Title

Role of noncoding sequence variants in cancer

Darren Burgess 4/16/2015 4:31 PM

Comment [1]: 4555 words, 0B, 4F, 2T, 94 refs

Authors

Ekta Khurana[1,2,3,4, #],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

[To MG: Did not work on abstract yet]

[Au: your abstract is about twice as long as we can accommodate (please shorten and aim for as close to 100 words as possible). The goal of the abstract is to entice the reader and give them a broad overview of what your review covers, rather than to explicitly list every take-home message. Below your abstract, I have adapted text from your abstract to form a shortened version – please re-edit if you prefer different emphasis, but hopefully you can see the conciseness that I am aiming for. If any really key points are removed due to space, you could always make sure they are clearly mentioned elsewhere in the relevant parts of the main text.] 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.

[Au: an example alternative version, shortened to fit:

Cancer patients carry somatic sequence variants in their tumor in addition to the germline variants in their inherited genome. Although variants in protein-coding regions have received the majority of attention, most somatic and germline variants occur in noncoding portions of the genome. We review our latest understanding of non-coding variants in cancer, including the great diversity in the mutation types — from single nucleotide variants to large genomic rearrangements — and in the wide range of mechanisms by which they disrupt gene expression to promote tumorigenesis, such as affecting transcription factor binding sites or functions of non-coding RNAs. We highlight distinctions between somatic and germline variants, and how they can be interpreted through computational and experimental tools.**]**

Ekta Khurana 5/24/2015 5:18 PM
Formatted: Highlight

Ekta Khurana 5/24/2015 5:18 PM
Formatted: Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Font:Bold, Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Font:Bold, Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Font:Bold, Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Font:Bold, Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Font:Bold, Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Highlight

Ekta Khurana 5/22/2015 7:07 PM
Formatted: Font:Bold, Highlight

Darren Burgess 4/16/2015 3:21 PM
Formatted: Font:Bold

Introduction

Exome sequencing of tens of thousands of tumor samples has revealed the landscape of somatic mutations in protein-coding genes [[REF.]]. Most previous studies of cancer genomes used exome rather than whole-genome sequencing due to lower costs and focus on regions considered to be most functionally relevant. However, the decreasing costs of sequencing have led to whole-genome sequencing (WGS) 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). One of the most important benefits of whole-genome sequencing is identification of variants in noncoding regions of the genome. Indeed, most of the variants obtained from WGS of tumor genomes lie in noncoding regions (Figure 1). 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 analyze noncoding variants in ~2500 tumor and matched normal whole-genomes. One of the biggest challenges of analyzing noncoding variants is to identify drivers from passengers, similar to the challenge of analyzing coding variants.

The link between inherited germline variants and complex disorders has been probed previously by numerous genome-wide association studies (GWAS) using DNA from non-disease cells (usually blood). These studies have revealed that most loci associated with complex traits, including those 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 are also sites of somatic driver events [[REF.]]; if such a relationship holds for noncoding regions, those harboring cancer-associated germline variants identified by GWAS may also contain somatic drivers. Indeed, the list of known noncoding, germline cancer-risk variants might be able to guide the discovery of novel somatic drivers, and vice versa.

In this review, we discuss our current understanding of the role of noncoding sequence variants in cancer development and growth. We first describe distinctions in the nature of somatic versus germline sequence variants and then provide brief overviews of the various noncoding annotations. We then discuss diverse molecular mechanisms by which somatic and germline variants are known to lead to tumorigenesis, including their functional interplay. Finally, to interpret the wealth of non-coding variants that are being linked to cancer, we describe how bioinformatics and experimental approaches can be used to prioritize and validate the functional relevance of the variants. Throughout our review we focus on effects of DNA sequence variants in noncoding regions. However, we acknowledge that besides sequence alterations, other changes can occur in noncoding regions of cancer genomes, such as epigenetic changes at regulatory elements [4] and transcriptional dysregulation of noncoding RNAs (ncRNAs) [5, 6]; for further information on these phenomena, the reader is referred to [[ADD REF.]]....

NOTE

Genomic sequence variants

Deleted: [Au: as a general comm ... [1]

Ekta Khurana 5/22/2015 7:07 PM
Formatted ... [2]

Ekta Khurana 5/23/2015 3:44 PM
Deleted: [Au: I think that this ... [3]

Ekta Khurana 5/22/2015 7:09 PM
Formatted ... [4]

Ekta Khurana 5/23/2015 3:31 PM
Formatted ... [5]

Ekta Khurana 5/17/2015 12:32 AM
Deleted: The first tumor whole-gen ... [6]

Ekta Khurana 5/17/2015 12:36 AM
Deleted: M

Darren Burgess 4/14/2015 4:24 PM
Formatted ... [7]

Darren Burgess 4/14/2015 10:19 AM
Moved down [1]: In this review we ... [8]

Ekta Khurana 5/17/2015 12:37 AM
Deleted: We note that most previou ... [9]

Darren Burgess 4/14/2015 10:03 AM
Formatted ... [10]

Ekta Khurana 5/17/2015 12:37 AM
Deleted: t…ere is an increased ... [11]

Ekta Khurana 5/23/2015 3:35 PM
Formatted ... [12]

Ekta Khurana 5/21/2015 5:22 PM
Deleted: Genetic susceptibility for

Darren Burgess 4/14/2015 11:07 AM
Formatted ... [13]

Ekta Khurana 5/23/2015 3:45 PM
Deleted: [Au: to make the conce ... [14]

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [15]

Darren Burgess 4/14/2015 11:01 AM
Deleted: -…raits loci… including t ... [16]

Darren Burgess 4/14/2015 10:01 AM
Formatted ... [17]

Darren Burgess 4/14/2015 11:44 AM
Deleted: also contain

Ekta Khurana 5/23/2015 3:45 PM
Formatted ... [18]

Darren Burgess 4/14/2015 12:37 PM
Deleted: .

Ekta Khurana 5/23/2015 3:45 PM
Deleted: [Au:OK? Presumably n ... [19]

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [20]

Darren Burgess 4/14/2015 12:34 PM
Deleted: Thus, …oncoding region ... [21]

Ekta Khurana 5/21/2015 6:00 PM
Deleted: s…may also contain som ... [22]

Darren Burgess 4/16/2015 4:12 PM
Formatted ... [23]

Ekta Khurana 5/21/2015 6:25 PM
Formatted ... [24]

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [28]

Darren Burgess 4/14/2015 10:19 AM
Moved (insertion) [1] ... [29]

Ekta Khurana 5/23/2015 3:50 PM
... [25]

Darren Burgess 4/14/2015 1:07 PM

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [26]

Ekta Khurana 5/23/2015 3:50 PM
Formatted ... [27]

Ekta Khurana 5/21/2015 6:32 PM
... [30]

Darren Burgess 4/16/2015 3:46 PM
Formatted ... [31]

Darren Burgess 4/14/2015 12:43 PM

1st

We discuss general properties of DNA sequence variants since most of them occur in noncoding regions. They 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 germline 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]. Most studies of somatic variants have focused on the ones in tumor tissues, since they are relatively rare in normal tissues [[REF]]. Hence, in this article, we refer to somatic variants as the ones specific to tumor cells. Somatic mutation frequency varies considerably across different cancer types [33, 34]. 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). This could be due to the higher number of passengers in tumors with high total numbers of mutations, and most noncoding mutations corresponding to passenger events.

