



# HHS Public Access

Author manuscript

*Curr Opin Syst Biol.* Author manuscript; available in PMC 2018 October 26.

Published in final edited form as:

*Curr Opin Syst Biol.* 2017 February ; 1: 9–15. doi:10.1016/j.coisb.2016.12.017.

## Non-coding genetic variation in cancer

Tawny N. Cuykendall<sup>1,2</sup>, Mark A. Rubin<sup>3,4,5</sup>, and Ekta Khurana<sup>1,2,3,5</sup>

<sup>1</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, 10065, USA

<sup>2</sup>Institute for Computational Biomedicine, Weill Cornell Medicine, New York, New York, 10021, USA

<sup>3</sup>Meyer Cancer Center, Weill Cornell Medicine, New York, New York, 10065, USA

<sup>4</sup>Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, New York, 10065, USA

<sup>5</sup>Institute for Precision Medicine, Weill Cornell Medicine, New York, New York, 10065, USA

### Abstract

The vast majority of somatic variants in cancer genomes occur in non-coding regions. However, progress in cancer genomics in the past decade has been mostly focused on coding regions, largely due to the prohibitive cost of whole genome sequencing (WGS). Recent technological advances have decreased sequencing costs leading to the current acquisition of thousands of tumor whole genome sequences which has led to a hunt for non-coding drivers. The most well characterized regulatory drivers are in the *TERT* promoter and have been identified in many cancer types. Despite the larger fraction of somatic variants occurring in non-coding regions, the number of non-coding drivers identified so far is much less than the number of coding region drivers. Here we discuss reasons that may hinder the detection of non-coding drivers. We also examine the relationship between non-coding genetic variation and epigenetic state in tumor cells and assert the need for additional epigenetic data sets as a prerequisite for understanding the rewiring of regulatory networks in cancer.

### Introduction

Non-coding elements encompass *cis*-regulatory regions (promoters, enhancers, insulators and silencers) (Figure 1A) as well as ncRNAs. Many of these elements have been identified by a combination of functional genomics approaches and in silico methods based on sequence conservation [1]. In addition, non-coding regions exhibit histone modification patterns characteristic of specific functional elements. For example, trimethylated Lys4 of histone 3 (H3K4me3) marks active promoters while H3K27me3 is associated with repressed

Corresponding Author: Correspondence to E.K., [ekk2003@med.cornell.edu](mailto:ekk2003@med.cornell.edu).

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

regions [2]. As in coding sequence (CDS), non-coding regions can be disrupted by single nucleotide variants (SNVs), indels (insertions and deletions less than 50 base pairs (bp)), and larger structural variants (SVs). These alterations are characteristic of both germline and somatic variants. While non-coding germline variants are clearly important and have been associated with multiple cancer types [3–6], we focus on somatic variants. Sequence variants occurring within functional non-coding elements, such as promoters and enhancers, have the potential to alter gene expression and affect epigenetic states [7].

A major goal in cancer genomics is to identify driver mutations, which are variants that provide a selective advantage to tumor cell growth and proliferation. Over the past few years, the number of tumor whole genomes has increased substantially, leading to a flurry of studies to identify drivers in non-coding regions. However, distinguishing driver from passenger mutations in non-coding regions is challenging for several reasons: 1) there is a larger number of mutations in non-coding elements than coding regions, 2) non-coding regions are incompletely annotated, and 3) non-coding regions generally function within complex regulatory networks and therefore, current methods that were developed based on CDS properties may be less robust to non-coding driver identification.

