(5000 words maximum)

Title

Role of noncoding variants in cancer

Preface (100 words)

Tumor genomes contain numerous somatic 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. These parts play a role in genome structure organization and contain various regulatory elements (such as promoters, enhancers and noncoding RNAs) that modulate gene expression. Variability of epigenetic marks (like DNA methylation and histone modifications) across cellular states makes many regulatory elements cell-type specific. Thus, noncoding variants can have functional consequences causing tumor progression by effecting gene expression in a tissue-specific manner. 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[1]. As a result of the decreasing costs, whole-genomes of thousands of tumors have since been sequenced by TCGA (The Cancer Genome Atlas, tcga-data.nci.nih.gov) and ICGC (International Cancer Genome Consortium, icgc.org). 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 and ICGC, 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 complex-trait loci, including the ones associated with cancer susceptibility, lie in noncoding regulatory regions of the genome [2, 3]. Previous studies have found that protein-coding regions harboring germline variants linked with increased cancer risk also contain somatic driver events [REF]. Thus, noncoding regions with cancer GWAS variants may also contain somatic drivers. In this review,

Ekta Khurana 12/21/2014 7:04 PM
Deleted: Many regulatory elements exhibit cell-type specificity due to dynamic epigenetic marks, like DNA methylation and histone modifications.

Ekta Khurana 12/21/2014 7:04 PM
Deleted: to variable extent

Ekta Khurana 12/11/2014 5:01 PM
Formatted: Highlight

Ekta Khurana 12/21/2014 7:27 PM
Deleted: (The Cancer Genome Atlas)

Ekta Khurana 12/21/2014 7:27 PM
Deleted: (International Cancer Genome Consortium)

Ekta Khurana 12/22/2014 7:58 PM
Deleted:

Ekta Khurana 12/21/2014 8:21 PM
Deleted: we postulate that

we discuss the intricate relationship between germline polymorphisms and somatic variants that leads to tumorigenesis.

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 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) [4] and the NIH Roadmap Epigenomics Mapping Consortium[5] have been launched to create a comprehensive map of these regions. The GTEx project aims to provide an atlas of gene expression across multiple tissues [6]. Thus, these efforts aim to provide genome-wide functional annotations across multiple cell- and tissue-types.

The various classes of noncoding annotations can be 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. The raw signals from these experiments are processed using computational algorithms to yield functional annotation blocks [7]. In particular, TFs bind to specific DNA sequences within the larger peak regions identified using ChIP-Seq assays. DNase I fingerprinting can also help identify TF occupancy at nucleotide resolution within the larger DNase I hypersensitive sites [9, 10]. Variability in chromatin conformation and epigenetic marks across various cellular states leads to cell-type specific TF binding events. The dynamic annotation of noncoding regions across various cellular states may be thought of as turning gene regulation switches on and off using epigenetic marks. 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. Several histone modifications are associated with specific putative functions: H3K4me3 for promoters, H3K27ac for active promoters and enhancers, H3K27me3 for repressive regions, etc[11]. While most sequence-specific TFs and some chromatin marks lead to highly localized ChIP-Seq signals, other marks (such as H3K9me3 and H3K36me3) are associated with large genomic domains that can cover up to XX bp. Besides these cis-regulatory regions where TFs bind, the genome contains different types of noncoding RNAs that play a major role in gene regulation. These include tRNAs, rRNAs, snoRNAs, snRNAs, miRNAs, lncRNAs (>200bp), etc [12] . All these RNAs act via different mechanisms to modulate gene expression and many are well known to play an important role in cancer biology [13].

Ekta Khurana 12/22/2014 7:58 PM
Deleted: .

Ekta Khurana 12/22/2014 7:58 PM
Formatted: Indent: First line: 0"

Ekta Khurana 12/22/2014 7:58 PM
Deleted: .

Ekta Khurana 12/5/2014 12:10 AM
Deleted: were once thought to be junk DNA but are now well known to

Ekta Khurana 12/22/2014 12:20 AM
Deleted: T

Ekta Khurana 12/22/2014 7:58 PM
Deleted: . ... [1]

Ekta Khurana 12/22/2014 10:50 PM
Formatted: Indent: First line: 0.5"

Ekta Khurana 12/21/2014 11:09 PM
Deleted: TFs

Ekta Khurana 12/21/2014 10:53 PM
Deleted: in regions of open chromatin and can be divided into general TFs, chromatin remodelers and sequence-specific TFs [8]. [[TO MG: sometimes only sequence-specific TFs are called TFs: do we want to change how we define here?]] Sequence-specific TFs bind to specific DNA motifs within the

Ekta Khurana 12/22/2014 10:50 PM
Deleted:

Ekta Khurana 12/22/2014 7:56 PM
Deleted: generate

Transcriptome sequencing using RNA-Seq also yields functional insight into the genome. Besides revealing noncoding transcripts, gene expression studies help identification of eQTLs (expression quantitative trait loci) in noncoding regions, which in turn point to the putative functional role of the region [14]. Gene expression studies across various tissues reveal regulatory regions associated with tissue-specific expression [6].