TYPICALLY

[Au: if you include this, please mention its relevance. If the excess of noncoding mutations in tumours with high mutation loads is just a consequence of functionally irrelevant passenger mutations, it's not clear why this correlation is useful.
Also note that the correlation is across 10, not 11 tumour types according to the figure 1 legend (the outlier PA is excluded from the analysis). It's also not clear why these particular 10 tumour types are chosen for the correlation analysis, given the availability of mutation data from a much wider range of tissue types. Is this analysis attempting to highlight tissue types for which noncoding mutations deserve special attention? If so this issue is partially explored in your conclusions section ("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"**), so perhaps a more consolidated discussion there could also work?]**
[To MG: Based on the first sentence of Editor's comment in the highlighted part above, perhaps better to remove correlation from text and just keep in figure legend. Since we are saying that excess of noncoding mutations in tumors with high mutation loads is a consequence of functionally irrelevant mutations, sort of also going against the theme of article that important to look at noncoding mutations. Also don't think fits in conclusion so if mention that would be here.]

A discussion of germline variants is important since cancer is known to have a familial component and several noncoding variants are known to play a role in cancer development. Rare, noncoding germline variants with high penetrance may be directly responsible for tumorigenesis (e.g. as observed in familial cancer cases [35]), while variants with low penetrance may modulate the effects of somatic variants [36]. [To MG: Do not know of other high penetrance noncoding examples, need to check]. With the exception of pediatric cancers, most cancer cases occur at an older age. Thus, the germline variants associated with increased cancer

2HIT

Ekta Khurana 5/21/2015 7:09 PM
Deleted: DNA

Ekta Khurana 5/21/2015 7:09 PM
Deleted: sequence variants

Ekta Khurana 5/21/2015 6:55 PM
Deleted: , while

Ekta Khurana 5/22/2015 7:10 PM
Formatted: Highlight

Ekta Khurana 5/24/2015 2:34 PM
Formatted: Indent: First line: 0.5"

Darren Burgess 5/24/2015 1:43 PM
Comment [2]: A single reference of TERT-associated melanoma seems a bit restrictive here. Are there any other examples of noncoding components to other familial cancers, e.g. BRCA1/2, TP53 (Li Fraumeni), APC, VHL etc?

Ekta Khurana 5/24/2015 2:47 PM
Formatted: Highlight

REL-TO THE REF

susceptibility for non-pediatric cancers do not typically have a fitness effect at reproductive age, which is perhaps the reason for the prevalence of such variants in the population.

The number of germline variants per individual 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. First, the majority of ~4 million germline variants are single-nucleotide polymorphisms (SNPs), even though indels and SVs overall account for more nucleotide differences among humans as they cover larger segments of the genome [37]. In contrast, a higher fraction of somatic variants consists of large genomic rearrangements. (REF) Recurrent fusion events between distant genes have been observed in many cancer types but are relatively rare in germline sequences. Complex genomic rearrangements including chromoplexy[38] and chromothripsis[39] are known to occur in cancer cells. Chromosomal aneuploidy, where an entire chromosome may be lost or gained, is also often observed in cancer [40]. Second, various phenomena, such as *kataegis* (localized hypermutation) [41] and other mutational signatures [33] 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. Third, unlike germline variants, somatic variants are not inherited. Thus, they are not subject to the recombinatorial effects of meiosis and hence do not show linkage disequilibrium or association of alleles at multiple loci. Fourth, somatic sequence variants may not be shared by all cells in the tumor tissue. Such tumor heterogeneity makes interpretation of somatic variants more complex.

[Au: this paragraph seems to be quite choppy and missing sufficient context/explanation. It also seems out of context due to its focus on expression disruption in a section that is otherwise strictly about sequence variants. This text would be best if moved elsewhere in your article; for example, If you are proposing new nomenclature, it could fit well in your conclusions section. Note also that I wonder whether the NcMut-driver classification you propose is already covered by the other classifications. Assuming that Mut-driver specifically refers to coding mutations, then would NcMut-driver just be a subset of the Epi-driver genes for which an underlying sequence variant was known to contribute to the aberrant expression?]

[[Editor: We agree that such classification might be confusing for readers and we have removed this part]]

Noncoding annotations

In order to understand the effect of sequence variants in noncoding regions, we need to first understand the role of various noncoding functional elements. We discuss these elements and the approaches used to annotate them in the genome. Noncoding elements can play diverse roles in regulation of protein-coding genes. Broadly speaking, they consist of cis-regulatory regions where TFs bind and noncoding RNAs. These elements are generally identified by

Deleted: To Mark

Ekta Khurana 5/24/2015 1:37 PM
Formatted ... [39]

Ekta Khurana 5/21/2015 6:57 PM
Deleted: a tumor genome contain ... [40]

Ekta Khurana 5/23/2015 3:52 PM
Deleted: **Second is that your** ... [42]

Ekta Khurana 5/23/2015 3:52 PM
Formatted ... [41]

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [43]

Ekta Khurana 5/24/2015 2:38 PM
Deleted: also

Darren Burgess 4/15/2015 11:17 AM
Deleted:

Ekta Khurana 5/23/2015 3:52 PM
Deleted: …**[Au: unfortunately, j** ... [44]

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [45]

Darren Burgess 4/15/2015 11:17 AM
Deleted: (i) T

Ekta Khurana 5/21/2015 7:17 PM
Deleted: although

Darren Burgess 4/15/2015 11:39 AM
Formatted ... [46]

Darren Burgess 4/15/2015 11:59 AM
Moved (insertion) [2] ... [49]

Ekta Khurana 5/23/2015 3:52 PM
Deleted: **[Au: I have moved the** ... [47]

Darren Burgess 4/15/2015 11:59 AM
Deleted: (iv) V

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [48]

Unknown
Field Code Changed ... [50]

Ekta Khurana 5/22/2015 7:07 PM
Formatted ... [51]

Darren Burgess 4/15/2015 11:35 AM
Deleted: (ii) U

Ekta Khurana 5/21/2015 7:29 PM
Deleted: arise during mitotic cell ... [52]

Darren Burgess 4/15/2015 11:47 AM
Formatted ... [53]

Ekta Khurana 5/23/2015 3:53 PM
Deleted: **[Au: to my eyes, this s** ... [54]