## Non-coding drivers in cancer

### How many non-coding drivers have been identified?

Several methods exist to identify non-coding drivers and are generally based on identifying mutations that are highly recurrent [8] or those predicted to have a high functional impact [9–11]. The most well characterized regulatory driver mutations in cancer occur in the *TERT* promoter [12,13]. Recurrent mutations in the *TERT* promoter were originally identified in melanoma, but have now been identified in multiple cancers [6,12,14,15]. *TERT* encodes the reverse transcriptase subunit of telomerase [16] and is normally silenced in somatic cells. The promoter mutations create novel binding sites for the ETS family of transcription factors, resulting in *TERT* up-regulation [12,15]. Recently, Chiba et al. showed that the recurrent *TERT* promoter mutations cause increased telomere length in differentiated cells. Increased telomeres lead to reduced rates of apoptosis, pointing towards a molecular mechanism for *TERT*'s role in tumorigenesis [17]. Following the identification of *TERT* promoter mutations, the hunt was on for analogous mutations in other regulatory regions. While additional regulatory regions have been identified as recurrently mutated [10,18–23], no element has yet been found that is recurrently mutated to the level of the *TERT* promoter (up to 79% in Myxoid Liposarcoma [15]). The alternative approach, identifying mutations with high functional impact, has also yielded some regulatory drivers. There are several computational tools to annotate and prioritize functional non-coding variants [1]. One such method is FunSeq2, which assigns a weighted functional score to each variant based on several sequence features, such as functional annotation, conservation, and whether the variant breaks or creates a transcription factor (TF) binding motif [9,20]. Recently, a method utilizing the functional impact of mutations to identify non-coding drivers was applied to a set of 505 tumors representing 19 different cancers [11]. They identified *TERT* as the most significant element, as well as 54 additional candidates (q-value < 0.05) in promoters, intronic splice sites and the 5' and 3' UTRs [11]. In contrast, they detected 156 protein-

coding driver genes (q-value < 0.05). Consistent with these results, the Cancer Gene Census contains 573 genes (cancer.sanger.ac.uk) implicated in cancer [24,25], while the number of known non-coding driver elements is much smaller. Collectively, these observations prompt two possibilities: 1) there are fewer non-coding drivers compared to coding drivers or 2) there are additional non-coding drivers yet to be discovered.

We suggest there are additional non-coding drivers that have not yet been identified. This is partly because the genetic basis of a large number of tumors remains unexplained based on driver genes. For example, Rubio-Perez et al identified 454 driver genes in a pan-cancer cohort of ~4000 samples [26]. Of these samples, ~90% have at least one non-synonymous mutation, copy number alteration (CNA) or gene fusion in one of these genes [26]. We note that even 90% is probably an overestimate because not all of the alterations in driver genes are likely to be oncogenic. Thus, the genomic basis of a large number of tumors remains unexplained. Because only a handful of non-coding drivers have been discovered, it is possible that non-coding drivers exhibit distinct properties from known CDS drivers, which obscures their detection with current methods.

### **Non-coding regions exhibit distinct properties from coding regions**

Drivers differ in their magnitude of effect on tumorigenesis and can also be context-specific. Castro-Giner et al. distinguish two different classes of drivers: ‘major drivers’ and ‘mini drivers’ [27]. ‘Major drivers’ are mutations with a large effect size that confer a large selective advantage to the tumor cell [27,28]. In contrast, ‘mini drivers’ are mutations with a small effect size and therefore provide only a small selective advantage [27]. Castro-Giner et al. assert the possibility that many somatic mutations in cancer might actually be ‘mini drivers’, rather than passengers, and that the composite effect of multiple ‘mini drivers’ can have a large effect on tumorigenesis [27]. We anticipate that a subset of non-coding mutations in cancer are in fact ‘mini drivers.’ For example, a variant in a transcription factor binding site (TFBS) may alter TF binding affinity, while not completely obliterating it, leading to a subtle change in gene expression that is only somewhat advantageous to the cell (Figure 1B). It is possible that the majority of individual non-coding drivers have a smaller impact on tumorigenesis than the ‘major’ known CDS drivers. We note that even for coding genes, the magnitude of impact on tumorigenesis is not clear for most drivers since they have largely been identified by the presence of sequence variants (e.g. Cancer Gene Census) [25].

In order to identify the ‘mini drivers’ in non-coding regions, we need to understand their functional impact. Several studies have looked for associations between mutations in regulatory regions and gene expression. There is a positive correlation between mutations in the *TERT* promoter and gene expression, though this correlation is not significant across all analyses [18,21–23]. In fact, only a few non-coding regions with recurrent mutations actually show significant associations with gene expression [18,21–23]. While it is possible that these mutations are not functional and are recurrent due to underlying mutational signatures or mutation rate co-variates (such as impaired DNA repair in TFBS [29–31]), it is also possible that the architecture of the regulatory network leads to this lack of association. For example, because multiple enhancers can regulate a gene, a mutation in any one of the

enhancers could disrupt TF binding and cause a slight change in gene expression (Figure 1B). Bailey et al. identified the set of regulatory elements targeting *ESR1* in breast cancer and found that collectively, these elements contain a significant enrichment of mutations [32]. All of the somatic non-coding mutations are predicted to modulate binding of *ESR1* regulators and furthermore, the majority of those tested showed a significant effect on *ESR1* gene expression [32]. Thus, the authors speculate that non-coding drivers may be characterized by selection acting on a phenotypic outcome (e.g. *ESR1* expression), rather than on a specific variant, as has been observed for CDS drivers [32].