Evolutionary conservation of genomic sequence across multiple species is also used to annotate noncoding regions [15, 16]. It is estimated that ~5% of the genome is more conserved between human and mouse than would be expected by neutral evolution [17]. Since only ~1.2% of the genome codes for proteins, the remaining ~3.8% conserved regions likely contain regulatory elements. Furthermore, 481 segments that are at least 200 bp long are 100% conserved between human, mouse and rat. These regions, termed ultra-conserved elements, cover ~107 kb of the genome and also exhibit high conservation among vertebrates [18]. Transcribed ultra-conserved regions exhibit aberrant expression in tumorigenesis and indeed can be used to differentiate cancer types [19, 20]. Hundreds of evolutionarily conserved regions (including ultra-conserved elements) have been tested for their *in vivo* activity as enhancers and are available from the VISTA database [21]. Besides selection constraint across multiple species, noncoding elements also exhibit conservation among humans further pointing to their functional roles [11, 22, 23].

Linking the noncoding functional elements to their target protein-coding genes in the three-dimensional (3D) chromatin structure 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 (3C) technology [24, 25], correlation of histone marks at enhancer regions and target gene expression across multiple cell lines [26], etc. The resulting linkages can then be studied as a comprehensive regulatory network [27] (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 4 million sequence variants relative to the reference human genome [28], while a tumor genome contains thousands of variants relative to the germline DNA (Figure 1) [29]. While the majority of ~4 million germline variants are SNPs; indels and SVs overall account for more nucleotide differences among humans as they cover larger segments of the genome [30]. The number of variants per individual also varies by ethnicity and individuals from different populations show varied profiles of rare and common variants [28]. Unlike germline variants, somatic variants arise during mitotic cell divisions. Due to their different biological origins, they do not share many properties of germline variants, such as linkage disequilibrium or association of alleles at multiple loci due to limited recombination between them. Instead, somatic mutations show other characteristic patterns. For example: (i) A higher fraction of somatic variants contain large genomic rearrangements. Recurrent fusion events between distant genes have been observed

Ekta Khurana 12/22/2014 7:58 PM
Deleted: .

Ekta Khurana 12/21/2014 11:24 PM
Deleted: linear

Ekta Khurana 12/22/2014 12:07 AM
Deleted: transcription factor (TF) binding

Ekta Khurana 12/22/2014 7:58 PM
Deleted: .

Ekta Khurana 12/22/2014 8:11 PM
Deleted: .

in many cancer types but are relatively rare in germline sequences (REF). Complex genomic rearrangements including chromoplexy[31] and chromothripsis[32] 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)[33] and other mutational signatures[29] are characteristic only of somatic variants. More than 20 mutational signatures have been identified in 30 different cancer types. Some signatures (such as the one associated with the APOBEC family of cytidine deaminases) are common across many different cancer types, while others (such as the one observed in malignant melanoma and linked with ultraviolet-light) are specific to cancer types [29].

[[EKtoMG: Could not find any reference for C & M types of cancer but we were already discussing this under 'Different types of cancer' below. Please see that section.]]

Ekta Khurana 12/23/2014 12:50 AM
Formatted: Highlight

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

Noncoding 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 and large deletions may delete entire TF binding sites/enhancer elements (Figure 3). In this section, we discuss some known cases of somatic variants and their likely role in oncogenesis. We note that very few studies have tried to explore the role of noncoding somatic variants in cancer development and only a handful of studies have tried this for large-scale analysis of many different cancer types [22, 34, 35]. Thus, we expect this list to grow as more whole cancer genomes are sequenced and analyzed. We are also likely to see new types of mutational effects, for example, most known point mutations related to oncogenesis lead to gain of TF motif and we expect to see examples of mutations leading to loss of motif. Vogelstein et al had previously 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[36]. 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 may be effected by somatic changes --

a) Gain of TF binding sites.

