**Working title**

Role of noncoding variants in cancer

**Outline of main text:**

### *Abstract (~100 words)*

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

### *Introduction*

In this part we will discuss that this topic is very timely because whole-genome sequencing of tumors is possible now and there is huge interest in knowing what the variants mean. We will discuss the following points: [[EKtoMG: The points that Mark R. raised here are discussed in conclusions/perspective and I elaborated them more. I think it is ok to discuss them in conclusions/perspectives. What do you think ??]].

1) Noncoding regions play varied roles in cancer – e.g. besides sequence variants in these regions, epigenetic changes and expression changes of noncoding RNAs can also drive cancer. We will make the reader aware that our article focuses only on sequence variants in noncoding parts.

2) We will discuss that noncoding mutations are much more abundant than coding by providing a sense of the scale of genomic coverage. [[EKtoMG: I agree with Mark R. that we can include this idea in the figure]]

3) We will discuss that many germline mutations in promoters and enhancers are known to be causal for inherited diseases and we are just beginning to explore the role of noncoding somatic mutations in oncogenesis. Recent studies show 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). Thus, the combined effect of small changes in expression due to noncoding mutations in cancer might be huge. In relation to this, the current idea of binary classification of somatic mutations, i.e. drivers and passengers, may not necessarily be true. This is because the tendency of mutation to cause tumor growth

Ekta Khurana 8/16/14 5:03 PM
**Deleted:**

Mark Rubin 8/16/14 9:06 AM
**Comment [1]:** As an overview should we list out where the major WGS efforts are occurring with some metrics from the pan cancer ICGC efforts? For example: While WGS has only limited clinical application currently, large international programs such as the ICGC have focused the majority of their sequencing efforts on Whole Genome Sequencing. (Or something like that?) If we think about this, there are not a lot of major efforts besides ICGC for cancer. This of course is not true for constitutional germlines but that is again not for cancer care…Not sure if we should point our that TCGA efforts focus to 95% on the coding genome with exome sequencing and only a minority of cases undergo WGS…

Mark Rubin 8/16/14 9:07 AM
**Comment [2]:** This would make a nice graphical figure

Ekta Khurana 8/16/14 5:19 PM
**Deleted:** So cumulative

Ekta Khurana 8/16/14 5:20 PM
**Deleted:** can

Ekta Khurana 8/16/14 5:21 PM
**Formatted:** Not Highlight

may not be the extreme 100% (driver) or 0% (passenger) but rather anywhere in between. While some somatic variants may have a direct role (such as *TERT* promoter mutations found in many different cancer types), others may indirectly modulate important cancer pathways (such as genomic rearrangements perturbing androgen receptor binding sites in a subset of prostate cancers[1, 2]). We discuss various examples under 'Known cases of somatic variants playing a role in tumor development and growth' that illustrate this point.

4) We will also discuss germline variants that have been associated with increased cancer susceptibility, specially the cases where there is an intricate relationship between germline polymorphisms and somatic variants.

### *Main sections*

*1) Noncoding annotations.*

a) What are the various noncoding annotations: transcription factor binding sites, DNase I hypersensitivity sites, noncoding RNAs, etc. We will discuss that the dynamic nature of the epigenome (including various histone marks and DNA methylation) leads to differential activity of regulatory regions in different cellular states. We will also discuss large-scale efforts to annotate functional elements in the genome, such as ENCODE [3] and Roadmap Epigenomics project [4]. This section will also include a discussion of evolutionary conserved regions in noncoding genome e.g. ultraconserved elements [5] and ultrasensitive regions [6].

b) Regulatory regions are often cell-type/tissue specific and thus sequence variants in these regions are more likely to exhibit tissue-specific effects.

c) Multiple approaches are currently used to link cis-regulatory regions to their target genes. For example: different variations of chromosome conformation capture technology [7, 8], correlation of transcription factor (TF) binding and expression across multiple cell lines [9], etc. The resulting linkages can then be studied as a comprehensive regulatory network [10].