### Why we think additional non-coding drivers are yet to be revealed

**Larger sample sizes and higher sequencing depth**—Non-coding functional elements, with the exception of the *TERT* promoter, appear to contain lower frequency variants than many driver protein-coding genes [10,18,20,21,23,33]. It is possible that a large portion of putative non-coding drivers are analogous to the ‘hills’ of cancer genes [34], those that are infrequently mutated in cancer (e.g., *PRDMI* and *HRAS*) [28]. In contrast, because *TERT* mutations are frequent across multiple cancer types, they are analogous to the ‘mountains’ of cancer genes that are mutated at high frequencies (such as *TP53* and *KRAS*) [34]. We note that most currently known ‘hills’ of cancer genes were found through the analysis of ~5000 exomes [28]. In contrast, the current largest collection of tumor whole genomes is only ~2800 by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). Thus, identifying the equivalent non-coding ‘hills’ in cancer genomes will likely require additional tumor WGS. However, even with increased sample sizes, variant calling in non-coding regions can be challenging. For example, promoters tend to be GC-rich and therefore exhibit lower sequencing coverage than the genomic average [35]. Therefore, in order to accurately call variants in GC-rich regulatory regions, higher sequencing depth is essential.

**New computational approaches**—We propose a difference in thinking for non-coding regions may be required to fully harness the power of WGS data to identify non-coding drivers. As discussed above, multiple enhancers can regulate one gene. In fact, nearly half of enhancers regulate more than one gene [36]. Typically, methods to identify drivers look for recurrent mutations in single elements. Therefore, new methods are needed that account for the architecture of tissue-specific regulatory networks. As demonstrated by Bailey et al. (described in the section Non-coding regions exhibit distinct properties from coding regions), pooling elements, such as enhancers, can increase power to detect non-coding drivers [32]. However, this is currently hindered because it is challenging to 1) identify all the enhancers regulating a gene and 2) link enhancers to their target genes. Multiple methods have been used to define enhancer-promoter interactions. For example, the correlation of eQTLs or histone marks at enhancer regions with target gene expression can be used to infer linkages [37,38]. Another approach is Hi-C, the high throughput derivative of the chromosome conformation (3C) technology, which provides a readout of the physical interactions within the three-dimensional (3D) chromatin structure [39]. However, this technology presents its own computational challenges [40–43]. Furthermore, it is unfeasible to experimentally validate all predicted interactions. Thus, linking distal regulatory elements to their target genes remains an active area of research.

As the number of tumor whole genomes increases, and the methods to identify non-coding drivers are refined, we anticipate the discovery of several additional regulatory element drivers across tumor types.