Recurrent mutations have been observed in the promoter of the *TERT* gene in many different cancer types[37-40]. These mutations create binding motifs for Ets/TCF 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[39]. The high occurrence of these mutations points to their role as drivers as opposed to passengers.

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[41]. 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 [42].

b) Fusion events due to genomic rearrangements.
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 [43]. This leads to *ERG* overexpression further disrupting androgen receptor (AR) signaling. Genomic rearrangements are also significantly associated with androgen receptor (AR) binding sites in a subset of prostate cancers, indicating that AR binding may drive the formation of structural rearrangements [44, 45].

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 [46] (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) 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 [47] and copy-number amplification of long ncRNA, lncUSMycN, is thought to contribute to neuroblastoma progression [48, 49]. 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 [50].

d) Role of pseudogenes in modulation of the expression of parent gene.
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 [51] (Figure XX).

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

Cancer is known to have a familial component and several loci associated with increased cancer risk have been identified by GWAS. Many of these loci lie in noncoding regions. Rare germline variants with high penetrance may be directly responsible for tumorigenesis (e.g. as observed in familial cancer cases) while common variants with low penetrance may modulate the effects of somatic variants. Several cases indicate that cancer results from a complex interplay of inherited germline and acquired somatic mutations. The case of 'two-hit' hypothesis demonstrates one such scenario where one allele is disrupted by a germline variant and the second one by a somatic mutation leading to oncogenesis. In another scenario, in hormone-regulated cancers (such as breast and prostate), the effects of altered hormonal generation during an individual's lifetime might be different depending on the germline genotypes of other genes and regulatory elements in the hormone-regulated pathway (Supplementary Figure XX) (personal communication with FD).

Unlike somatic variants, germline variants exhibit distinct characteristics like linkage disequilibrium and they occur in all tissues of the body. However, their functional effect might not be manifested in all tissues, e.g. if they occur in regions of closed chromatin or if they disrupt a binding site of a TF that is not expressed in the tissue, etc. We discuss a few examples of noncoding germline variants related to cancer susceptibility here.

a) Gain of TF binding site in *TERT* promoter.

Besides their somatic recurrence, germline mutations in *TERT* promoter are associated with familial melanoma[38]. Similar to the effect of somatic mutations, these mutations create binding motifs for Ets/TCF TFs. The functional effects of these mutations are more likely to be exhibited in the tissues where these TFs are expressed. Elevated expression of the TCF *ELK1* gene is observed in female specific tissues, such as ovary and placenta. Horn et al. reasoned that besides melanoma, this may be related to the increased ovarian cancer risk in women who are carriers of the mutation[38].

It was reported in another study that a common SNP (rs2853669) at another location in the *TERT* promoter modifies the effects of somatic *TERT* promoter mutations in bladder cancer on patient survival [52]. If the patients with somatic lesions in the *TERT* promoter carried this SNP, they showed better survival. From a mechanistic viewpoint, the common SNP might weaken the effect of somatic mutations since it disrupts a pre-existing Ets2 binding site.

The multiple germline and somatic variants in the *TERT* promoter demonstrate their complex relationship with cancer susceptibility, oncogenesis and patient survival.

b) SNPs in enhancers.

Multiple SNPs in a gene desert on chromosome 8q24 upstream of *MYC* are related with increased risk for many cancer types (breast, prostate, ovarian, colon and bladder cancers and chronic lymphocytic leukemia) [53]. Several observations, such as histone methylation and acetylation marks and 3C assays, suggest that these 8q24 SNPs occur in regions that act as enhancers for *MYC* in a tissue-specific manner. In another example, a prostate cancer risk associated SNP occurs in a cell-type specific enhancer and leads to increased *HOXB13* binding. This in turn upregulates *RFX6* and is linked to increased prostate cancer susceptibility [54].

(d) Noncoding RNAs (ncRNAs).

While most cancer associated polymorphisms are related to increased risk, some of them can also be beneficial and reduce susceptibility. A SNP in miR-27a impairs the processing of pre-mir-27a to its mature version. The reduced miR-27a level results in increased expression of its target *HOXA10*, which reduces susceptibility to gastric cancer [55].

(e) Intronic splice site mutations.

A rare mutation in the intron of *BRCA2* causes aberrant splicing and is related with Fanconi anemia (a rare recessive disorder involving high cancer risk) [56].