*2) Genomic sequence variants.*

a) What are various types of sequence variants: single nucleotide substitutions, small insertion and deletions (indels), and larger structural variants (SVs).

b) There are many differences in patterns of somatic variants and inherited germline variants: (i) A higher fraction of somatic variants contain large genomic rearrangements. Chromosomal aneuploidy is also often observed in cancer cells. (ii) Tumor heterogeneity makes interpretation of somatic variants more complicated. (iii) Various phenomena, such as kataegis [11], chromoplexy [12] and chromothripsis [13] are characteristic only of somatic cancer variants.

*3) Known cases of somatic variants playing a role in tumor development and growth.*

A discussion of how mutations could effect gene expression, e.g. point mutations in transcription factor binding motifs and miRNAs, small insertions or deletions and larger structural variants.

Mark Rubin 8/16/14 9:17 AM

**Comment [3]:** This highlights an important concept that can be viewed in at least two ways. 1) non-coding mutations/variants that act as direct modulators such as TERT (plus other examples). 2) non-coding variants/mutations that can indirectly modulate important cancer pathways. This might effect the regulation of hormone synthesis leading to individuals who are more sensitive to estrogen or androgen due to mutations/variants. –I know that we have examples from francecsca's work that variants alter AR biosynthesis-we have theorized that this could lead to higher risk for cancer I think the ICGC paper from Jan's group also suggested this-rearrangements that we also originally reported in Berger et al, occur in a manner that effects AR binding modulating subsets of cancers. 3)Indirect effects related potentially to conformational alterations in the DNA that then may effect transcription. I am not sure if I know an example of a non-coding gene that leads to this. 4) not sure where to put this but mutations in non-coding promoter areas may effect regulation. Example from Lin et Neoplasia paper (Mir 31) show how mutations in the AR promotor site alter the regulatory control of AR leading to cancer progression…

Ekta Khurana 8/16/14 6:51 PM

**Deleted:** [1, 2]but rather anywhere in between. [[Mark R.: What are your thoughts about the highlighted sentences.]]

Ekta Khurana 8/16/14 6:59 PM

**Deleted:** We will also discuss how there is an increased interest in the cancer community to analyze noncoding mutations because many recent studies point to the role of *TERT* promoter mutations in many different cancer types.

Ekta Khurana 8/17/14 3:21 PM

**Deleted:**

Mark Rubin 8/16/14 9:26 AM

**Comment [4]:** I would also add Chromoplexy as well (Baca et al, Cell). Importantly, we have been thinking about the accumulation of patterns of genomic alterations –it is hard to define for sure at this point the idea might be that an initiating mutation in a DNA damage repair gene is arising early and leads to a cascade of somatic alterations in coding and non-coding regions. These alterations may sometimes be random but also highly recurrent. We have examples in prostate cancer but certainly other cancers t... [1]

This discussion will be combined with examples listed below which we will also display in a Table.

a) Promoters: Recurrent *TERT* promoter mutations observed in many different cancer types[14-17].
b) Enhancers: 'Enhancer hijacking' in medulloblastoma where somatic SVs juxtapose coding sequences of *GFI1* or *GFI2* proximal to active enhancers [18].
c) UTRs: Fusion of 5' UTR of *TMPRSS2* with ETS genes frequently observed in prostate cancer [19].
d) Other TF binding sites: Genomic rearrangements significantly associated with androgen receptor binding sites in a subset of prostate cancers [1, 2].
e) ncRNAs:
-- Expression change of ncRNA can be due to somatic variants like CNVs of ncRNAs. For example: *MALAT1*, which is frequently up-regulated in cancer, was found to be significantly mutated in bladder cancer [20] and amplification of long ncRNA, lncUSMycN, contributes to neuroblastoma progression [21, 22]
-- Pseudogene deletion can effect competition for miRNA binding with the parent gene, which in turn could effect expression of the parent gene [23].
-- Mutations in miRNA binding sites can also effect their binding, e.g. mutations in miR-31 binding site can effect androgen receptor regulation in prostate cancer [24]

*4) Germline inherited variants in noncoding regions that alter cancer susceptibility or patient survival.*

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

(i) SNPs in enhancers on chr 8q24 upstream of *MYC* are related with increased risk for multiple cancer types [25].
(ii) A SNP in *RFX6* gene intron effects *HOXB13* binding and is linked to increased prostate cancer susceptibility [26].
(iii) A SNP in miR-27a gene reduces susceptibility to gastric cancer [27].
(iv) A common SNP in *TERT* promoter modifies the effects of somatic *TERT* promoter mutations in bladder cancer on patient survival [28].
(v) Splice site mutation in the intron of *BRCA2* has implications for familial breast cancer [29].