## **Relationship between genetic and epigenetic changes in non-coding regions in cancer**

### **Epigenetic states in cell-of-origin dictate the mutational landscape in non-coding regions in cancer**

In order to accurately identify drivers, it is necessary to understand the underlying mutational processes that shape the cancer genome. There are thousands of somatic variants present in a typical tumor genome; these variants are the combined result of mutational processes and the breakdown of the DNA repair machinery. Alexandrov et al. identified several mutational signatures, which are characterized by their unique composition of trinucleotide sequence changes [44]. The recent accumulation of tumor WGS has also revealed that the epigenetic state of the tumor cell-of-origin leaves characteristic footprints in the genome (Figure 1B). Polak et al. analyzed 173 WGS from 8 different cancers and correlated hundreds of epigenetic features, including DNAase I hypersensitive sites (DHSs) and histone marks, with mutation density. They showed that the epigenetic state of the tumor cell-of-origin, in conjunction with replication timing, explains up to 86% of the variance in mutation density [45] and confirmed that regions of open chromatin exhibit decreased mutation density compared to closed regions [46,47]. This decrease in mutation density is likely due to the increased ability of the DNA repair machinery to access and repair DNA damage in active regulatory regions. However, two studies earlier this year revealed that the transcriptional machinery can actually inhibit DNA repair, leading to increased mutation density in TFBS. Sabrinathan et al. analyzed WGS from both melanoma and lung cancer samples and found an increase in mutation frequency at active TFBS; they showed that this is caused by deficient nucleotide excision repair (NER) in these regions [31]. Similarly, Perera et al. analyzed WGS from 14 different cancers and found that the centers of active promoters exhibit increased mutation density [30]. Thus, while broad DHS regions have decreased mutation rates as reported previously [45,46], these recent studies at higher resolution show that repair is hindered at the sites of actual TF binding [29–31]. Because drivers are generally identified by their frequency, knowing how the underlying molecular mechanisms co-vary with mutation rate is a prerequisite for developing accurate methods to detect drivers [8]. It is reasonable to hypothesize that with increasing tumor whole genome sequences, we will uncover additional novel links between DNA repair and other cellular processes that will enable more accurate modeling of background mutation rates.

### **Non-coding sequence variants can affect the epigenetic state of the cancer cell**

While somatic mutations reflect the epigenetic state of the tumor cell-of-origin, the resulting mutations can then in turn alter the epigenetic state of the tumor cells (Figure 1C). A common functional consequence of non-coding variation in promoters is a change in gene expression. In prostate cancer, the ERG transcription factor is frequently overexpressed due to fusion with the 5' UTR of *TMPRSS2* [48] (Figure 1B). Because ERG regulates many

downstream genes, Rickman et al. tested whether its overexpression has an effect on chromatin structure by performing Hi-C in prostate epithelial cell lines overexpressing ERG [7]. Indeed, they found that the ERG-overexpressed and control cells had significant differences in *cis* (intrachromosomal) as well as *trans* (interchromosomal) interactions, though the latter are less reliable. This result demonstrated that ERG overexpression causes topological changes in chromatin [7] (Figure 1C). These results demonstrate that the fusion of non-coding sequence from one gene (*TMPRSS2*) with the coding region of another gene (*ERG*) can lead to epigenetic changes in the tumor cell.

In another example, Taberlay et al. found that while topologically associated domains (TADs) are mostly conserved between prostate cancer cell lines and normal cells, cancer cells exhibit novel sub-domains and that the boundaries of these domains are enriched for CTCF binding [49]. CTCF is implicated in mediating long-range interactions in the genome and therefore, influences the 3D chromatin structure [50]. Multiple studies have found that CTCF binding sites are enriched for mutations in cancer [51,52] and recently Hnisz et al. showed that DNA mutations altering CTCF binding sites can perturb normal DNA looping (which can in turn activate proto-oncogenes) [53]. Therefore, mutations in CTCF binding sites likely mediate the altered chromatin topology observed in cancer cells.

These studies suggest that non-coding variants can alter epigenetic states in tumor cells. In fact, this is likely a more general phenomenon that also occurs in normal cells. Hashimoto et al. recently showed that at least one measure of the epigenetic state of a cell, i.e. chromatin accessibility, can be inferred directly from the DNA sequence [54]. However, beyond the examples discussed above, there are relatively few examples in the literature of somatic mutations in cancer causing epigenetic changes. This could be due to the current dearth of epigenetic data sets in cancer (as discussed below).

## Concluding remarks

As discussed in this review, increased numbers of tumor whole genomes are required for the detection of additional non-coding drivers. One step in this direction is the effort by The Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium, a collaboration between TCGA and ICGC, to analyze non-coding variants in ~2800 tumor and matched normal whole genomes. We anticipate that as the numbers of tumor whole genomes increase even more in the future, novel non-coding drivers will be uncovered.

ENCODE and the Roadmap Epigenomics consortia have generated a comprehensive data set of epigenomic profiles, including histone modification patterns and DNA accessibility, in multiple tissue types and cell lines (including some cancer cell lines). While methylation and transcriptome data are available for a large fraction of cancer samples from TCGA and ICGC, other types of epigenetic data (such as histone modifications) are not available at this large-scale for cancer samples. This is likely due to the requirement for large amounts of tissue for ChIP-seq experiments, as well as the fact that these experiments are generally more technically challenging and time intensive compared to RNA-seq and DNA methylation assays. This makes it hard to obtain an accurate map of epigenetic states in cancer samples, which is essential for understanding the impact of non-coding genetic



variants on rewiring of the regulatory network as cells transform from normal to malignant state. Creation of the tumor-specific epigenetic maps with new technologies (such as Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) [55,56]) holds the promise to bridge this gap and allow a more complete understanding of the role of non-coding regions in cancer.

