(5000 words maximum)

Title

Role of noncoding variants in cancer

Preface (100 words)

Tumor genomes contain numerous somatic sequence variants. These include single nucleotide mutations, small insertions and deletions and larger sequence rearrangements. A large majority of these variants occur in noncoding parts of the genome. Noncoding variants can effect gene expression to variable extents and may have major functional consequences causing tumor progression. Although most previous studies have focused on the identification of functional variants in protein-coding genes, many recent studies suggest that the repertoire of noncoding somatic variants contains driver events playing an important role in tumor growth. Furthermore, numerous noncoding germline variants are known to play a role in cancer susceptibility. In many instances, tumor growth relies on an intricate balance between inherited germline and acquired somatic variants. In this review, we discuss the current understanding of the role of noncoding somatic and germline variants in cancer.

Introduction

The first tumor whole-genome was sequenced in 2008 (REF). As a result of the decreasing costs, whole-genomes of thousands of tumors have since been sequenced. The numbers of cancer patients that have undergone whole-genome sequencing (WGS) is only going to increase as precision medicine approaches are increasingly being adopted in the clinic (REF). Most of the variants obtained from WGS of tumor genomes lie in noncoding regions (Figure 1). In this review we provide an overview of the current understanding of the role of noncoding sequence variants in cancer development and growth. We note that most previous studies of somatic cancer variants have focused on exomes. However, there is an increased realization of the importance of noncoding variants in cancer and an ongoing collaboration between TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium), called Pan-Cancer Analysis of Whole Genomes (PCAWG), aims to identify noncoding mutations of functional consequence in ~2500 tumor and matched normal whole-genomes.

Genetic susceptibility for complex disorders has been probed previously by numerous genome-wide association studies (GWAS). These studies have revealed that most loci associated with complex traits lie in noncoding regions of the genome (REF). Many studies have also explored the link between inherited germline variants and cancer susceptibility. In agreement with other complex traits, these studies also revealed many noncoding loci associated with altered cancer risk (REF eg from Francesca). Thus, noncoding regions play an important role in cancer not only due to the somatic aberrations in tumor cells, but also the inherited germline variants they contain. In this review, we also discuss germline variants that have been associated with increased cancer susceptibility, specially the cases where there is an intricate relationship between germline polymorphisms and somatic variants.

Besides sequence alterations, other changes in the noncoding regions such as epigenetic and transcriptional variation can also influence cancer development. For example, many noncoding RNAs are known to be misregulated in various cancers (REF), H3K4me1 sites can be lost or gained in cancer cells relative to matched normal (REF), etc. However, in this review, we focus on effects of DNA sequence variants in noncoding regions and suggest reviews such as XX and XX for discussions of other cancer associated changes.

Before we go into the details of effects of sequence variants in noncoding regions, we first provide brief overviews of the various noncoding annotations and different kinds of sequence variants.

Noncoding annotations

The noncoding parts of the genome were once thought to be junk DNA but are now well known to contain many different types of regulatory elements that modulate expression of protein-coding genes. These elements are generally identified by sequence conservation or functional genomics approaches and often display cell- and tissue-type specificity (Figure 2). Several large-scale efforts such as ENCODE (Encyclopedia of DNA Elements) [1] and the NIH Roadmap Epigenomics Mapping Consortium[2] have been launched to create a comprehensive map of these regions. These efforts aim to provide genome-wide functional annotations across multiple cell- and tissue-types.

The various classes of noncoding annotations are identified using several functional genomics assays. For example, DNase I hypersensitivity for regions of open chromatin, ChIP-Seq for binding peaks of transcription factors (TFs) and histone marks, RNA-Seq for noncoding RNAs, etc. Evolutionary conservation of genomic sequence is also used to annotate noncoding regions[3, 4]. The dynamic annotation of these regions across various cellular states may be thought as turning gene regulation switches on and off using epigenetic marks. For example, as shown in the schematic in Figure 2, differential H3K27ac marks across various tissues indicate variable enhancer loci although the sequence at these loci where TFs bind stays the same. As a result, sequence variants in these loci are likely to exhibit tissue-specific effects on gene expression. This makes the functional interpretation of noncoding variants even more complex.

Linking the linear noncoding functional elements to their target protein-coding genes is of great importance and crucial to understand the effects of sequence variants in them. Multiple approaches are used to link cis-regulatory regions to their target genes. For example: different variations of chromosome conformation capture technology [5, 6], correlation of transcription factor (TF) binding and expression across multiple cell lines [7], etc. The resulting linkages can then be studied as a comprehensive regulatory network [8] (Figure 2).

