**Title**

 Role of noncoding sequence variants in cancer

**Authors**

Ekta Khurana1,2,3,4, #,Yao Fu 5, Dimple Chakravarty 2,7, Francesca Demichelis 2,3,8, Mark A. Rubin 1,2,7, #, Mark Gerstein 5,6,9, #

1 Meyer Cancer Center, Weill Cornell Medical College, New York, NY

2 Institute for Precision Medicine, Weill Cornell Medical College, New York, NY

3 Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY

4 Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY

5 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT

6 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT

7 Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY

8 Centre for Integrative Biology, University of Trento, Trento, Italy

9 Department of Computer Science, Yale University, New Haven, CT

#Corresponding authors: ekk2003@med.cornell.edu, rubinma@med.cornell.edu, pi@gersteinlab.org

**Preface**

Cancer patients carry somatic sequence variants in their tumor in addition to the germline variants in their inherited genome. Most somatic and germline variants occur in noncoding portions of the genome. Most common germline variants linked with cancer susceptibility and identified through genome-wide association studies show small effect sizes while rare variants with large effect sizes have been identified in familial cancer cases. On the extreme are driver somatic events with direct consequences on tumor growth and progression. Furthermore, the range of variants can vary from single nucleotide mutations to those affecting wide regions, e.g. genomic rearrangements. Functional effects of noncoding variants can be interpreted using annotations of regulatory regions, e.g. transcription-factor binding sites and noncoding RNAs. Variability of epigenetic marks across cellular states makes many of these regulatory elements tissue-specific. In this review, we provide a number of case studies of germline and somatic variants in noncoding regions associated with cancer. These variants often manifest themselves through change of expression levels of cancer-associated genes. We also show that early studies suggest that the prevalence of noncoding variants is different in various cancer types with some types, such as lung cancer, having proportionately more noncoding mutations.

**Introduction**

The first tumor whole-genome was sequenced in 20081. The decreasing costs of sequencing have led to whole-genome sequencing of thousands of tumors by individual research groups and efforts such as 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. 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 the 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 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. Thus, noncoding regions with cancer GWAS variants may also contain somatic drivers. In this review, 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 (ncRNAs) are known to be dysregulated in various cancers 4, 5, H3K4me1 sites can be lost or gained in cancer cells relative to matched normal 6, etc. However, in this review, we focus on effects of DNA sequence variants in noncoding regions. 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) 7 and the NIH Roadmap Epigenomics Consortium 8, 9 were launched to create a comprehensive map of these regions. The GTEx project aims to provide an atlas of gene expression across multiple tissues and many individuals 10, and this can be analyzed to determine potential regulatory regions. 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 and RNA-Seq for ncRNAs. The raw signals from these experiments are processed using computational algorithms to yield functional annotation blocks 11. 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 12, 13. 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; examples include, H3K4me3 for promoters, H3K27ac for active promoters and enhancers and H3K27me3 for repressive regions 14. While most sequence-specific TFs and some chromatin marks lead to highly localized ChIP-Seq signals (hundreds of nucleotides), other marks (such as H3K9me3 and H3K36me3) are associated with large genomic domains that can cover up to a few megabases. Besides these cis-regulatory regions where TFs bind, the noncoding genome contains different types of ncRNAs that play a major role in gene regulation. These include tRNAs, rRNAs, snoRNAs, snRNAs, miRNAs, lncRNAs (>200bp), etc 15 . All these RNAs act via different mechanisms to modulate gene expression and many are well known to play an important role in cancer biology 4. Pseudogenes may be considered a category of ncRNAs that act via different modes. They can generate endo-siRNAs and play a role in gene regulation via the RNA interference pathway 16, 17. They can also act as molecular sponges and compete with parent gene mRNA for miRNA binding 4.

Transcriptome sequencing using RNA-Seq yields functional insights into the genome. Besides revealing noncoding transcripts, correlation of expression with the occurrence of sequence variants helps in the identification of eQTLs (expression quantitative trait loci) in noncoding regions, which in turn point to the putative functional role of the region 18. Moreover, gene expression studies across various tissues can reveal regulatory regions associated with tissue-specific expression 10.