## Acknowledgments

The authors wish to thank Cassandra Burdziak and Matthew MacKay for helpful comments on the manuscript as well as Priyanka Dhingra, Alexander Martinez Fundichely, Eric Minwei Liu and members of PCAWG for insightful discussion of the concepts covered.

## References

- 1. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. *Nat Rev Genet.* 2016; 17:93–108. Comprehensive recent review of non-coding genetic variation in cancer. [PubMed: 26781813]
2. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]
3. Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, Arva NC, Bargonetti J, Bartel F, Taubert H, Wuerl P, et al. A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell.* 2004; 119:591–602. [PubMed: 15550242]
4. Bond GL, Levine AJ. A single nucleotide polymorphism in the p53 pathway interacts with gender, environmental stresses and tumor genetics to influence cancer in humans. *Oncogene.* 2007; 26:1317–1323. [PubMed: 17322917]
5. Grisanzio C, Freedman ML. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer.* 2010; 1:555–559. [PubMed: 21779458]
6. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, et al. TERT promoter mutations in familial and sporadic melanoma. *Science.* 2013; 339:959–961. [PubMed: 23348503]
7. Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, et al. Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A.* 2012; 109:9083–9088. [PubMed: 22615383]
8. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013; 499:214–218. [PubMed: 23770567]
9. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* 2014; 15:480. [PubMed: 25273974]
10. Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res.* 2015; 43:8123–8134. [PubMed: 26304545]
11. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* 2016; 17:128. [PubMed: 27311963]
12. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. *Science.* 2013; 339:957–959. [PubMed: 23348506]
13. Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, et al. Frequency of TERT promoter mutations in human cancers. *Nat Commun.* 2013; 4:2185. [PubMed: 23887589]
14. Heidenreich B, Rachakonda PS, Hemminki K, Kumar R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev.* 2014; 24:30–37. [PubMed: 24657534]

15. Killela PJ, Reitman ZJ, Jiao Y, Bettegowda C, Agrawal N, Diaz LA Jr, Friedman AH, Friedman H, Gallia GL, Giovanella BC, et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc Natl Acad Sci U S A*. 2013; 110:6021–6026. [PubMed: 23530248]
16. Martinez P, Blasco MA. Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins. *Nat Rev Cancer*. 2011; 11:161–176. [PubMed: 21346783]
17. Chiba K, Johnson JZ, Vogan JM, Wagner T, Boyle JM, Hockemeyer D. Cancer-associated TERT promoter mutations abrogate telomerase silencing. *Elife*. 2015:4.
18. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet*. 2014; 46:1258–1263. [PubMed: 25383969]
19. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraiishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet*. 2016; 48:500–509. [PubMed: 27064257]
- 20. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013; 342:1235587. One of the first methods for the identification of non-coding drivers. [PubMed: 24092746]
21. Melton C, Reuter JA, Spacek DV, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet*. 2015; 47:710–716. [PubMed: 26053494]
22. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. 2016; 534:47–54. [PubMed: 27135926]
23. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet*. 2014; 46:1160–1165. [PubMed: 25261935]
24. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015; 43:D805–811. [PubMed: 25355519]
25. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. *Nat Rev Cancer*. 2004; 4:177–183. [PubMed: 14993899]
26. Rubio-Perez C, Tamborero D, Schroeder MP, Antolin AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A, Lopez-Bigas N. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*. 2015; 27:382–396. [PubMed: 25759023]
27. Castro-Giner F, Ratcliffe P, Tomlinson I. The mini-driver model of polygenic cancer evolution. *Nat Rev Cancer*. 2015; 15:680–685. [PubMed: 26456849]
28. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014; 505:495–501. [PubMed: 24390350]
29. Khurana E. Cancer genomics: Hard-to-reach repairs. *Nature*. 2016; 532:181–182. [PubMed: 27075092]
- 30. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JW. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*. 2016; 532:259–263. One of the first papers to demonstrate how interplay between DNA repair and transcription initiation affects the mutation rate. [PubMed: 27075100]
- 31. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*. 2016; 532:264–267. One of the first papers to demonstrate how interplay between DNA repair and transcription initiation affects the mutation rate. [PubMed: 27075101]
- 32. Bailey SD, Desai K, Kron KJ, Mazrooei P, Sinnott-Armstrong NA, Treloar AE, Dowar M, Thu KL, Cescon DW, Silvester J, et al. Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. *Nat Genet*. 2016 The authors propose a method to detect non-coding driver elements that tests for enrichment of mutations in the set of all known regulatory elements of a target gene.