Ekta Khurana 5/22/2015 7:10 PM
Formatted ... [55]

Darren Burgess 4/15/2015 11:47 AM
Formatted ... [56]

Darren Burgess 4/15/2015 11:40 AM
Deleted: (iii)…sS ... [57]

Ekta Khurana 5/22/2015 3:59 PM
Deleted: due to clonal evolution

Darren Burgess 4/15/2015 11:59 AM
Moved up [2]: . (iv) Various phen ... [58]

Darren Burgess 4/15/2015 11:48 AM
Formatted ... [59]

Darren Burgess 4/15/2015 11:59 AM
Deleted: …[41] and other mutationa ... [60]

Ekta Khurana 5/24/2015 1:38 PM
Formatted ... [61]

Ekta Khurana 5/24/2015 11:44 AM
Formatted ... [62]

Ekta Khurana 5/22/2015 4:14 PM
... [63]

Ekta Khurana 5/24/2015 11:44 AM
Formatted ... [64]

Ekta Khurana 5/23/2015 3:55 PM
... [65]

Ekta Khurana 5/22/2015 7:11 PM
Formatted ... [66]

Darren Burgess 4/15/2015 4:44 PM

functional genomics approaches or sequence conservation and often display cell- and tissue-type specificity (Figure 2).

Cis-regulatory regions include promoters and distal elements (enhancers, silencers and insulators) [[REF]]. These are the regions where TFs bind and regulate gene expression. TF binding sites can be identified using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) assays. TFs bind to specific DNA sequences (motifs) within their larger regions of occupancy (peaks) identified by ChIP-Seq. They bind DNA in regions of open (non-nucleosomal) chromatin. These regions can be identified using DNase I hypersensitivity assays and DNase I footprinting can also help identify high-resolution TF binding sites within the larger DNase I hypersensitive sites (DHSs) [12, 13]. Furthermore, DNA methylation and histone modifications can modulate TF accessibility to DNA. Indeed, several histone marks are associated with specific putative functions, for example: H3K4me3 with promoters, H3K27ac with active promoters and enhancers and H3K27me3 with repressive regions [14]. Sites of histone modifications can also be identified using ChIP-Seq assays. Due to their inherent properties, 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. Thus, overall, epigenetic changes can alter TF accessibility in different cellular states and act as gene regulation switches resulting in cell-type specific TF binding events. Additionally, distal regulatory elements regulate gene expression by interacting with promoters in the three-dimensional (3D) structure of the genome. Linking the distal elements to their target protein-coding genes in the 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). Several large-scale efforts such as ENCODE (Encyclopedia of DNA Elements) [8] and the NIH Roadmap Epigenomics Consortium [9, 10] were launched to create a comprehensive map of these regions.

ELT ON & OFF

EQTL

Transcriptome sequencing using RNA-Seq yields functional insights into the genome. Correlation of gene 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]. Gene expression studies across various tissues can reveal regulatory regions associated with tissue-specific expression [7]. The Genotype–Tissue Expression (GTEx) project has provided an atlas of gene expression across multiple tissues and many individuals [7], which can be used to identify potential regulatory regions. [[UPDATE REF]]

RNA-Seq also reveals noncoding transcripts, which can be further confirmed to not code for proteins by the absence of open reading frame or proteomic analysis. Certain histone modifications can also indicate noncoding RNA activity, such as H3K4me3 associated with promoters and H3K36me3 associated with actively transcribed regions. ncRNAs can be divided into several categories: tRNAs, rRNAs, snoRNAs, snRNAs, miRNAs, lncRNAs (>200bp), etc [15]. All these RNAs act via different mechanisms to modulate gene expression and many are well

REWORD

Ekta Khurana 5/22/2015 8:40 PM
Deleted: [Au:OK? To match the order of the discussions that follow]

Ekta Khurana 5/22/2015 7:11 PM
Formatted: Highlight

Ekta Khurana 5/22/2015 8:40 PM
Deleted: [Au: I think that it is clearer to describe the assays and concepts first, and then mentioning relevant large consortia that can put all this together, so I suggest moving the ENCODE, Roadmap and GTEx mentions instead below to the relevant studies/assays, so that it is clearer what particular approaches / data types they use/integrate. For example, you mention the GTEx approach later in this section, "gene expression studies across various tissues can reveal regulatory regions associated with tissue-specific expression [7]**" (without actually mentioning its name there) so mentioning the project by name there would be more appropriate than here.]** Several large-scale efforts such as ENCODE (Encyclopedia of DNA Elements) [8] and the NIH Roadmap Epigenomics Consortium [9, 10] were launched to create a comprehensive map of these regions. The Genotype–Tissue Expression (GTEx) project aims to provide an atlas of gene expression across multiple tissues and many individuals [7], 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.

Ekta Khurana 5/22/2015 7:11 PM
Formatted ... [71]

Ekta Khurana 5/22/2015 8:41 PM
Deleted:

Ekta Khurana 5/22/2015 7:55 PM
Formatted ... [72]

Ekta Khurana 5/23/2015 4:12 PM
Formatted ... [73]

Ekta Khurana 5/23/2015 1:53 PM
Deleted: The various classes of noncoding annotations can be identified using several functional genomics assays. For example, DNase I hypersensitivity for regions of open (non-nucleosomal) chromatin, chromatin immunoprecipitation followed by sequencing (ChIP-seq) for binding peaks of transcription factors (TFs) and histone marks, and high-throug ... [74]

Ekta Khurana 5/22/2015 7:18 PM
Formatted ... [75]

known to play an important role in cancer biology [5]. ¶ In particular, transcribed pseudogenes are a particular type of ncRNA that bear a clear resemblance to a functioning protein-coding gene. They do not code for proteins due to disabling mutations but can regulate the expression of their parent genes, for example, by generating endo-siRNAs and participating in RNA interference pathway [16, 17] or by acting as molecular sponges competing with parent gene mRNA for miRNA binding [5].

TOO COMPACT

Evolutionary conservation of genomic sequence across multiple species is also used to annotate noncoding regions [19, 20]. Comparative analysis of human with mouse, rat and dog genomes showed that at least ~5% of the genome is conserved [21, 44, 45]. Since only ~1.5% of the genome codes for proteins, the remaining ~3.5% conserved regions **[Au: this assumes that the entire protein-coding genome is conserved. Is that the case?] [Editor:Yes]** 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. Analysis of the sequence variants in these noncoding, ultra-conserved elements is important since they have been shown to play a role in cancer biology. Some noncoding, ultra-conserved elements are transcribed and act as ncRNAs that exhibit aberrant expression in tumorigenesis and indeed can be used to differentiate cancer types [23, 24]. ¶ 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]. Furthermore, functional activity of evolutionary conserved regions can be tested using various assays. For example, 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].