 Evolutionary conservation of genomic sequence across multiple species is also used to annotate noncoding regions 19, 20. It is estimated that ~5% of the genome is more conserved between human and mouse than would be expected by neutral evolution 21. Since only ~1.2% of the genome codes for proteins, the remaining ~3.8% conserved regions likely contain regulatory elements and ncRNAs. 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 22. 370 of these 481 ultra-conserved elements do not overlap protein-coding exons. Transcribed noncoding, ultra-conserved regions exhibit aberrant expression in tumorigenesis and indeed can be used to differentiate cancer types 23, 24. 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 25. Besides selection constraint across multiple species, noncoding elements also exhibit conservation among humans. Negative selection among humans can be estimated using various metrics, such as enrichment of rare alleles, and further points to the functional role of these elements 14, 26, 27.

 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 28, 29 and correlation of histone marks at enhancer regions and target gene expression across multiple cell lines 30.The resulting linkages can then be studied as a comprehensive network 31 (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, also called genomic rearrangements, can be copy-number aberrant (such as deletions and duplications) or copy-number neutral (such as inversions and translocations). An average human genome contains roughly 4 million sequence variants relative to the reference human genome 32, while a tumor genome contains thousands of variants relative to the same individual’s germline DNA (Figure 1) 33. The number of germline variants per individual also differs by ethnicity and individuals from different populations show varied profiles of rare and common variants 32. Germline and somatic variants show many distinct features. (i) The majority of ~4 million germline variants are SNPs, although indels and SVs overall account for more nucleotide differences among humans as they cover larger segments of the genome 34. In contrast, a higher fraction of somatic variants consists of large genomic rearrangements. Recurrent fusion events between distant genes have been observed in many cancer types but are relatively rare in germline sequences. Complex genomic rearrangements including chromoplexy35 and chromothripsis36 are known to occur in cancer cells. Chromosomal aneuploidy, where an entire chromosome may be lost or gained, is also often observed in cancer 37. (ii) Unlike germline variants, somatic variants arise during mitotic cell division. 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. (iii) Somatic sequence variants may not be shared by all cells in the tumor tissue due to clonal evolution. Such tumor heterogeneity makes interpretation of somatic variants more complex. (iv) Various phenomena, such as *kataegis* (localized hypermutation)38 and other mutational signatures33 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 particular cancer types 33. We discuss the patterns of somatic variants in different cancer types in more detail below.

 Dysregulation of gene expression is a prominent cancer signature. Logsdon et al reported a method to identify driver genes based only on expression profiles 39. Genes can be dysregulated in tumors due to various reasons. 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 in the tumor, respectively40. 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.

**Germline variants in noncoding regions altering cancer susceptibility or survival**

Cancer is known to have a familial component and several loci associated with increased cancer risk have been identified by GWAS. Most of these lie in noncoding regions. Rare germline variants with high penetrance may be directly responsible for tumorigenesis (e.g. as observed in familial cancer cases 41), while variants with low penetrance may modulate the effects of somatic variants 42. With the exception of pediatric cancers, most cancer cases occur at an older age. Thus, the germline variants associated with increased cancer susceptibility for non-pediatric cancers do not have a fitness effect at reproductive age, perhaps the reason for the prevalence of such variants in the population.

Unlike somatic variants, germline variants 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. Furthermore, noncoding 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). We discuss a few examples of noncoding germline variants related to cancer susceptibility below.

1. Promotor Mutations: Gain of TF binding site in *TERT* promoter.

Germline mutations in the *TERT* promoter are associated with familial melanoma 41. These mutations create binding motifs for Ets TFs and ternary complex factors (TCFs) (Figure 3B). 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 41.

b) SNPs in enhancers.

Multiple SNPs in a gene desert on chromosome 8q24 upstream of *MYC* are related to increased risk for many cancer types (breast, prostate, ovarian, colon and bladder cancers and chronic lymphocytic leukemia) 43. 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 44.

Another example illustrates that in hormone-regulated cancers (such as prostate, breast, ovary and endometrial), germline polymorphisms in enhancers can alter the strength of binding of nuclear TFs (such as androgen receptor, AR or estrogen receptor, ER). This can affect the expression of target tumor suppressor genes and contribute to carcinogenesis 45 (Figure 3B).

(c) SNPs in 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* and reduced susceptibility to gastric cancer 46.

(d) Variants in introns.