[[EKtoMG: You mentioned talk about germline variants and aging here. Not sure what we discussed but cancer occurs late mostly due to accumulation of somatic]]

Ekta Khurana 12/11/2014 3:50 PM
Deleted: :

Ekta Khurana 12/11/2014 3:50 PM
Deleted: :

Ekta Khurana 12/23/2014 12:50 AM
Formatted: Highlight

Ekta Khurana 12/22/2014 7:58 PM
Deleted:

We note that the examples above do not include an exhaustive list of all known cases of noncoding germline variants associated with altered cancer risk, but are meant to illustrate the diverse ways in which many regulatory polymorphisms exhibit their functional effects. Various other methods of identifying variants with potential functional consequences, such as expression quantitative trait loci (eQTL) and allele-specific expression analyses, have been used to interpret GWAS cancer loci [57-59]. Such studies reveal germline determinants of gene expression in tumors and help establish a link between noncoding risk loci and their target coding genes.

Different types of cancer

Somatic mutation frequency varies considerably across different cancer types [29, 60]. In general, slow growing tumors, such as carcinoid tumors and prostate cancer, harbor fewer mutations as compared to rapidly growing melanomas, bladder cancer and lung cancer. However, growth rate is not the only determinant and some rapidly growing tumors, such as acute myeloid leukemia (AML), Ewing sarcoma and neuroblastoma, are on the lower spectrum of somatic mutations. Many slow growing tumors have canonical oncogenic drivers that might obviate the need for a cancer to acquire mutations as a mechanism for selective advantage. Specifically, some of the tumors listed with the lowest mutations rate harbor defining genomic alterations and gene fusions: rhabdoid tumors harbor SMARCB1 deletions, Ewing sarcoma harbor a recurrent ETS gene fusion (EWS-FLI1), thyroid cancers harbor common RET mutations and fusions RET/PTC1, neuroblastomas harbor amplification of NMYC, and prostate cancers harbor common ETS gene fusions (most commonly TMPRSS2-ERG). We expect most mutations in tumors with high total numbers of mutations to be passenger events with no functional consequence. We also expect that a higher fraction of noncoding mutations would be passengers with little or no functional consequence as compared to coding mutations. In agreement with this hypothesis, we observe that the fraction of noncoding mutations is positively correlated with the total numbers of mutations across 11 (or 12) cancer types (Figure XX; Spearman correlation between total number of mutations and noncoding fraction=0.32, p val=2.20e-15).

Computational methods to identify noncoding somatic variants with functional consequences

A number of computational tools have been developed to annotate and prioritize potentially functional noncoding variants. A list of these tools with corresponding references is provided in Table 2. In general, these tools can be summarized into three categories.

1) Annotation tools, such as SeattleSeq, SNPnexus, ANNOVAR, VEP, GEMINI and OncoCis. These tools try to predict the functional impact of variants by annotating them with various genomic annotations.

2) Scoring methods, such as RegulomeDB, SInBaD, CADD, GWAVA and FunSeq. These tools go beyond annotation of variants and provide a score for each input variants for its likely deleterious effect. Most of these methods integrate multiple layers of knowledge including functional annotations and conservation.

Ekta Khurana 12/22/2014 7:59 PM
Formatted: Indent: First line: 0.5"

Ekta Khurana 12/11/2014 4:45 PM
Deleted: ... [2]

Ekta Khurana 12/11/2014 4:54 PM
Deleted: [To MG: I think we should include in Figure caption that this result is when we exclude pilocytic astrocytoma which shows a lot of variability in number of mutations and has been hypothesized to be a single pathway disease]

Ekta Khurana 12/11/2014 4:57 PM
Deleted: ... [3]

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

3) Tools designed specifically for common GWAS variants, such as FunciSNP, Haploreg, and GWAS3D. These tools also use functional annotations but go further to identify candidate regulatory SNPs that are in linkage disequilibrium with GWAS SNPs. They identify putative causal variants for complex disorders including cancer susceptibility.

Experimental approaches to understand the functional effects of noncoding mutations

[[EKtoMG: Dimple's new text still does not include cost & scale, need to discuss with her. Also need to make it more clear I think.]]