Ekta Khurana 5/23/2015 4:19 PM
Formatted: Indent: First line: 0"

Ekta Khurana 5/23/2015 7:17 PM
Formatted: Highlight

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

[Au: for this section (highlighted in grey) I'm struggling to work out how this fits conceptually into your article. I suggest removal, unless it can be reworked to more clearly emphasize its relevance. Even then, it would probably be best as a box, because in its current position it is breaking the flow of the sections above and below it, which should ideally be juxtaposed so that the (often parallel) biological consequences of somatic versus germline mutations can be explored in adjacent sections]

[Editor: We have moved a couple of sentences to section 'Genomic sequence variants' and removed the rest of the section.]

Ekta Khurana 5/24/2015 1:29 PM
Formatted: Highlight

Ekta Khurana 5/24/2015 1:30 PM
Formatted: Highlight

Roles for somatic variants in cancer

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, 46-48]. 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.

[Au: similar to my comments on the equivalent discussions of germline variants, I like these examples, but it would be useful to include a bit more information on the affected oncogenes/TSGs, any explanations for the observed tissue specificity, and citations of any additional examples from each type of mutation]

Ekta Khurana 5/24/2015 2:48 PM
Formatted: Highlight

Gain of TF binding sites. TRANS

Telomerase reverse transcriptase (*TERT*) is the catalytic subunit of the enzyme telomerase. Telomerase lengthens telomeres allowing cells to escape apoptosis and become cancerous. *TERT* is generally repressed in normal somatic cells and its overexpression has been observed in cancer, thereby making it an oncogene [REF.]. In the last few years, numerous studies have reported recurrent mutations in the promoter of the *TERT* gene in many different cancer types [35, 49-51]. 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 mutations [50]. The high occurrence of these mutations points to their role as drivers as opposed to passengers.

Ekta Khurana 5/24/2015 4:59 PM
Formatted: Font:Italic

Ekta Khurana 5/24/2015 3:35 PM
Formatted: Font:Italic

Ekta Khurana 5/24/2015 3:38 PM
Formatted: Highlight

Ekta Khurana 5/24/2015 3:33 PM
Formatted: Highlight

Gain of TF binding sites has also been observed for enhancers, which constitute important distal cis-regulatory elements and play a major role in gene transcription. In particular, super-enhancers are regions that recruit many TFs and drive expression of genes that define cell identity [52]. 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* (T-cell acute lymphocytic leukemia 1) gene resulting in its overexpression [53]. *TAL1* is an oncogene that codes for a basic helix-loop-helix TF, which plays an important role in erythroid differentiation and is implicated in hemopoietic malignancies [REF.]

Ekta Khurana 5/24/2015 3:51 PM
Formatted: Font:Italic

Ekta Khurana 5/24/2015 4:57 PM
Formatted: Font:11 pt

Ekta Khurana 5/24/2015 4:57 PM
Formatted: Font:11 pt, Highlight

Fusion events due to genomic rearrangements.

Multiple examples of fusion events placing active regulatory elements next to oncogenes are known. For example, the 5' UTR of *TMPRSS2* is frequently fused with Ets genes (e.g., *ERG* and *ETV1*) in prostate cancer [54]. This leads to *ERG* overexpression further disrupting androgen receptor (AR) signaling. [Need one more sentence may be from Mark R.] Genomic

Ekta Khurana 5/24/2015 5:03 PM
Formatted: Highlight

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 [55, 56].

In another example, it was reported that somatic SVs juxtapose coding sequences of *GFI1* or *GFI2* proximal to active enhancers (called ‘enhancer-hijacking’) in medulloblastoma [57] (Figure 3C). In these cases, even if the SV affects the coding sequence, its functional impact occurs due to the activity of the enhancer region. Similarly, in T-ALL, *TAL1* coding sequence is fused with the promoter of ubiquitously expressed *SIL* (SCL-interrupting locus) gene, leading to overexpression of *TAL1* [58]. This rearrangement is found in 25% of cases of human T-ALL.

[[To MG: READ UNTIL THIS POINT]]

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 upregulated in cancer, was found to be significantly mutated in bladder cancer [59] and copy-number amplification of a long ncRNA, lncUSMycN, is thought to contribute to neuroblastoma progression [60, 61]. Mutations in miRNA binding sites can also affect their binding, e.g. mutations in miR-31 binding site **[Au: literally in the AR mRNA, or is the effect mediated through other genes?]** can lead to overexpression of AR in prostate cancer [62] (Figure 3D).

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

Transcribed pseudogenes are a particular type of ncRNA that bears a clear resemblance 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 affect competition for miRNA binding with the parent gene, which in turn could affect expression of the parent gene. This is observed in certain cancers where *PTENP1* pseudogene is deleted, thereby leading to downregulation of the parent *PTEN* tumor-suppressor gene [63] (Figure 3E).

Roles for germline variants in cancer

[Au: somewhere in your article (either here, or in the section above where you compare germline versus somatic sequence variants) it would be useful to highlight whether the detection methods have an influence on the identification or interpretation of germline variants. In particular, my understanding is that most cancer GWASs have been typically performed using SNP-sensitive microarrays, whereas it is not clear from your article how the germline variants have been probed. If our understanding of how germline variants differ from somatic variants might be skewed by different methodologies (e.g. microarrays for germline variants versus sequencing for somatic variants) this would be worthy of mention, particularly as you highlight in your article numerous differences between germline and somatic variants, so it would be important to know that they were biologically relevant rather than just a consequence of the methodology. Presumably bias can be minimized if WGS is performed on both the tumour and patient-matched

Ekta Khurana 5/24/2015 5:01 PM
Formatted: Indent: First line: 0.5"

Ekta Khurana 5/24/2015 5:06 PM
Formatted: Font:Italic

Ekta Khurana 5/24/2015 5:11 PM
Formatted: Font:Italic

Ekta Khurana 5/24/2015 5:14 PM
Formatted: Font:Italic

Ekta Khurana 5/24/2015 5:17 PM
Formatted: Highlight

Darren Burgess 5/23/2015 3:57 PM
Comment [3]: Could this also be relevant to germline variants, e.g. Targeted resequencing of the microRNAome and 3'UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer.
Chen X, Paranjape T, Stahlhut C, McVeigh T, Keane F, Nallur S, Miller N, Kerin M, Deng Y, Yao X, Zhao H, Weidhaas JB, Slack FJ.
Oncogene. 2014 Jun 9. doi: 10.1038/onc.2014.117

Darren Burgess 4/16/2015 4:31 PM
Comment [4]: OK? Note that primary headings (i.e. the bold titles in your article) need to be short to fit on 1 line in print. If you prefer an alternative, please keep it equally short

Darren Burgess 4/15/2015 4:49 PM
Deleted: G

Darren Burgess 4/15/2015 4:48 PM
Deleted: in noncoding regions altering

Darren Burgess 4/15/2015 4:49 PM
Deleted: susceptibility or survival

Darren Burgess 4/14/2015 11:42 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 4:31 PM
Comment [5]: e.g. http://www.nature.com/icogs/primer/cogs-project-and-design-of-the-icogs-array/

Darren Burgess 4/14/2015 11:42 AM
Formatted: Font:Bold

Darren Burgess 4/14/2015 11:42 AM
Formatted: Font:Bold

normal tissue, in order to extract both germline and somatic variants from the sequencing data?, But whether this is routinely done is not clear.