Variants in introns can affect splice sites and also cause loss of repressor elements. For instance, 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) 47. Also, germline copy number variants spanning intronic inhibitor regulatory elements can lead to the overexpression of target transcripts potentially modulating cell proliferation or migration. The loss of an intronic regulatory element in the α-1,3-mannosyl-glycoprotein 4-β-N acetylglucosaminyltransferase C (*MGAT4C*) gene was found to be associated with increased risk of developing aggressive prostate cancer in a population-based study 48.

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. 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 49-51. Such studies reveal germline determinants of gene expression in tumors and help establish a link between noncoding risk loci and their target coding genes.

**Somatic variants in different types of cancer**

Somatic mutation frequency varies considerably across different cancer types 33, 52. In general, slow growing tumors, such as carcinoid tumors and prostate cancer, harbor fewer mutations as compared to rapidly growing melanomas and bladder 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-end of spectrum in terms of the number of somatic mutations. Specifically, some of the tumors listed with the lowest mutation 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 rearrangements (most commonly TMPRSS2-ERG). In fact, most tumors can be divided into two classes: M class that harbor mostly point mutations and C class with mostly copy-number alterations 53. However, *TP53* mutations are an exception to this trend and are strongly enriched in the C class, consistent with early *TP53* mutations causing genomic instability for copy-number aberrations.

 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 eleven cancer types (Figure 1; Spearman correlation between total number of mutations and noncoding fraction=0.32, p val=2.20e-15).

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

In this section, we discuss some known cases of somatic variants and their likely role in oncogenesis. We note that although many studies have explored the link between noncoding germline variants and cancer, 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 across many different cancer types 26, 54, 55. Based on the prevalence of noncoding germline variants associated with cancer susceptibility, we expect the list of noncoding somatic variants related to tumorigenesis will grow as more whole cancer genomes are sequenced. 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. Different noncoding elements may be affected by somatic changes.

1. Gain of TF binding sites.

Recurrent mutations have been observed in the promoter of the *TERT* gene in many different cancer types41, 56-58. These mutations create binding motifs for the Ets TFs and TCFs leading to their binding and subsequent up-regulation of *TERT* (Figure 3B). Tumors in tissues with relatively low rates of self-renewal (including melanomas, urothelial carcinomas and medulloblastomas) tend to exhibit higher frequencies of *TERT* promoter mutations57. 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 identity59. Recently, it was reported that somatic mutations create MYB binding motifs in T-cell acute lymphoblastic leukemia (T-ALL) forming a super-enhancer upstream of the *TAL1* oncogene resulting in its overexpression 60.

1. Fusion events due to genomic rearrangements.

Genomic lesions affecting UTRs are also known to be associated with cancer. The 5’ UTR of *TMPRSS2* is frequently fused with Ets genes (e.g., *ERG* and *ETV1*) in prostate cancer 61. This leads to *ERG* overexpression further disrupting AR signaling. Genomic rearrangements are also significantly associated with AR binding sites in a subset of prostate cancers, indicating that AR binding may drive the formation of structural rearrangements 62, 63.

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 64 (Figure 3C). In this case, even if the SV effects the coding sequence, its functional impact occurs due to the activity of the enhancer region.

1. ncRNAs and their binding sites.

Dysregulation 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 65 and copy-number amplification of a long ncRNA, lncUSMycN, is thought to contribute to neuroblastoma progression 66, 67. 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 68 (Figure 3D).

d) Role of pseudogenes in modulation of the expression of parent gene.

Transcribed pseudogenes are a particular type of ncRNA that bears a clear resemble to a functioning protein-coding gene. Due of this resemblance, transcribed pseudogenes are thought to have a natural way to affect and regulate their parent gene. In particular, 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 69 (Figure 3E).

**Interplay between germline and somatic variants**

Several cases discussed in this review indicate that cancer results from a complex interplay of inherited germline and acquired somatic mutations. Knudson’s ‘two-hit’ hypothesis is widely known, where one allele is disrupted by a germline variant and the second through somatic mutation leading to oncogenesis. In a contrasting scenario, a common SNP (rs2853669) in *TERT* promoter weakens the effects of somatic *TERT* promoter mutations. This SNP modifies the effects of somatic *TERT* promoter mutations in bladder cancer on patient survival 70. 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. Thus, the multiple germline and somatic variants in the *TERT* promoter particularly demonstrate the complex relationship of regulatory variants with cancer susceptibility, oncogenesis and patient survival.