Noncoding mutations in regulatory elements in the genome can modulate the net transcriptional output resulting in a divergent transcriptome due to altered binding of TFs and other chromatin regulatory complexes. Several recent studies have explored methods to annotate and functionally assess these mutations. Experimental strategies to understand the effects of noncoding mutations on cellular functions are outlined in Figure XX. To capture the full spectrum of biological effects will require a combination of discovery and targeted approaches. In the discovery approach, unknown, noncoding mutations can be annotated and functionally validated using high-throughput sequencing and reporter assays. This can be achieved by optimal size specific shearing of total genomic DNA followed by ligation of synthetic adaptor DNA sequences to 5' and 3' ends of the sonicated DNA and cloning in transcription reporter constructs to generate promoter/enhancer libraries (PMID: 23328393). These cloned libraries are then transfected into eukaryotic cells and poly-A RNA produced from transcription competent constructs are isolated. Total poly-A RNA is reverse transcribed to obtain cDNA and further amplified using PCR utilizing reverse complimentary primers that hybridize to the adaptor sequences. This is followed by massively parallel paired-end sequencing of amplified DNA to generate sequencing longer reads with high read depth for comprehensive characterization of the widest range of structural variants. Paired end reads are mapped to reference genome and annotated using bioinformatics approach. Sequencing of various cancer datasets using this approach can not only provide a genome-wide annotation of non-coding mutations but can also predict if these mutations are associated with functional activity. This combined analysis and validation approach is useful to capture global changes in non-coding mutations and to decipher their role in tissue-specific evolution of cancer and cancer subtypes.

Noncoding mutations that modulate regulatory output from specific regions of genome can be expected to have a hierarchical organization. To understand the role of driver mutations and to rule out false positives derived from sequencing approach, a direct validation is imperative. This can be achieved by generating single or combinatorial mutations in genome using CRISPR-Cas9 system or by site-directed mutagenesis. Functional evaluation of the WT and mutants *in vivo* can provide relevance of mutations in the biological context. The second, targeted approach provides an opportunity for direct validation of known noncoding mutations using synthetic transcription reporter constructs that have regulatory sequences upstream of the reporter gene. Transfection based reporter gene assays can provide a direct *in vivo* assessment of the impact of non-coding mutations *in vitro* using various cell line model systems. To derive a

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

Ekta Khurana 12/23/2014 12:00 AM
Formatted: Normal, Space After: 12 pt, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Ekta Khurana 12/11/2014 4:59 PM
Formatted: Font:(Default) Arial, 11 pt

Ekta Khurana 12/11/2014 4:59 PM
Deleted:

Ekta Khurana 12/23/2014 12:00 AM
Formatted: Font:11 pt, Not Bold

Ekta Khurana 12/11/2014 4:59 PM
Deleted: (a) Discussion of currently available computational methods to predict noncoding driver mutations from whole-genome sequencing data, for example, FunSeq [22], CADD [61] and GWAVA [62]. We will also list these in Table 3 with associated website links.

Ekta Khurana 12/23/2014 12:00 AM
Formatted: Normal

Ekta Khurana 12/11/2014 5:04 PM
Deleted: 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.

Ekta Khurana 12/23/2014 12:31 AM
Formatted: Highlight

Ekta Khurana 12/23/2014 12:30 AM
Formatted: Highlight

Ekta Khurana 12/23/2014 12:42 AM
Formatted: Indent: First line: 0.5"

more direct *in vivo* relevance of noncoding mutations analyzed by both discovery and targeted approach will require biologic validation in cell lines, and in mouse and zebra fish models.

Conclusions

Cancer arises because of accumulation of multiple driver mutations[36] -- some of these drivers could be noncoding. This could be particularly true for certain cancer types, such as non-small cell lung cancer where coding drivers have not been identified in major subpopulations [63]. Currently, 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. Furthermore, 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[64]). Thus, 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[39]), others may indirectly modulate important cancer pathways (such as genomic rearrangements perturbing androgen receptor binding sites in a subset of prostate cancers[44, 45]). The various cases discussed in this article show that the effects of somatic mutations on tumorigenesis depend on the existing germline variants and their binary classification into drivers and passengers does not capture this complexity.

Currently, 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 [65] and muscular dystrophy in mice [66]). However, identification of noncoding germline variants associated with increased cancer susceptibility is also very important for risk assessment and potentially for preventive approaches.