As a more general comment, before discussing the identified variants themselves (here, but also in the equivalent somatic variants section, as well as in the somatic-germline interplay section), it would be helpful to include a paragraph of study design considerations, such as what are the current and emerging strategies, technologies and challenges (especially those that are more apparent when seeking non-coding rather than protein-coding mutations). This would provide more specific information than the brief overview in the article introduction. I have also suggested a new flow chart figure as a schematic for the study strategies (please see my comments after your tables).]

Unlike somatic variants, germline variants occur in all tissues of the body. **[Au: to bring together concepts from elsewhere in your article, I think that it would be worth expanding this point to explain how the fact that germline variants must be compatible with organismal viability and reproduction when present in every tissue and developmental stage (which is not a requirement for somatic mutations). Thus, this might provide a functional explanation for why cancer-associated germline variants are typically less major/disruptive than the frequent large-scale chromosomal rearrangements that occur somatically.]** 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. Furthermore, noncoding variants can affect 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.

[Au: these are a great range of molecular examples (for both this germline section and the separate somatic section). As general guidance, I think the information is a bit brief, so would benefit from expansion along the following lines:

- **You mention disruption of numerous key oncogenes and tumour suppressor genes, but I think too much reader knowledge is assumed about the roles of these genes. Please add some more information about the biological roles for these oncogenes/TSGs in cancer (even just an extra sentence for a given gene would be really helpful). Then it will be clearer how these alterations are functionally linked to cancer biology.**
- **Tissue specificity is a recurring theme in your article, so where possible please mention reasons for why particular variants affect a certain cancer type (e.g. that the affected oncogene/TSG is itself tissue-specific, or rather that the binding TFs are). You do this a bit (e.g. why TERT mutations are linked to ovarian cancer risk, but the main link to melanoma isn't clear).**
- **I like your use of seminal case studies, as there certainly isn't room for a detailed discussion of every known example of different classes of variants. However, for completion, if there are additional examples it would be helpful to at least cite them, so that readers can get a sense of how widespread the mechanisms are**

Darren Burgess 4/14/2015 11:42 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 4:31 PM
Comment [6]: A single reference of TERT-associated melanoma seems a bit restrictive here. Are there any other examples of noncoding components to other familial cancers, e.g. BRCA1/2, TP53 (Li Fraumeni), APC, VHL etc?

Darren Burgess 4/14/2015 11:42 AM
Formatted: Font:Bold

Ekta Khurana 5/24/2015 2:39 PM
Deleted: 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 [35]), while variants with low penetrance may modulate the effects of somatic variants [36]. 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 typically have a fitness effect at reproductive age, which is perhaps the reason for the prevalence of such variants in the population.**[Au:OK?]**

Darren Burgess 4/15/2015 4:03 PM
Formatted: Highlight

Darren Burgess 4/15/2015 4:09 PM
Formatted: Font:Bold

Darren Burgess 4/15/2015 4:14 PM
Deleted: , etc

Darren Burgess 4/15/2015 4:14 PM
Deleted: e

Darren Burgess 4/15/2015 5:03 PM
Formatted: Font:Bold

Darren Burgess 4/15/2015 4:53 PM
Formatted: Bulleted + Level: 1 + Aligned at: 0.25" + Indent at: 0.5"

Darren Burgess 4/15/2015 5:03 PM
Formatted: Font:Bold

beyond the few examples you mention, and they can then also seek out the cited papers if they want more information.]

Promoter Mutations.

Germline mutations in the *TERT* promoter are associated with familial melanoma [35]. 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 [35]. **[Au: I recall some publications on a promoter SNP in MDM2 (SNP309) but I'm not sure whether that has since stood up to confirmatory analyses – it might be an example worth mentioning if sufficiently supported by studies]**

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) [64]. 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 [65].

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-receptor TFs **[Au:OK? As all TFs must be nuclear to function anyway]** (such as androgen receptor (AR) or estrogen receptor (ER)). This can affect the expression of target tumor suppressor genes and contribute to carcinogenesis [66] (Figure 3B). **[Au: is there a distinction that you are trying to make between the non-hormonal and hormonal TFs? If so, please could you clarify it, as the message of altered TF binding sites seems to be very similar between the 2 paragraphs in this section. Also, I am a bit confused about the suggestion that AR and ER drive tumour suppressors – my limited understanding is that these TFs are generally growth promoting, hence the use of estrogen- and androgen-blocking therapies in breast and prostate cancer, respectively]**

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 [67].

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) [68]. Also, germline copy

Darren Burgess 4/15/2015 4:15 PM
Formatted: Font:Italic

Darren Burgess 4/14/2015 10:01 AM
Deleted: *o*

Darren Burgess 4/15/2015 4:15 PM
Formatted: No bullets or numbering

Darren Burgess 4/23/2015 1:23 PM
Formatted: Font:Bold

Darren Burgess 4/15/2015 4:15 PM
Deleted: *b)*

Darren Burgess 4/15/2015 4:15 PM
Formatted: Font:Italic

Darren Burgess 4/15/2015 5:06 PM
Formatted: Font:Bold

Darren Burgess 4/15/2015 5:05 PM
Deleted: ,

Darren Burgess 4/15/2015 5:06 PM
Deleted: ,

Darren Burgess 4/15/2015 5:11 PM
Formatted: Font:Bold

Darren Burgess 4/15/2015 5:11 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:23 PM
Deleted:

Darren Burgess 4/15/2015 4:15 PM
Deleted: *(c)*

Darren Burgess 4/15/2015 4:15 PM
Formatted: Font:Italic

Darren Burgess 4/15/2015 4:15 PM
Deleted: *(d)*

Darren Burgess 4/15/2015 4:15 PM
Formatted: Font:Italic

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 [69].