**Computational methods to identify noncoding 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. The various features of these tools are also provided in the Table. Most of these tools can interpret both SNVs and indels, while some tools (e.g. ANNOVAR, VEP and GEMINI) also analyze SVs. Many tools first annotate variants with various functional annotations and evolutionary conservation. Some tools are designed specifically for common GWAS variants (e.g. FunciSNP, Haploreg and GWAS3D) and try to identify candidate regulatory SNPs that are in linkage disequilibrium with GWAS SNPs. Thus, they identify putative causal variants for complex disorders including cancer susceptibility. Some tools also use a scoring scheme to provide a score for each variant (e.g. RegulomeDB, CADD, FunSeq and FitCons). Most of the methods that score variants integrate multiple layers of functional and conservation knowledge. Additionally, some methods (such as FunSeq) analyze recurrence of somatic variants from tumor samples in functional elements, similar to the burden-tests strategy used for association of rare germline variants with complex traits 71. We note that methods that try to identify driver noncoding elements (i.e. elements undergoing positive selection in tumor) need to account for genomic mutation rate covariates (such as chromatin accessibility and replication timing), similar to the driver analyses for coding genes 52, 54, 55, 72.

 **Experimental approaches to understand the functional effects of noncoding mutations**

Several recent studies have explored methods to annotate and functionally assess noncoding mutations. Experimental approaches to understand the effects of noncoding mutations on cellular functions are outlined in Figure 4, which shows the main elements of the strategies: (A) creating sequence variants, (B) high- and low-throughput functional assays to understand their transcriptional effects, and (C) direct biological validation. Specifically, mutations can first be introduced in DNA using site-directed mutagenesis or CRISPR-Cas9 system 73 (Figure 4A). Oligos containing the mutations may also be synthesized directly for high-throughput screening. Then, the functional effects of noncoding mutations can be probed through massively parallel high-throughput assays and/or low- to medium-throughput luciferase reporter assays (Figure 4B). High-throughput assays involve ligation of synthetic adaptor DNA sequences to 5’ and 3’ ends of the wild type or mutant DNA and cloning in transcription reporter constructs to generate promoter/enhancer libraries 74. These cloned libraries are then transfected into eukaryotic cells and poly-A RNA produced from transcription competent constructs is isolated. Total poly-A RNA is reverse transcribed to obtain cDNA and further amplified using PCR utilizing reverse complementary primers that hybridize to the adaptor sequences. This is followed by massively parallel sequencing of amplified DNA and subsequent mapping to the genome. 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. Reporter assays using synthetic transcription reporter constructs that have regulatory sequences upstream of the reporter gene provide an opportunity for direct validation of known noncoding mutations.

To understand the biological role of driver mutations and to rule out false positives derived from sequencing approaches, a direct validation of oncogenic properties of the mutations is imperative (Figure 4C). Functional evaluation of the WT and mutants *in vitro* (using various cell line model systems) and *in vivo* (in model organisms, such as zebrafish and mouse) can provide relevance of mutations in the biological context.

Functional validation of potential noncoding variants is extremely important to understand their biological consequence. High throughput analysis of variants significantly reduces the cost per variant (Figure 4D). However, among all the current methods for functional validation of variants, the cost of biologic validation is the highest. Hence, prioritization of putative functional mutations is critical prior to the establishment of these *in vivo* systems given considerations of lengthy developmental time (years) and costs.

 **Conclusions**

Cancer arises because of accumulation of multiple driver mutations40 -- some of these drivers can be noncoding. This can be particularly the case for certain cancer types, such as non-small cell lung cancer where coding drivers have not been identified in major subpopulations 75. Currently, there is a bias in the literature against driver mutations in noncoding regions because researchers have not 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 76). Thus, the combined effect of small changes in expression due to noncoding mutations in cancer might be more significant than currently appreciated. Thus, genomic variants could contribute to oncogenesis with varying probabilities, as opposed to the binary classification of mutations into drivers and passengers. The effects of somatic variants also depend on the existing genetic background, for example the presence of risk alleles in inherited germline DNA. While some somatic variants may have a direct role (such as *TERT* promoter mutations found in many different cancer types57), others may indirectly modulate important cancer pathways. The various cases discussed in this review 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 reserved for research purposes and have not been incorporated into precision-medicine cancer care approaches. 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. Although the use of CRISPR/Cas9 system for targeted editing of tumor DNA has not been explored, CRISPR has shown promising *in vivo* results, e.g. for prevention of muscular dystrophy in mice 77 and to generate a mouse model of lung cancer with a specific chromosomal rearrangement 78. However, identification of noncoding germline variants associated with increased cancer susceptibility is also very important for risk assessment and potentially for preventive approaches.