(b) eQTL analysis has been used to interpret risk loci [30, 31]. We will also discuss why usually there is no eQTL analysis for somatic variants (since cancer is heterogeneous so these variants are rare). Cryptic effects of noncoding mutations have also been noted where germline variants exhibit allelic effects in tumor [32].

These examples illustrate how the effect of noncoding mutations and interplay between germline and somatic variants can be complex. We will discuss the relevance of two hit hypothesis (where one allele is disabled by a germline variant and the other by somatic variant)

Mark Rubin 8/16/14 9:27 AM
**Comment [5]:** Francesca has this type of data that we might be able to include.

Ekta Khurana 8/17/14 4:27 PM
**Deleted:** UTR/promoter juxtaposed to gene, e.g. *TMPRSS2-ERG* fusion [[Mark R.: does this sound correct (does the UTR fuse or promoter) and best reference to cite??]]

Ekta Khurana 8/17/14 5:15 PM
**Formatted:** Font:Italic

Ekta Khurana 8/17/14 5:15 PM
**Formatted:** Font:Italic

Ekta Khurana 8/17/14 4:24 PM
**Formatted:** Font:Italic

Ekta Khurana 8/17/14 4:22 PM
**Formatted:** Font color: Black

Ekta Khurana 8/16/14 6:48 PM
**Deleted:** c

Ekta Khurana 8/17/14 5:19 PM
**Deleted:** e

Ekta Khurana 8/17/14 5:19 PM
**Deleted:** r

Ekta Khurana 8/17/14 5:19 PM
**Deleted:** na

Ekta Khurana 8/17/14 5:19 PM
**Deleted:**

Ekta Khurana 8/17/14 5:22 PM
**Formatted:** Font:Not Italic

Ekta Khurana 8/17/14 5:22 PM
**Formatted:** Font:Not Italic

Ekta Khurana 8/17/14 5:47 PM
**Formatted:** Font:11 pt

Ekta Khurana 8/17/14 5:47 PM
**Deleted:** .

Mark Rubin 8/16/14 9:41 AM
**Comment [6]:** As mentioned above, the is the Lin et al paper from Noeplasia showing mutations in the AR regulatory area that occur modulating miR31 regulatation PMID:23233736
. There are also some nice papers reporting on NMYC regulation by lncRNAs(PMID:24906397) and PMID: 22936790–Dimple is aware of this –similar class as the NEAT1 story (in press).

Ekta Khurana 8/17/14 6:50 PM
**Deleted:** .

for noncoding regions. We will also use the above examples to discuss how the notion of driver mutations may not be binary since somatic mutations can influence cancer growth to varied extent based on the presence of other germline and somatic variants.

*5) Different types of cancer*

(a) Discussion of total numbers of mutations and numbers of noncoding vs coding mutations in different types of cancer. For example, tumors of self-renewing tissues (such as colorectal) contain more mutations than non-self-renewing ones (such as glioblastomas and pancreatic cancers) [33].

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

*6) Computational methods to identify noncoding somatic variants with functional consequences.*

(a) Discussion of currently available computational methods to identify noncoding driver mutations from whole-genome sequencing data, for example, FunSeq [6], CADD [34] and GWAVA [35]. We will also depict these in a Table with associated website links.

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

Finally, we will discuss experimental ways to test which noncoding mutations have functional effects (e.g. genome editing using CRISPR, luciferase reporter assays, high-throughput assays, etc). We will also discuss the scale and cost of all the techniques and summarize them in a Table.