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 eQTL and allele-specific expression analyses, have been used to interpret GWAS cancer loci [70-72]. Such studies reveal germline determinants of gene expression in tumors and help establish a link between noncoding risk loci and their target coding genes. **[Au: maybe it's too early to tell, but is there yet an emerging sense of whether cancer-relevant non-coding variants largely affect the expression of classic predisposition genes (and hence are consolidating well-known roles for classic oncogenic or TSG pathways) or whether they instead point to many new and unexpected cellular alterations in cancer]**

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 [74]. 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. **[Au: are there any further examples of interplay, given that TERT is only 1 example (albeit an interesting one). Extra case-studies would help to strengthen this section. For example, is there appreciable overlap between noncoding regions affected by cancer-associated somatic versus germline variants? Even if many results aren't yet known, an exploration of the efforts in this area would be great to enhance this section (e.g. are there aims of the Pan-cancer project to investigate this interplay, and, if so, what are the strategies to achieve this?) So, overall, I think that this section could be developed a lot more. Also, without extra examples, the mention of the Knudson two-hit hypothesis seems a bit confusing and out of context, as the TERT example does not really follow that model.]**

[Au: please see my suggestion above for bringing your 'Noncoding annotations' discussions here, as I think the fit will be better than in their current position, and

Darren Burgess 4/15/2015 5:14 PM
Deleted: expression quantitative trait loci (…QTL) ... [76]

Darren Burgess 4/23/2015 1:20 PM
Formatted ... [77]

Darren Burgess 4/16/2015 4:31 PM
Comment [7]: OK? Same reasoning as above, i.e. headings must be short to fit page layout formats

Darren Burgess 4/16/2015 4:31 PM
Comment [8]: Could this also be relevant to germline variants, e.g. Targeted resequencing of the microRNAome and 3'UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer.
Chen X, Paranjape T, Stahlhut C, McVeigh T, Keane F, Nallur S, Miller N, Kerin M, Deng Y, Yao X, Zhao H, Weidhaas JB, Slack FJ.
Oncogene. 2014 Jun 9. doi: 10.1038/onc.2014.117

Ekta Khurana 5/24/2015 11:46 AM
Deleted: ... [78]

Darren Burgess 4/15/2015 5:27 PM
Formatted ... [79]

Darren Burgess 4/16/2015 10:10 AM
Formatted ... [80]

Darren Burgess 4/16/2015 10:14 AM
Formatted ... [81]

Darren Burgess 4/16/2015 10:21 AM
Formatted ... [82]

this introduction to annotating variants will lead nicely into the computational methods that leverage them]

Computational methods for functional interpretation

[Au: for context, it would be helpful to give a sense of the scale of the challenge here. That is, it is already difficult to prioritise protein-coding variants to distinguish functionally relevant lesions from passenger events. But that the problem is confounded for non-coding variants due to their greater abundance and less-obvious ways of modifying biological function. This doesn't really come across in the current text] 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. **[Au: please could you expand this overall section to provide more explanation and context? Currently the concepts are just mentioned in passing with minimal explanations: I think they will only be accessible to readers with strong familiarity in this area. I have added some specific examples below where more information would be useful but a stronger overall narrative to provide broader context would be useful.]** 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 **[Au: such as what types of 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. **[Au: what is the reason for this? Is it based on the assumption that most SNPs identified by GWASs are non-functional variants that merely tag a region harbouring a nearby causal variant? (And hence the rationale to examine nearby variants in LD?)]** 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.**[Au: and this score is thus a measure of the likely degree of functional impact of the variant?]** 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 [75]. **[Au: that is, if the same element is affected in multiple tumours (of the same type? / of different types?) it is more likely to be functionally relevant?]** 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 [34, 46, 47, 76].

Experimental approaches for functional validation

Several 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

Darren Burgess 4/16/2015 4:31 PM
Comment [9]: OK?

Darren Burgess 4/16/2015 4:02 PM
Deleted: to identify

Darren Burgess 4/16/2015 4:02 PM
Deleted: noncoding variants with functional consequences

Darren Burgess 4/14/2015 12:55 PM
Formatted: Font:Bold

Darren Burgess 4/14/2015 12:55 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:45 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:45 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:45 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:45 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:45 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:45 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:46 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:47 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:47 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:47 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:48 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:49 AM
Formatted: Highlight

Darren Burgess 4/16/2015 10:58 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 10:50 AM
Formatted: Highlight

Darren Burgess 4/16/2015 4:31 PM
Comment [10]: OK?

Darren Burgess 4/16/2015 4:25 PM
Deleted: to understand the

Darren Burgess 4/16/2015 4:25 PM
Deleted: effects of noncoding mutations

introduced in DNA using site-directed mutagenesis or CRISPR-Cas9 system [77] (Figure 4A). Oligonucleotides 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). **[Au: it would be helpful to explicitly mention that these assays are typically designed to capture effects on gene expression caused by variant promoters and enhancers (i.e. only a small subset of the types of variants you discuss elsewhere in the manuscript). Following the descriptions of these transcription assays, it would be useful to at least briefly mention if there are equivalent assays for other types of variants that you discuss above, such as SVs, variant ncRNAs, variant introns, variant UTRs etc. This will make your article more cohesive and rounded. Otherwise it will be unclear to readers why you only mention assays for a subset of the variants that you discuss elsewhere.]** 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 [78]. 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. This approach can provide a genome-wide annotation of noncoding mutations and predict if they are associated with gene expression changes. **[Au: the highlighted text seems a bit convoluted. Perhaps phrase along the lines of RNA-seq is used to assess the resulting expression level of the reporter driven by each variant element, and that this can be achieved in bulk because the promoter/enhancer region is included in the sequence reads, thus serving as an identity tag? Note that I don't completely follow the logic of the approach you describe. You cite STARR-seq, which allows the control element to be sequenced in the RNA because they are placed downstream of the reporter and thus transcribed. However, my understanding is that this only works for enhancers (promoters won't drive a reporter when downstream of it) whereas you imply here and in Fig 4 that it also works for promoters.]** 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. **[Au: is this referring to the LUC reporter shown in the right of fig 4B? If so, briefly mentioning the distinctions from the sequencing-based approaches would be useful, e.g. that throughput is achieved by testing constructs individually in multi-well plates rather than in bulk? (And hence why you describe them above as** "low- to medium-throughput luciferase reporter assays")?**]**

To understand the biological role of driver mutations and to rule out false positives derived from sequencing **[Au: and luciferase?]** approaches, it is imperative to move beyond demonstrations of effects on gene expression to provide a direct biological validation of oncogenic properties of the mutations (Figure 4C). **[Au: OK? Just to emphasize the different stages of validation]** 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.