 Moreover, to interpret the functional effects of regulatory variants, it is important to know the links between cis-regulatory regions and their target genes. Although many approaches exist (as discussed in this review), this remains an active and important area of research, especially the development of high-throughput chromosomal capture technologies. 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**

Germline variants: Heritable variants that are transmitted to offspring. These variants are constitutional, i.e. present in all cells of the body.

Somatic variants: Variants that are not inherited from a parent and are not transmitted to offspring.

Cis-regulatory regions: Regions of DNA that regulate gene expression via TF binding. These include enhancers and promoters.

**Figure captions**

**Figure 1.** Somatic mutations in various cancer types. Bar plot denotes the average number of SNVs. Box plot shows the fraction of noncoding variants (based on Gencode 19). As shown in the pie charts, noncoding variants are further classified into different categories according to ENCODE annotations (mean values are reported). Variants are assigned to these categories with the following order: 'ncRNA', 'Pseudogene' > 'DHS' > 'Histone' > 'Unannotated'.  'AML' - acute myeloid leukemia; 'MB' - medulloblastoma; 'DLBC' - B cell lymphoma; 'STAD' - gastric cancer; 'BRCA' - breast cancer; 'PAAD' - pancreatic cancer; 'PRAD' - prostate cancer; 'LIHC' - liver cancer; 'PA' -pilocytic astrocytoma; 'CLL' - chronic lymphocytic leukemia**;**'LUAD' - lung adenocarcinoma. 'DHS' - Dnase1 hypersensitive site; 'Histone' - histone modification peaks. Spearman correlation between total number of mutations and noncoding fraction=0.32, p val=2.20e-15. 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.** Identification of regulatory elements using functional genomics assays and evolutionary conservation. Even though the sequence motif is same, regulatory elements can vary across tissues due to variability in regions of open chromatin (DHS) or histone marks (ChIP-seq) in turn leading to variability in TF binding (ChIP-Seq). Some elements may not show activity in limited functional genomics experiments and are identified by evolutionary conservation only. The elements can be connected to target coding genes, which can then be compiled into networks.

**Figure 3**. Effect of sequence variants in noncoding regions in oncogenesis. (A) Overview of the noncoding elements that can be effected. Specific cases are shown in (B) to (E). (B) Mutations can lead to loss- or gain- of TF binding motifs. The effects of a SNP that reduces nuclear receptor (NR) binding affinity to DNA are observed at lower NR levels as a result of reduced hormone levels. (C) SVs juxtaposes proto-oncogene (GFI1/GFI1B) next to regulatory element (super enhancer). Deletions, tandem duplications, inversions, translocations or other complex SVs can juxtapose the gene next to enhancer leading to its transcription. Either enhancer or gene can overlap SVs. (D) Mutations in miRNA binding sites prevent miRNA binding leading to increased target gene expression. (E) PTEN pseudogene loss. Pseudogene deletion leads to more miRNAs binding to the parent gene further leading mRNA silencing through its degradation or translational repression.

**Figure 4.** Methods for functional validation of noncoding variants. (A) Mutations in cloned DNA fragments can be generated using site-directed mutagenesis or by the CRISPR-CAS system. Additionally synthetic oligos with WT or mutant sequence can be chemically synthesized. (B) Functional output of the noncoding mutations can be determined either using a single or combinatorial approach involving high-throughput sequencing and/or luciferase (LUC) reporter assays. In the former method DNA fragments are cloned in expression polyA tagged constructs to generate promoter/enhancer libraries. RNA transcripts from these transcribed libraries are used for cDNA synthesis and further amplified using PCR, followed by massively parallel paired-end sequencing of amplified DNA. For the LUC reporter assays, DNA fragments cloned into the reporter vectors are transfected in cells followed by measuring the reporter activity. (C) Oncogenic properties, such as cell proliferation, migration and invasion can be tested *in vitro* using cell lines and tumorigenesis can also be tested *in vivo* using model organisms. (D) The cost of functional validation per mutation changes with the techniques used and is the highest when *in vitro* and *in vivo* biologic validation studies are included. Cost/variant for functional validation from 10 up to 100 variants is computed using a combination of site directed mutagenesis (SDM) and reporter luciferase assays. However, for functional validation of 1000 variants and above, cost per variant is optimized with oligo library synthesis with and without the mutation, cloning, transfection into cells, RNA extraction and high-throughput sequencing and reporter assays. The dotted line includes the cost for biological (*in vitro* and *in vivo* tumorigenic assays) validation of 10 variants.