### *Conclusions/perspective*

(a) Cancer arises because of accumulation of multiple mutations -- some of these drivers could be noncoding. There is a bias in literature for driver noncoding mutations because people haven't explored these regions for cancer drivers to the same extent, for example most studies have been focusing on exomes including the majority of TCGA studies. There is an increased realization of this and there is an ongoing collaborative effort between TCGA and ICGC called Pan-Cancer Analysis of Whole Genomes (PCAWG), which will try to identify important noncoding mutations in ~2500 tumor and matched normal whole-genomes.

(b) There is a debate in the community about whether we should look at noncoding/whole genomes vs exomes. Studies of somatic noncoding mutations are currently mostly for research purposes, as opposed to regular clinical use. This is primarily because current therapeutic approaches attempt to target proteins. It is possible that alternate methodologies, such as genome editing using CRISPR, may be used in future (e.g. CRISPR/Cas9 mediated editing has been used for HIV in cell lines [36] and muscular dystrophy in mice [37]). However, noncoding germline variants associated with increased cancer susceptibility should be important for risk assessment and potentially for preventive approaches.

Ekta Khurana 8/17/14 5:56 PM
**Deleted:** *[[ More text from Mark R.]]*

Ekta Khurana 8/17/14 6:16 PM
**Deleted:** N

Ekta Khurana 8/17/14 6:05 PM
**Deleted:** and may lie in noncoding regions

Mark Rubin 8/16/14 9:46 AM
**Comment [7]:** This is an important topic but I think the data will be highly sparse and imperfect. There is a good paper to site from Bert Vogelstein (PMID:23539594) suggesting that tumors have typically 2-less than 10 driver mutations. Since most of the studies have focused on canonical coding mutations, we might speculate that non-coding effects are certain to play a role but have yet been adequately described. For example, are tumors that have a paucity of coding mutations controlled more by non-coding and epigenetic alterations? This information should be coming out more and more with the increased WGS from the ICGC integrated with other datasets….

Ekta Khurana 8/17/14 6:58 PM
**Deleted:** [33]

Ekta Khurana 8/17/14 6:58 PM
**Formatted:** Indent: Left: 0"

Mark Rubin 8/16/14 9:48 AM
**Comment [8]:** Dimple has worked on this. I think that describing Crisper might be nice. The future might even see therapeutics with this type of editing.

Ekta Khurana 8/17/14 6:27 PM
**Deleted:** . *[[Mark R.: should we add something else here ?]]*

Ekta Khurana 8/17/14 6:46 PM
**Deleted:** was

Ekta Khurana 8/17/14 6:45 PM
**Deleted:** and proposed for therapeutic application

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

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

## Proposed display items

(1) Table of noncoding annotations
(2) Table of somatic sequence variants important in cancer
(3) Table of germline sequence variants related to altered cancer susceptibility
(4) Table of computational methods to prioritize noncoding mutations with functional effects
(5) Table of experimental techniques to validate them
(6) Schematic for role of various noncoding annotations and sequence variants in them in oncogenesis

## Key references

1. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
2. 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).
3. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
4. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317-24 (2012).
5. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321-5 (2004).
6. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
7. 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).
8. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189-91 (2012).
9. 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).
10. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).

11. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
12. Baca, S.C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-77 (2013).
13. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
14. Huang, F.W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-9 (2013).
15. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-61 (2013).
16. 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).
17. Heidenreich, B., Rachakonda, P.S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr Opin Genet Dev* **24**, 30-7 (2014).
18. Northcott, P.A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-34 (2014).
19. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).
20. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-9 (2013).
21. 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).
22. Buechner, J. & Einvik, C. N-myc and noncoding RNAs in neuroblastoma. *Mol Cancer Res* **10**, 1243-53 (2012).
23. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8 (2010).
24. 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).
25. Grisanzio, C. & Freedman, M.L. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer* **1**, 555-9 (2010).
26. 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).
27. 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).
28. 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).
29. 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).
30. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
31. Xu, X. et al. Variants at IRX4 as prostate cancer expression quantitative trait loci. *Eur J Hum Genet* **22**, 558-63 (2014).
32. Ongen, H. et al. Putative cis-regulatory drivers in colorectal cancer. *Nature* (2014).
33. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
34. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
35. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).

36. Hu, W. et al. RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proc Natl Acad Sci U S A* (2014).

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