We summarize the various sources of noncoding annotations with the web links for file downloads in Table 1.

Genomic sequence variants

DNA sequence variants range from single nucleotide variants (SNVs) to small insertions and deletions less than 50bp in length (indels) to larger structural variants (SVs). SVs comprise of deletions and duplications that lead to copy-number aberrations and inversions and translocations that are copy-number neutral. An average human genome contains roughly 3 million sequence variants relative to the reference human genome [9], while a tumor genome contains thousands of variants relative to the germline DNA (Figure 1)[10]. Unlike germline variants, somatic variants arise during mitotic cell divisions. Due to their different biological origins, somatic mutations tend to show distinct genomic patterns than germline variants. For example: (i) A higher fraction of somatic variants contain large genomic rearrangements. Recurrent fusion events between distant genes have been observed in many cancer types but are relatively rare in germline sequences (REF; confirm this is correct). Complex genomic rearrangements including chromoplexy[11] and chromothripsis[12] are known to occur in cancer cells. Chromosomal aneuploidy, where an entire chromosome may be lost or gained, is also often observed in cancer (REF). (ii) Somatic sequence variants may not be shared by all cells in the tumor tissue due to clonal evolution (REF). Such tumor heterogeneity makes interpretation of somatic variants more complex. (iii) Various phenomena, such as kataegis (localized hypermutation)[13] and other mutational signatures[10] are characteristic only of somatic variants.

Known cases of somatic variants playing a role in tumor development and growth

Somatic variants can effect gene expression in many different ways, e.g. point mutations in binding motifs of sequence-specific TFs may disrupt their binding, large deletions may delete entire TF binding sites/enhancer elements, etc (Figure 3). In this section, we discuss some known cases of somatic variants and their likely role in oncogenesis. We note that Vogelstein et al introduced the concept of Mut-driver and Epi-driver protein-coding genes, those that contain driver mutations and those that show aberrant expression providing selective growth advantage due to epigenetic changes, respectively[14]. Here we introduce an additional category, NcMut-driver genes, those that show aberrant expression providing selective growth advantage due to mutations in their noncoding regulatory regions. The examples discussed below correspond to such NcMut-driver genes. Different noncoding elements are effected by somatic changes --

a) Promoters: Recurrent mutations have been observed in the promoter of the *TERT* gene in many different cancer types[15-18]. These mutations create a binding motif for an *ETS* TF leading to its binding and subsequent up-regulation of *TERT* (Figure XX). Tumors in tissues with relatively low rates of self-renewal (including melanomas, urothelial carcinomas and medulloblastomas) tend to exhibit higher frequencies of *TERT* promoter mutations[17]. Germline mutations in this promoter are also observed and are related with familial melanoma[16]. The high occurrence of these mutations points to their role as drivers as opposed to passengers.
b) Enhancers: Enhancers constitute important cis-regulatory elements and play a major role in gene transcription. Super-enhancers are regions that recruit many TFs and drive expression of genes that define cell identity[19]. Recently, it was reported that somatic mutations create MYB binding motifs in T-cell acute lymphoblastic leukemia (T-ALL) which results in formation of a super-enhancer upstream of the *TAL1* oncogene resulting in its overexpression [20]. In another study, it was reported that somatic SVs juxtapose coding

sequences of *GFI1* or *GFI2* proximal to active enhancers (called 'enhancer-hijacking') in medulloblastoma [21] (Figure XX). In this case, although the SV effects the coding sequence, its functional impact occurs due to the activity of the enhancer region.

c) UTRs: Genomic lesions hitting UTRs are also known to be associated with cancer. The 5' UTR of *TMPRSS2* is frequently fused with ETS genes (*ERG* and *ETV1*) in prostate cancer [22]. This leads to *ERG* overexpression further disrupting androgen receptor (AR) signaling.

d) Other TF binding sites: Genomic rearrangements significantly associated with androgen receptor binding sites in a subset of prostate cancers [23, 24]. Basically this shows that AR binding drives the formation of structural rearrangements in some sub-types of cancer. [[To MG: I don't think this shows the functional role of noncoding sequence variants – infact noncoding sequence variants are the result of AR binding there. So I think we should exclude but need to discuss with Mark R.]]

e) Noncoding RNAs (ncRNAs) and their binding sites: Mis-regulation of ncRNAs is a cancer signature, and at least in some cases it could be due to the presence of somatic variants in them. For example, *MALAT1*, which is frequently up-regulated in cancer, was found to be significantly mutated in bladder cancer [25] and copy-number amplification of long ncRNA, lncUSMycN, is thought to contribute to neuroblastoma progression [26, 27]. In another scenario, pseudogene deletion can effect competition for miRNA binding with the parent gene, which in turn could effect expression of the parent gene. This is observed in certain cancers where *PTENP1* pseudogene is deleted, thereby leading to down-regulation of the parent *PTEN* tumor-suppressor gene [28] (Figure XX). Mutations in miRNA binding sites can also effect their binding, e.g. mutations in miR-31 binding site can lead to overexpression of AR in prostate cancer [29].