**Table 1: Noncoding annotations.**

Weblinks: GENCODE (<gencodegenes.org>), FANTOM (<fantom.gsc.riken.jp>), ENCODE (<encodeproject.org>), Roadmap Epigenomics (roadmapepigenomics.org). DHS, DNase I hypersensitivity.

|  |  |
| --- | --- |
| **Annotation** | **Resource**  |
| Transcription start sites |  GENCODE, FANTOM  |
| Transcription factor binding sites and motifs | ENCODE, Roadmap Epigenomics, JASPAR (jasper.genereg.net), Transfac (biobase-international.com/products), CIS-BP (cisbp.ccbr.utoronto.ca) |
| DHS sites (regions of open chromatin) | ENCODE, Roadmap Epigenomics |
| Histone marks | ENCODE, Roadmap Epigenomics |
| Integrated chromatin states (including enhancers) | ENCODE & Roadmap Epigenomics (derived from methods such as ChromHMM and Segway), FANTOM |
| Enhancer-Promoter linkages | ENCODE, Roadmap Epigenomics, FunSeq2 (funseq2.gersteinlab.org) |
| TF-Target gene linkages  | ENCODE (Derived from ChIP-Seq: encodenets.gersteinlab.org and DHS: regulatorynetworks.org), Roadmap Epigenomics |
| Topologically associated domains from HiC | ENCODE  |
| Various types of ncRNAs | GENCODE, additional miRNAs at [mirbase.org](http://www.mirbase.org), snoRNAs at www-snorna.biotoul.fr, tRNAs at gtrnadb.ucsc.edu and lncRNAs at mitranscriptome.org |

**Table 2: Computational methods to prioritize noncoding variants with functional effects**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tool** | **Variant type** | **Functional annotation** | **Conservation** | **LD calculation** | **Somatic mutation recurrence** | **Scoring scheme** | **Weblink** |
| SeattleSeq | SNV, Indel | Y | Y  | N | N  | N  | snp.gs.washington.edu/SeattleSeqAnnotation138 |
| SNPnexus | SNV, Indel | Y  | Y  | N  | N  | N  | snp-nexus.org (79, 80) |
| ANNOVAR | SNV, Indel, SV | Y  | Y  | N  | N  | N  | openbioinformatics.org/annovar/ (81) |
| VEP | SNV, Indel, SV | Y  | N  | N  | N  | N  | ensembl.org/info/docs/tools/vep/ (82) |
| OncoCis | SNV, Indel | Y  | Y  | N  | N  | N  | powcs.med.unsw.edu.au/OncoCis/ (83) |
| GEMINI | SNV, Indel, SV | Y  | Y  | N  | N  | N  | github.com/arq5x/Gemini (84) |
| FunciSNP | SNP | Y  | N  | Y  | N  | N  | bioconductor.org (85) |
| HaploReg | SNP, Indel | Y  | Y  | Y  | N  | N  | compbio.mit.edu/HaploReg (86) |
| GWAS3D | SNP | Y  | Y  | Y  | N  | Y  | jjwanglab.org/gwas3d (87) |
| is-rSNP | SNV | N  | N  | N  | N  | Y  | genomics.csse.unimelb.edu.au/is-rSNP (88) |
| RegulomeDB | SNV | Y  | N  | N  | N  | Y  | RegulomeDB.org (89) |
| SInBaD | SNV | N  | Y  | N  | N  | Y  | tingchenlab.cmb.usc.edu/Sinbad (90) |
| CADD | SNV, Indel | Y  | Y  | N  | N  | Y  | cadd.gs.washington.edu (91) |
| FunSeq | SNV, Indel | Y  | Y  | N  | Y  | Y  | funseq2.gersteinlab.org (26, 92) |
| GWAVA | SNV, Indel | Y  | Y  | N  | N  | Y  | sanger.ac.uk/resources/software/gwava/ (93) |
| FitCons | SNV | Y  | Y  | N  | N  | Y  | (94) |

**References**

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. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov* **1**, 391-407 (2011).