Darren Burgess 4/16/2015 11:48 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:48 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 4:31 PM
Comment [11]: For example, these types of assays could be worth citing:
Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control.
Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, Segal E.
PLoS Genet. 2015 Apr 15;11(4):e1005147. doi: 10.1371/journal.pgen.1005147. eCollection 2015 Apr.
PMID:25875337

Deciphering the rules by which 5'-**UTR** sequences affect protein expression in yeast.
Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E.
Proc Natl Acad Sci U S A. 2013 Jul 23;110(30):E2792-801. doi: 10.1073/pnas.1222534110. Epub 2013 Jul 5. ... [83]

Darren Burgess 4/16/2015 11:48 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:48 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:12 AM
Formatted: Highlight

Darren Burgess 4/16/2015 11:48 AM
Formatted: Font:Bold, Highlight

Darren Burgess 4/16/2015 11:48 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:34 AM
Formatted: Font:(Default) Arial, Font color: Custom Color(RGB(26,26,26))

Darren Burgess 4/16/2015 11:49 AM
Formatted ... [84]

Darren Burgess 4/16/2015 11:49 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:49 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:51 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:53 AM
Deleted: is

Darren Burgess 4/16/2015 11:51 AM
Deleted: imperative

Darren Burgess 4/16/2015 11:54 AM
Formatted: Font:Bold

Functional validation of noncoding variants is extremely important to understand their biological consequence. High throughput analysis of variants substantially reduces the cost per variant tested (Figure 4D). However, among all the current methods for functional validation of variants, the biological validation for oncogenic properties is the most important but also the most costly. **[Au:OK? Just to clarify why this step can't be just dropped due to expense]** Hence, high-throughout prioritization of putative functional mutations is critical prior to the testing of the most promising candidates in *in vivo* systems, given the lengthy developmental time (years) and costs of *in vivo* assays. **[Au:OK? Just for clarification. Also, as many in vivo assays already exist, could the mention of 'years' be misleading? Presumably most variants could be tested in vivo in a matter of weeks?]**

Conclusions

Cancer arises because of accumulation of multiple driver mutations [43], and 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 of patients[79]. 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. **[Au: this was already mentioned in the article introduction. Any other biases, such as a poorer ability to interpret the functional consequences of non-coding variants?]**

[Au: new paragraph OK? The 'furthermore' part doesn't really follow on from the reasons for coding bias] 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 [80]). 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 types [50]), 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 in the clinic. **[Au:OK?]** 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. **[Au: to me, this seems quite far-fetched for a few reasons: i) it contradicts your idea of many ncMuts acting cooperatively (i.e. you would**

Darren Burgess 4/16/2015 11:56 AM
Deleted: ignificantly

Darren Burgess 4/16/2015 11:58 AM
Deleted: cost of

Darren Burgess 4/16/2015 11:57 AM
Deleted: highest

Darren Burgess 4/16/2015 11:58 AM
Formatted: Font:Bold

Darren Burgess 4/16/2015 11:59 AM
Deleted: establishment of these

Darren Burgess 4/16/2015 12:01 PM
Deleted: considerations of

Darren Burgess 4/16/2015 12:00 PM
Formatted: Font:Italic

Darren Burgess 4/16/2015 12:00 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:00 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:02 PM
Deleted:

Darren Burgess 4/16/2015 12:02 PM
Deleted: --

Darren Burgess 4/16/2015 12:15 PM
Deleted:

Darren Burgess 4/16/2015 12:17 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:17 PM
Deleted: Furthermore, r

Darren Burgess 4/16/2015 12:18 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:18 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:18 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 4:31 PM
Comment [12]: There is already emerging appreciation and research of this phenomenon of cumulative combinatorial effects (at least in the setting of CNVs), such as:

Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome.
Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ. ... [85]

Darren Burgess 4/16/2015 12:24 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:38 PM
Formatted: Font:Bold

probably need to correct many variants simultaneously), ii) protein-targeted therapies are still compatible with ncMuts, because many of your discussions have highlighted how ncMuts dysregulate protein-coding driver genes iii) issues of in vivo delivery in humans and iv) it seems a strange strategy to seek to 'correct' cancer cells rather than just to kill or resect them (perhaps all-trans retinoic acid in APL to induce differentiation is the only example that springs to my mind of a 'corrective' approach). If you include CRISPR ideas, please could you give an example of a type of therapeutic strategy it might be used in? Or alternatively, would this be just used to generate and study noncoding mutations in cancer models?] 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 [81] and to generate a mouse model of lung cancer with a specific chromosomal rearrangement [82]. **[Au: in terms of elaborating on the exome versus WGS debate, is there a case to be made for moving to WGS + RNA-seq? As long as it wasn't cost-prohibitive, this could potentially be used to find driver genes for therapeutic targeting: beyond driver genes harbouring coding mutations, such an approach might find dysregulated drivers based on transcriptional misexpression and/or mutations in regulatory regions. Or is such an approach not practicable for a large number of cases? Either way, I think that it is worth mentioning that sequencing non-coding regions could still be informative for regular, protein-based therapeutics and not just for completely novel future therapeutic approaches such as CRISPR.]** 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), **[Au: although mentioned briefly when discussing 3C approaches in your functional annotations section (and possibly your eQTL mentions), I didn't get a strong sense of this in your review. Perhaps in your later functional effects sections you could emphasise more how variants can be functionally linked to their target genes?]** 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.

Darren Burgess 4/16/2015 12:38 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:38 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:38 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:50 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:50 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:50 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:50 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:50 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:50 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:41 PM
Formatted: Font:Bold

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

[Au: please include glossary definitions for these additional suggested glossary terms]

Darren Burgess 4/16/2015 1:30 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:30 PM
Formatted: Font:Bold

precision medicine

Darren Burgess 4/16/2015 1:30 PM
Formatted: Not Highlight

exomes

genome-wide association studies

Darren Burgess 4/16/2015 1:30 PM
Formatted: Not Highlight

DNase I footprinting

Pseudogenes

Darren Burgess 4/16/2015 1:30 PM
Formatted: Not Highlight

endo-siRNAs

Darren Burgess 4/16/2015 1:30 PM
Formatted: Not Highlight

Negative selection

Darren Burgess 4/16/2015 1:30 PM
Formatted: Not Highlight

chromosome conformation capture

penetrance

single-nucleotide polymorphisms **[Au: glossary definition would be useful to highlight the difference in meaning between SNPs and SNVs, both of which are used in the manuscript (I.e that SNP implies a germline variant of appreciable frequency (<5%??) in the population]**

Darren Burgess 4/16/2015 1:30 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:30 PM
Formatted: Font:Bold

chromoplexy

chromothripsis

kataegis

Darren Burgess 4/16/2015 1:30 PM
Formatted: Font:Not Italic

burden-tests strategy

positive selection

[Au: please add relevant literature citations to your figure legends, even if they are the same ones as are included where the illustrated concepts are discussed in the main text.]

Darren Burgess 4/16/2015 1:03 PM
Formatted: Font:Bold, Not Italic

Figure captions

Darren Burgess 4/16/2015 1:03 PM
Formatted: Font:Bold

Figure 1. Somatic mutations in various cancer types. **[Au: is there a reason for the particular tissue types of cancer that you have chosen to include? Also, could they be arranged in a more logical order left to right, as they don't seem to be ordered according to any of the classifications or mutation data shown. Perhaps ordering according to mutation load would be clearer?]** 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.

