots to show the location of each mutation within the significantly mutated insulator and the number of mutated samples at each position. The regions of CTCF ChIP-seq peaks are labeled in blue color. Here, we show driver candidates with recurrent mutations in the insulator regions. 5 out of 21 predicted drivers do not contain CTCF motifs in the convergent orientation at original disrupted boundary pairs and were excluded from loop predictions. Detailed information of genes involved in the predicted loop-rewiring events is in supplementary table S8.

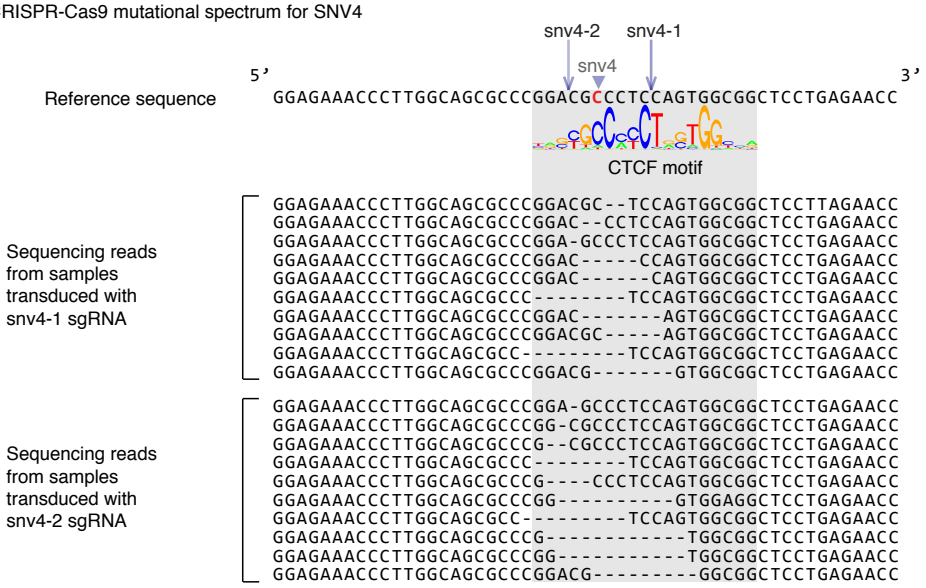


	Num of Samples	Mut frequency	Mut frequency at hotspots
✳ Relapse	20	15%	10%
▲ Primary	59	12%	8%
● Metastasis	125	19%	17%
Total	204	16%	13%

Figure S12. Mutations in TGFB1 associated insulator, Related to Figures 1 and 2

(A) Box plot of TGFB1 differential expression between tumor samples with mutations (MUT) and those without mutations (WT) in the significantly mutated insulator (chr19:41,767,305-41,771,623) in pan-cancer analysis. (B) The zoom-in to the significantly mutated insulator in pan-cancer analysis. (C) Loop conformation in A375 melanoma cells around the TGFB1 associated insulator region B. (D) The zoom-in to the significantly mutated insulator B shows the corresponding CTCF peak and co-located ELK4 and YY1 peaks. Locations of the CTCF and ELK4 motifs are also shown. (E) Single nucleotide variant (SNV) positions are shown. SNVs are numbered starting from left to right. There are 7 hotspot positions (recurrently mutated in multiple samples) (chr19: 41,768,332; 41,769,771; 41,769,772; 41,769,800; 41,769,821; 41,770,065; and 41,770,083). The mutations predicted as clonal are shown in green and subclonal in purple. (F) Summary table of proportion of mutated samples in each melanoma cohort.

A Examples of CRISPR-Cas9 mutational spectrum for SNV4



B Examples of CRISPR-Cas9 mutational spectrum for SNV8

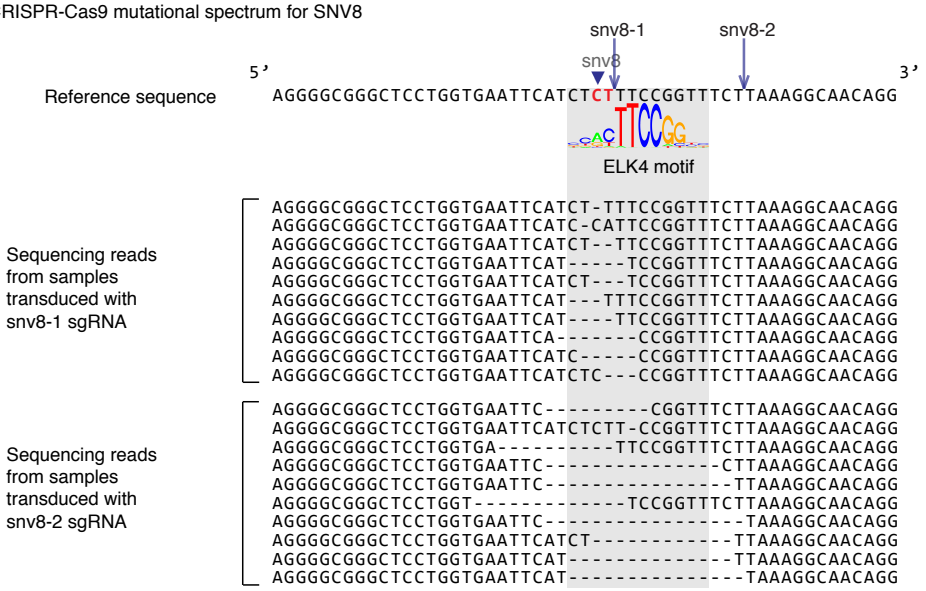


Figure S13. Deep sequencing of indel mutations at each CRISPR guided RNA target site, Related to Figure 3

(A-B) We targeted two mutations (SNV4 and SNV8) in the chr19 insulator and designed 2 sgRNAs for each. We performed deep sequencing to determine the resulting indel distribution of each sgRNA. (A) sgRNAs SNV4-1 and SNV4-2 target SNV4 located in a CTCF motif. (B) sgRNAs SNV8-1 and SNV8-2 target SNV8 which is located in an ELK4 motif.