5. Iyer, M.K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* (2015).

6. Akhtar-Zaidi, B. et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736-9 (2012).

7. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

8. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317-24 (2012).

9. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

10. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).

11. 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).

12. 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).

13. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).

14. Consortium, E.P. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

15. Morris, K.V. & Mattick, J.S. The rise of regulatory RNA. *Nat Rev Genet* **15**, 423-37 (2014).

16. Sasidharan, R. & Gerstein, M. Genomics: Protein fossils live on as RNA. *Nature* **453**, 729-731 (2008).

17. Tam, O.H. et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-8 (2008).

18. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).

19. 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).

20. Pennacchio, L.A. & Rubin, E.M. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**, 100-9 (2001).

21. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

22. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321-5 (2004).

23. Peng, J.C., Shen, J. & Ran, Z.H. Transcribed ultraconserved region in human cancers. *RNA Biol* **10**, 1771-7 (2013).

24. Calin, G.A. et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**, 215-29 (2007).

25. 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).

26. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).

27. Katzman, S. et al. Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).

28. 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).

29. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189-91 (2012).

30. 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).

31. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).

32. Consortium, G.P. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

33. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).

34. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).

35. Baca, S.C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-77 (2013).

36. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).

37. Holland, A.J. & Cleveland, D.W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat Rev Mol Cell Biol* **10**, 478-87 (2009).

38. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).

39. Logsdon, B.A. et al. Sparse expression bases in cancer reveal tumor drivers. *Nucleic Acids Res* (2015).

40. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

41. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-61 (2013).

42. Easton, D.F. & Eeles, R.A. Genome-wide association studies in cancer. *Hum Mol Genet* **17**, R109-15 (2008).

43. Grisanzio, C. & Freedman, M.L. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer* **1**, 555-9 (2010).

44. 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).

45. Garritano, S. et al. In-silico identification and functional validation of allele-dependent AR enhancers. *Oncotarget* (2015).

46. 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).

47. 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).

48. Demichelis, F. et al. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proc Natl Acad Sci U S A* **109**, 6686-91 (2012).

49. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).

50. Xu, X. et al. Variants at IRX4 as prostate cancer expression quantitative trait loci. *Eur J Hum Genet* **22**, 558-63 (2014).

51. Ongen, H. et al. Putative cis-regulatory drivers in colorectal cancer. *Nature* (2014).

52. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).

53. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133 (2013).

54. 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).

55. 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).

56. Huang, F.W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).

57. 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).

58. Heidenreich, B., Rachakonda, P.S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev* **24**, 30-7 (2014).

59. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).

60. Mansour, M.R. et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* (2014).

61. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).

62. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).

63. 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).

64. Northcott, P.A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-34 (2014).

65. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-9 (2013).

66. 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).

67. Buechner, J. & Einvik, C. N-myc and noncoding RNAs in neuroblastoma. *Mol Cancer Res* **10**, 1243-53 (2012).

68. 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).

69. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8 (2010).

70. 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).

71. Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).

72. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-4 (2015).

73. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* (2014).

74. Arnold, C.D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-7 (2013).

75. Network, C.G.A.R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-50 (2014).

76. 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).

77. Long, C. et al. Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. *Science* (2014).

78. Maddalo, D. et al. In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516**, 423-7 (2014).

79. Chelala, C., Khan, A. & Lemoine, N.R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655-61 (2009).

80. Dayem Ullah, A.Z., Lemoine, N.R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res* **40**, W65-70 (2012).

81. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

82. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

83. Perera, D. et al. OncoCis: annotation of cis- regulatory mutations in cancer. *Genome Biol* **15**, 485 (2014).

84. Paila, U., Chapman, B.A., Kirchner, R. & Quinlan, A.R. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput Biol* **9**, e1003153 (2013).

85. Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A. & Noushmehr, H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res* **40**, e139 (2012).

86. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).

87. Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C. & Wang, J. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res* **41**, W150-8 (2013).

88. Macintyre, G., Bailey, J., Haviv, I. & Kowalczyk, A. is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics* **26**, i524-30 (2010).

89. Boyle, A.P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).

90. Lehmann, K.V. & Chen, T. Exploring functional variant discovery in non-coding regions with SInBaD. *Nucleic Acids Res* **41**, e7 (2013).

91. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

92. Fu, Y. et al. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).

93. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).

94. Gulko, B., Hubisz, M.J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* (2015).