Darren Burgess 4/16/2015 12:55 PM
Formatted: Font:Bold

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.**[Au: this figure is useful, but the legend would benefit from expansion to more clearly walk the reader through what is being shown. Also, the 'TF binding motifs' that you show are all the same and don't match up with the coloured sequences below them.]**

Darren Burgess 4/16/2015 12:57 PM
Formatted: Font:Bold

Figure 3. Effect of sequence variants in noncoding regions in oncogenesis. **[Au: I think that this is the most useful figure of your set. We will need to have some back-and-forth about exactly what can be accommodated (figures can be a maximum of 1 full portrait page). For example, we could always remove the hormonal part in fig 3B if there wasn't room (it is largely equivalent to the loss-of-motif situation already shown in fig 3B)]** (A) Overview of the noncoding elements that can be affected. 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.

Darren Burgess 4/16/2015 1:01 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:01 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:01 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 12:58 PM
Deleted: e

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. **[Au: please see my comments in the main text about whether the STARR-seq approach (cited in the main text) can be used for promoter analysis]** (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.**[Au: where is this numerical data from? Please cite, or mention on what assumptions the costs are based. Given that 10-100 and 1,000-100,000 are calculated using very different criteria, I'm confused why the line is so straight, and why the 100 and 1000 points are joined up.]** 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 validation (*in vitro* and *in vivo* tumorigenic assays) of 10 variants.

Darren Burgess 4/16/2015 1:12 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:12 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:16 PM
Formatted: Font:Bold

Darren Burgess 4/16/2015 1:16 PM
Formatted: Font:Bold

[Au: these tables (1 and 2) are very useful]

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, snoRNAs at www-snorna.biotoul.fr, tRNAs at gtrnadb.ucsc.edu and lncRNAs at mitranscriptome.org

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

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 (83, 84)
ANNOVAR	SNV, Indel, SV	Y	Y	N	N	N	openbioinformatics.org/annovar/ (85)
VEP	SNV, Indel, SV	Y	N	N	N	N	ensembl.org/info/docs/tools/vep/ (86)
OncoCis	SNV, Indel	Y	Y	N	N	N	powcs.med.unsw.edu.au/OncoCis/ (87)
GEMINI	SNV, Indel, SV	Y	Y	N	N	N	github.com/arq5x/Gemini (88)
FunciSNP	SNP	Y	N	Y	N	N	bioconductor.org (89)
HaploReg	SNP, Indel	Y	Y	Y	N	N	compbio.mit.edu/HaploReg (90)
GWAS3D	SNP	Y	Y	Y	N	Y	jjwanglab.org/gwas3d (91)
is-rSNP	SNV	N	N	N	N	Y	genomics.csse.unimelb.edu.au/is-rSNP (92)
RegulomeDB	SNV	Y	N	N	N	Y	RegulomeDB.org (93)
SInBaD	SNV	N	Y	N	N	Y	tingchenlab.cmb.usc.edu/Sinbad (94)
CADD	SNV, Indel	Y	Y	N	N	Y	cadd.gs.washington.edu (95)
FunSeq	SNV, Indel	Y	Y	N	Y	Y	funseq2.gersteinlab.org (26, 96)
GWAVA	SNV, Indel	Y	Y	N	N	Y	sanger.ac.uk/resources/software/gwava/ (97)
FitCons	SNV	Y	Y	N	N	Y	(98)

Additional display item?

[Au: in total we can accommodate up to 7 display items (boxes, figures and tables). You have currently provided 6 (4 figures and 2 tables) so there is room for an additional display item if you choose to include one. One option could be a box containing a reworked version of the "Somatic variants in different types of cancer" section that I mentioned above. Other options that are likely to be even more useful are either: i) a flow chart figure showing some of the key study design aspects of identifying cancer-related mutations in non-coding regions – this would show

Darren Burgess 4/23/2015 12:52 PM
Formatted: Font:Bold

strategically the steps of the studies, highlight any particular methodological challenges that result from focusing on non-coding regions and would hopefully give a sense of both somatic and germline studies and their current/future integration; or ii) a table of the different examples of biologically validated non-coding mutations in cancer (ideally more comprehensive than the few case studies described in the main text). My favoured option would be the study designs flow-chart figure]

Darren Burgess 4/23/2015 12:52 PM
Formatted: Font:Bold

Darren Burgess 4/23/2015 12:52 PM
Formatted: Font:Bold

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. Akhtar-Zaidi, B. et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736-9 (2012).
5. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov* **1**, 391-407 (2011).
6. Iyer, M.K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* (2015).
7. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
8. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
9. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317-24 (2012).
10. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
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. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).
35. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-61 (2013).
36. Easton, D.F. & Eeles, R.A. Genome-wide association studies in cancer. *Hum Mol Genet* **17**, R109-15 (2008).
37. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
38. Baca, S.C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-77 (2013).
39. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
40. Holland, A.J. & Cleveland, D.W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat Rev Mol Cell Biol* **10**, 478-87 (2009).
41. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
42. Logsdon, B.A. et al. Sparse expression bases in cancer reveal tumor drivers. *Nucleic Acids Res* (2015).
43. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
44. Gibbs, R.A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).
45. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-19 (2005).

46. 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).
47. 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).
48. Smith, K.S. et al. Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. *Nucleic Acids Res* (2015).
49. Huang, F.W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).
50. 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).
51. Heidenreich, B., Rachakonda, P.S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev* **24**, 30-7 (2014).
52. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
53. Mansour, M.R. et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* (2014).
54. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).
55. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
56. 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).
57. Northcott, P.A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-34 (2014).
58. Breit, T.M. et al. Site-specific deletions involving the tal-1 and sil genes are restricted to cells of the T cell receptor alpha/beta lineage: T cell receptor delta gene deletion mechanism affects multiple genes. *J Exp Med* **177**, 965-77 (1993).
59. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-9 (2013).
60. 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).
61. Buechner, J. & Einvik, C. N-myc and noncoding RNAs in neuroblastoma. *Mol Cancer Res* **10**, 1243-53 (2012).
62. 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).
63. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8 (2010).
64. Grisanzio, C. & Freedman, M.L. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer* **1**, 555-9 (2010).
65. 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).
66. Garritano, S. et al. In-silico identification and functional validation of allele-dependent AR enhancers. *Oncotarget* (2015).
67. 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).
68. 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).

69. 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).

70. .

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

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

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

74. 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).

75. 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).

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

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

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

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

80. 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).

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

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

83. 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).

84. 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).

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

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

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

88. 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).

89. 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).

90. 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).

91. 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).
92. Macintyre, G., Bailey, J., Haviv, I. & Kowalczyk, A. is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics* **26**, i524-30 (2010).
93. Boyle, A.P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
94. Lehmann, K.V. & Chen, T. Exploring functional variant discovery in non-coding regions with SInBaD. *Nucleic Acids Res* **41**, e7 (2013).
95. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
96. Fu, Y. et al. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).
97. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).
98. 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).