[[To MG: BELOW TEXT IS STILL IN BULLET POINTS]]

Germline variants in noncoding regions that alter cancer susceptibility or patient survival

a) There is an enrichment of GWAS variants, including those associated with cancer susceptibility, in the noncoding genome; as we sequence more populations we will identify variants that are common in those populations and related to cancer susceptibility. We will discuss the following examples and summarize them in Table S1:

(i) SNPs in enhancers on chr 8q24 upstream of *MYC* are related with increased risk for multiple cancer types [30].

(ii) A SNP in *RFX6* gene intron effects *HOXB13* binding and is linked to increased prostate cancer susceptibility [31].

(iii) A SNP in miR-27a gene reduces susceptibility to gastric cancer [32].

(iv) A common SNP in *TERT* promoter modifies the effects of somatic *TERT* promoter mutations in bladder cancer on patient survival [33]. Also the germline TERT promoter mutations observed in familial melanoma.

(v) Splice site mutation in the intron of *BRCA2* has implications for familial breast cancer [34].

(b) eQTL analysis has been used to interpret risk loci [35, 36]. We will also discuss why usually there is no eQTL analysis for somatic variants (since cancer is heterogeneous so these variants are rare). Cryptic effects of noncoding mutations have also been noted where germline variants exhibit allelic effects in tumor [37].

These examples illustrate how the effect of noncoding mutations and interplay between germline and somatic variants can be complex. We will discuss the relevance of two hit hypothesis (where one allele is disabled by a germline variant and the other by somatic variant) for noncoding regions. We will also use the above examples to discuss how the notion of driver mutations may not be binary since somatic mutations can influence cancer growth to varied extent based on the presence of other germline and somatic variants.

Francesca's CN state figure and text

Different types of cancer

(a) Discussion of total numbers of mutations and numbers of noncoding vs coding mutations for different cancers stratified by tissue-type, patient age, etc. For example, tumors of self-renewing tissues (such as colorectal) contain more mutations than non-self-renewing ones (such as glioblastomas and pancreatic cancers), pediatric tumors generally contain fewer mutations than adult ones, etc [14]. Since the numbers of noncoding vs coding mutations have not been computed comprehensively for different cancer types, we will do these calculations for published whole-genome sequences[10, 11, 13, 23] and summarize the results in Figure 1.

-- If we exclude pilocytic astrocytoma, there is a significant positive correlation between total number of mutations and noncoding fraction (rho = 0.32, p val = 2.195e-15).

(b) Summary of cancers where driver mutations have been identified in protein-coding genes vs those where causal mutations have not been identified. In cases where causal mutations have not been identified, the answer might lie in the noncoding genome since most previous studies have focused on canonical coding mutations.

Computational methods to identify noncoding somatic variants with functional consequences

(a) Discussion of currently available computational methods to predict noncoding driver mutations from whole-genome sequencing data, for example, FunSeq [4], CADD [38] and GWAVA [39]. We will also list these in Table 3 with associated website links.

Experimental approaches to understand the functional effects of noncoding mutations

Finally, we will discuss experimental ways to test which noncoding mutations have functional effects (e.g. genome editing using CRISPR, luciferase reporter assays, high-throughput assays,

etc). We will also discuss the scale and approximate cost of all the techniques and summarize them in Figure 4.

Conclusions/perspective

Recent studies have shown that small changes in gene expression caused by noncoding mutations can have large phenotypic impact (e.g. a SNP in enhancer causing 20% change in *KITLG* expression is responsible for blond hair color[40]). We postulate that the combined effect of small changes in expression due to noncoding mutations in cancer might be huge. Under this notion, genomic variants contribute to oncogenesis with varying probabilities, as opposed to the binary classification of mutations into drivers and passengers. While some somatic variants may have a direct role (such as *TERT* promoter mutations found in many different cancer types[17]), others may indirectly modulate important cancer pathways (such as genomic rearrangements perturbing androgen receptor binding sites in a subset of prostate cancers[23, 24]).

(a) Cancer arises because of accumulation of multiple driver mutations[14] -- some of these drivers could be noncoding. There is a bias in the literature for driver noncoding mutations because people haven't explored these regions to the same extent as coding genes, for example, the majority of TCGA studies have focused on exomes.

(b) There is a debate in the community about whether we should analyze whole-genomes vs exomes. Studies of somatic noncoding mutations are currently mostly for research purposes, as opposed to regular clinical use. This is primarily because current therapeutic approaches attempt to target proteins. It is possible that alternate methodologies, such as genome editing using CRISPR, may be used in future (e.g. CRISPR/Cas9 mediated editing has been used for HIV in cell lines [41] and muscular dystrophy in mice [42]). However, noncoding germline variants associated with increased cancer susceptibility should be important for risk assessment and potentially for preventive approaches.

(c) In relation to (b), it is very important to know the links between cis-regulatory regions and their target genes. Although many approaches exist (as discussed under 'Main sections'), this remains a very active and important area of research, especially the development of high-throughput choromosomal capture technologies.

(d) Even when the links between regulatory regions and target genes are known, it is important to study effects of mutations in all elements controlling gene expression – thus network approaches will be important to understand the role of noncoding mutations in cancer. We might also be able to identify new pathways or novel participants in known pathways that are important in cancer.

Glossary

Proposed display items

Figure 1: Numbers of total and noncoding vs coding mutations for different cancer types (Yao)
Figure 2: Noncoding annotations (Ekta)
Figure 3: Effect of sequence variants in noncoding regions in oncogenesis (Ekta)
Figure 4: Experimental approaches used to understand the functional effects of noncoding variants (Dimple)

Table 1: Noncoding annotations (include FANTOM)
Table 2: Computational methods to prioritize noncoding mutations with functional effects

Supplementary Figures
Figure from Francesca ?

Possible Glossary terms
Germline variants
Somatic variants
Cis-regulatory regions

Key references (100 maximum)

1. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
2. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317-24 (2012).
3. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321-5 (2004).
4. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
5. Hughes, J.R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* **46**, 205-12 (2014).
6. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189-91 (2012).
7. Yip, K.Y. et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48 (2012).
8. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
9. Consortium, G.P. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
10. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
11. Baca, S.C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-77 (2013).
12. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
13. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
14. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

15. Huang, F.W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).
16. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-61 (2013).
17. Killela, P.J. 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* **110**, 6021-6 (2013).
18. Heidenreich, B., Rachakonda, P.S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev* **24**, 30-7 (2014).
19. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
20. Mansour, M.R. et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* (2014).
21. Northcott, P.A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-34 (2014).
22. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).
23. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
24. Weischenfeldt, J. et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159-70 (2013).
25. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-9 (2013).
26. Liu, P.Y. et al. Effects of a novel long noncoding RNA, lncUSMycN, on N-Myc expression and neuroblastoma progression. *J Natl Cancer Inst* **106** (2014).
27. Buechner, J. & Einvik, C. N-myc and noncoding RNAs in neuroblastoma. *Mol Cancer Res* **10**, 1243-53 (2012).
28. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8 (2010).
29. Lin, P.C. et al. Epigenetic repression of miR-31 disrupts androgen receptor homeostasis and contributes to prostate cancer progression. *Cancer Res* **73**, 1232-44 (2013).
30. Grisanzio, C. & Freedman, M.L. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer* **1**, 555-9 (2010).
31. Huang, Q. et al. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat Genet* **46**, 126-35 (2014).
32. Yang, Q. et al. Genetic variations in miR-27a gene decrease mature miR-27a level and reduce gastric cancer susceptibility. *Oncogene* **33**, 193-202 (2014).
33. Rachakonda, P.S. et al. TERT promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc Natl Acad Sci U S A* **110**, 17426-31 (2013).
34. Bakker, J.L. et al. A Novel Splice Site Mutation in the Noncoding Region of BRCA2: Implications for Fanconi Anemia and Familial Breast Cancer Diagnostics. *Human Mutation* **35**, 442-446 (2014).
35. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
36. Xu, X. et al. Variants at IRX4 as prostate cancer expression quantitative trait loci. *Eur J Hum Genet* **22**, 558-63 (2014).
37. Ongen, H. et al. Putative cis-regulatory drivers in colorectal cancer. *Nature* (2014).
38. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

39. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).
40. Guenther, C.A., Tasic, B., Luo, L., Bedell, M.A. & Kingsley, D.M. A molecular basis for classic blond hair color in Europeans. *Nat Genet* **46**, 748-52 (2014).
41. Hu, W. et al. RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proc Natl Acad Sci U S A* (2014).
42. Long, C. et al. Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. *Science* (2014).