To interpret the functional effects of regulatory variants, it is very important to know the links between cis-regulatory regions and their target genes. Although many approaches exist (as discussed in this article), this remains a very active and important area of research, especially the development of high-throughput chromosomal capture technologies. Indeed, there is dedicated NIH funding available for innovative approaches to study the 3D structure of the genome and its spatiotemporal dynamics. We note that even when the links between regulatory regions and target genes are known, it will be important to study effects of mutations in all elements controlling gene expression in a comprehensive fashion. 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

Possible Glossary terms

Ekta Khurana 12/23/2014 12:27 AM
Formatted: Font color: Custom Color(RGB(26,26,26))

Ekta Khurana 12/23/2014 12:52 AM
Deleted: /perspective

Ekta Khurana 12/23/2014 1:01 AM
Formatted: Highlight

Ekta Khurana 12/23/2014 1:04 AM
Deleted: R

Ekta Khurana 12/23/2014 12:44 AM
Deleted: We postulate that

Ekta Khurana 12/11/2014 4:28 PM
Formatted: Highlight

Ekta Khurana 12/23/2014 1:05 AM
Deleted:

Ekta Khurana 12/23/2014 1:05 AM
Deleted:
... [4]

Ekta Khurana 12/23/2014 1:07 AM
Deleted: should be

Ekta Khurana 12/23/2014 1:07 AM
Deleted:

Ekta Khurana 12/23/2014 1:07 AM
Deleted:
... [5]

Ekta Khurana 12/23/2014 1:08 AM
Deleted: under 'Main sections'

Ekta Khurana 12/23/2014 1:10 AM
Deleted: o

Ekta Khurana 12/23/2014 1:13 AM
Deleted:
... [6]

Ekta Khurana 12/23/2014 1:13 AM
Deleted: E

Ekta Khurana 12/23/2014 1:13 AM
Deleted: is

Ekta Khurana 12/23/2014 1:13 AM
Deleted: –

Ekta Khurana 12/23/2014 1:13 AM
Deleted: t

Germline variants
Somatic variants
Cis-regulatory regions

Proposed display items

Figure 1: Numbers of total and noncoding vs coding mutations for different cancer types (Yao). (Can also show coverage of different noncoding elements.)
Note this correlation is when we exclude pilocytic astrocytoma which shows a lot of variability in number of mutations and has been hypothesized to be a single pathway disease.
Figure 2: Noncoding annotations (Ekta)
Panel XX: 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.
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

Key references (100 maximum)

1. Ley, T.J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72 (2008).
2. Maurano, M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
3. Chen, C.Y., Chang, I.S., Hsiung, C.A. & Wasserman, W.W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med Genomics* **7**, 34 (2014).
4. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
5. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317-24 (2012).
6. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
7. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M.B. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**, 559-71 (2010).
8. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-63 (2009).
9. Galas, D.J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**, 3157-70 (1978).
10. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).

11. Consortium, E.P. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
12. Morris, K.V. & Mattick, J.S. The rise of regulatory RNA. *Nat Rev Genet* **15**, 423-37 (2014).
13. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov* **1**, 391-407 (2011).
14. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
15. Loots, G.G. et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-40 (2000).
16. Pennacchio, L.A. & Rubin, E.M. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**, 100-9 (2001).
17. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
18. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321-5 (2004).
19. Peng, J.C., Shen, J. & Ran, Z.H. Transcribed ultraconserved region in human cancers. *RNA Biol* **10**, 1771-7 (2013).
20. Calin, G.A. et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**, 215-29 (2007).
21. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).
22. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
23. Katzman, S. et al. Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
24. 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).
25. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189-91 (2012).
26. 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).
27. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
28. Consortium, G.P. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
29. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
30. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
31. Baca, S.C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-77 (2013).
32. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
33. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
34. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-5 (2014).

35. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**, 1258-63 (2014).
36. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
37. Huang, F.W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).
38. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-61 (2013).
39. 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).
40. Heidenreich, B., Rachakonda, P.S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev* **24**, 30-7 (2014).
41. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
42. Mansour, M.R. et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* (2014).
43. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).
44. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
45. 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).
46. Northcott, P.A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-34 (2014).
47. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-9 (2013).
48. 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).
49. Buechner, J. & Einvik, C. N-myc and noncoding RNAs in neuroblastoma. *Mol Cancer Res* **10**, 1243-53 (2012).
50. 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).
51. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8 (2010).
52. 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).
53. Grisanzio, C. & Freedman, M.L. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer* **1**, 555-9 (2010).
54. 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).
55. 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).
56. 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).
57. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
58. Xu, X. et al. Variants at IRX4 as prostate cancer expression quantitative trait loci. *Eur J Hum Genet* **22**, 558-63 (2014).

59. Ongen, H. et al. Putative cis-regulatory drivers in colorectal cancer. *Nature* (2014).
60. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).
61. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
62. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).
63. Network, C.G.A.R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-50 (2014).
64. 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).
65. 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).
66. Long, C. et al. Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. *Science* (2014).