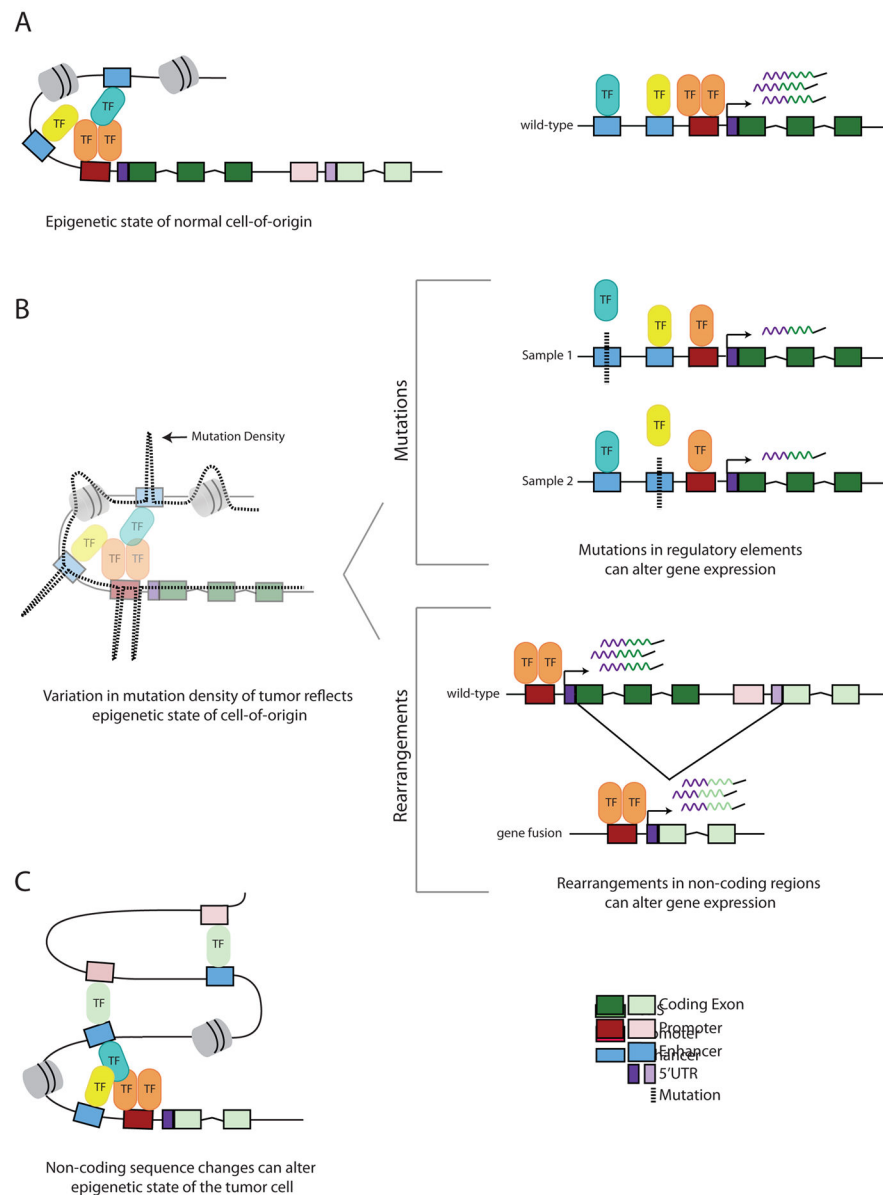


33. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012; 149:979–993. [PubMed: 22608084]
34. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science*. 2013; 339:1546–1558. [PubMed: 23539594]
35. Wang W, Wei Z, Lam TW, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep*. 2011; 1:55. [PubMed: 22355574]
36. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Roadmap Epigenomics C. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518:317–330. [PubMed: 25693563]
37. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015; 348:648–660. [PubMed: 25954001]
38. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol*. 2012; 13:R48. [PubMed: 22950945]
39. de Laat W, Dekker J. 3C-based technologies to study the shape of the genome. *Methods*. 2012; 58:189–191. [PubMed: 23199640]
40. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. *BMC Genomics*. 2012; 13:436. [PubMed: 22935139]
41. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in HiC data via Poisson regression. *Bioinformatics*. 2012; 28:3131–3133. [PubMed: 23023982]
42. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*. 2012; 9:999–1003. [PubMed: 22941365]
43. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*. 2011; 43:1059–1065. [PubMed: 22001755]
44. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. Signatures of mutational processes in human cancer. *Nature*. 2013; 500:415–421. [PubMed: 23945592]
- 45. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahovicek K, Stamatoyannopoulos JA, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*. 2015; 518:360–364. The authors show that variation of somatic mutation density is dictated by the cell of origin. [PubMed: 25693567]
46. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, Garraway LA, Mirkin S, Getz G, Stamatoyannopoulos JA, et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat Biotechnol*. 2014; 32:71–75. [PubMed: 24336318]
47. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*. 2012; 488:504–507. [PubMed: 22820252]
48. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005; 310:644–648. [PubMed: 16254181]
- 49. Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res*. 2016; 26:719–731. Using Hi-C, the authors characterize epigenetic remodelling in prostate cancer cell lines. [PubMed: 27053337]
50. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*. 2014; 15:234–246. [PubMed: 24614316]
51. Kaiser VB, Taylor MS, Semple CA. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet*. 2016; 12:e1006207. [PubMed: 27490693]

52. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet.* 2015; 47:818–821. [PubMed: 26053496]
53. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science.* 2016; 351:1454–1458. [PubMed: 26940867]
54. Hashimoto TB, Sherwood R, Kang DD, Rajagopal N, Barkal AA, Zeng H, Emons BJ, Srinivasan S, Jaakkola T, Gifford D. A synergistic DNA logic predicts genome-wide chromatin accessibility. *Genome Res.* 2016
55. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods.* 2013; 10:1213–1218. [PubMed: 24097267]
56. Rendeiro AF, Schmidl C, Strefford JC, Walewska R, Davis Z, Farlik M, Oscier D, Bock C. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun.* 2016; 7:11938. [PubMed: 27346425]

### Highlights

- Fewer non-coding drivers have been identified compared to coding region drivers
- The unique properties of non-coding regions can complicate driver detection
- Non-coding genetic variation in tumor cells can alter epigenetic states
- Larger sample sizes, higher sequencing depth, and epigenomic data sets for tumor samples are needed



**Figure 1. Schematic showing the interplay between non-coding genetic variation and epigenetic state in cancer**

**A)** The epigenetic state of the tumor cell-of-origin. Two genes and two nucleosomes (gray) are shown. On the right is a zoomed in diagram of one of these genes being actively transcribed. **B)** As the cell turns from normal to malignant, the variation in somatic mutation density reflects the epigenetic state of the tumor cell-of-origin. TFBS within active regions have elevated mutations rates due to inefficient NER (for example, in melanoma and lung cancer), while regions of open chromatin flanking active TFBS have decreased mutation rates. Regions of closed chromatin exhibit elevated mutation rates relative to accessible regions. Sequence variants can be point mutations or rearrangements. In the mutations panel, a mutation (dashed black line) in either enhancer causes a decrease in gene expression by modulating TF-binding affinity (e.g. only one TF binds the promoter). This regulatory

architecture can decrease power to detect associations between mutations in a single regulatory element and the target gene. In the rearrangements panel, there is an intergenic deletion resulting in a fusion between the 5' UTR of one gene with the exons of the other gene (e.g. *TMPRSS2-ERG*), resulting in overexpression of the fusion gene (e.g. *ERG* overexpression in prostate cancer). **C**) Non-coding sequence changes can alter epigenetic state in the cancer cell. Overexpression of the fusion gene product (as shown in **B**) (e.g. ERG, light green TF) can alter the 3D chromatin topology by mediating new promoter-enhancer linkages. Additional genes in this region are omitted for clarity.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript