Title	Prioritisation of non-coding somatic mutations in cancer
Authors(s)	Piraino, Scott William
Publication date	2017
Publisher	University College Dublin. School of Medicine
Link to online version	http://dissertations.umi.com/ucd:10149
Item record/more information	http://hdl.handle.net/10197/8697

Prioritisation of Non-coding Somatic Mutations in Cancer

Scott W. Piraino

Student number: 14206789

This thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Research Masters in Medicine and Medical Science

Conway Institute of Biomolecular and Biomedical Research, University College Dublin

Head of School: Prof. Patrick Murray

Supervisor: Dr. Simon J. Furney

Research Masters panel: Dr. Simon J. Furney, Prof Walter Kolch, Prof Kenneth Wolfe

Submission date: May 2017

Table of Contents

1 Abstract

The identification of somatic mutations that play a causal role in tumour development, so called “driver” mutations, is of critical importance for understanding how cancers form and how they might be treated. Several large whole exome sequencing projects have identified genes that are recurrently mutated in cancer patients, indicating a possible causal role in tumourogenesis. While the landscape of coding drivers has been extensively studied and many of the most prominent driver genes are well characterised, comparatively less is known about what driver mutations may reside in the non-coding regions of the genome. Using mutations identified in over 1300 whole cancer genomes, I have identified regions, both coding and non-coding, that are recurrent targets of somatic mutations in cancer. Using both recurrence and information on evolutionary conservation to score regions of the genome as potential driver mutations, I have identified putative driver regions that include both well known drivers as well as novel recurrently mutated regions.

2 Statement of Original Authorship

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

3 Acknowledgements

I would like to thank my advisor, Dr. Simon J. Furney, for giving my the opportunity to do this work and for the guidance he provided to me. I also thank the Conway community of professors, researchers, and students who made me feel welcome and provided an excellent environment to work in. I’m grateful to the friends I made during my time in Dublin, and to my family for their support in pursuing this accomplishment.

4 Introduction

4.1 Cancer genomics overview

The Catalogue of Somatic Mutations in Cancer (http://cancer.sanger.ac.uk/cosmic) contains over 21,000 genomes or exomes from cancer patients [1]. Many of these sequences come from large multi-institution consortia whose aim is to comprehensively characterise the molecular variations that occur in human cancers by identifying genes containing somatic mutations [2]. This task is complicated by the fact that most mutations within the genome of a cancer cell are “passenger” mutations which are not directly implicated in tumour development [3, 4]. Hence, it is not always clear whether a given mutation in a patient’s tumour is a passenger mutation or a “driver” mutation, which does confer a selective advantage to cancer cells and is therefore likely to be involved in pathogenesis. A major effort that has emerged in the field of cancer genomics is the systematic identification of cancer driver genes (genes that can contain driver mutations) [5, 6]. The identification of driver genes is critical both in understanding the molecular events that take place within cancer cells as well as for the prioritisation of targets for therapeutic intervention. In addition, the sequencing of thousands of cancer exomes and genomes has allowed the inference of distinct regional and global mutational signatures and processes from the genomic variations that have occurred during tumour development [7, 8].

Most cancer mutation studies have focused exclusively on variants that alter the amino acid sequences of protein coding genes (non-synonymous mutations), and have assumed that coding mutations that do not alter the amino acid sequence of a protein are passenger mutations. Hence, translational approaches using insights from genomic analyses to develop novel therapies or clinical genetic tests have focused on non-synonymous mutations in driver genes. However, several recent studies have demonstrated that driver mutations do not need to alter a protein’s amino acid sequence to drive cancer [9-12]. Furthermore, mutations that reside outside of coding sequence have been identified as putative driver mutations [9, 12] and the vast majority of somatic mutations in cancer are within non-coding regions, which comprise >98% of the of the genome. Given the diversity of functional elements that may reside within non-coding DNA, it is feasible that in addition to protein-coding driver genes, there exists a class of driver regions within the non-coding genome that can contribute to tumourigenesis. There are few genomic studies that have attempted to identify these non-coding driver regions, and future efforts will likely face unique challenges compared to the search for coding driver genes.

My aim was to develop a method to prioritise non-coding regions of the genome in terms of their potential to contain driver mutations. Based on the principle that driver regions should be recurrently mutated and have a higher likelihood of containing functional mutations, I developed a scoring method that uses both recurrence and conservation that can be used to identify putative driver mutations, both coding and non-coding. I used a set of over 1300 whole cancer genomes to identify recurrently mutated non-coding regions that may be under selection in cancer. Below, I review the several concepts that can be helpful in the identification of driver mutations, and then present the results of my method applied to this set of whole cancer genomes.

4.2 Somatic mutation rates in cancer genomes

Genome and exome sequencing studies have confirmed vast heterogeneity in the mutational rates and signatures in cancer genomes. Tumours exposed to mutagenic environmental factors such as the ultraviolet B component of sunlight in melanomas and tobacco smoke in lung cancers also have elevated mutation rates [13]. Furthermore, distinctive C>T (G>A) mutations at the 3’ end of dipyrimidines predominate in UV exposed melanoma genomes [14, 15], with G>T (C>A) transversions are prevalent in the lung cancer genomes of smokers [16, 17]. Thus, these mutational signatures are characteristic of the DNA damage caused by UVlight and tobacco smoke. Several signatures have been identified which correlate with features including patient age, exposure to mutagens such as tobacco smoke or certain chemotherapy agents, DNA repair deficiencies, as well as numerous other mechanisms that affect genomic mutation rates [7, 18-20].

The local somatic mutation rate also depends on factors that vary along the genome. The somatic mutation rate in windows along the genome is positively correlated with histone marks of closed chromatin, and negatively correlated with histone marks for open chromatin [8]. Consistent with these observations, DNase I hypersensitive sites within the genome, a measurement used to identify regulatory regions, have lower mutation rates compare to flanking DNA [21]. This decreased mutation rate in DNase hypersensitive sites is absent in individual genomes of tumours that have mutations in various DNA repair genes [21]. Decreased variation along the genome has also been observed at a global level in tumours that are deficient in mismatch repair [22], highlighting DNA repair processes as a central factor that contributes to genomic variation in somatic mutation rate. CTCF/cohesion-binding sites (CBSs) mapped in a colorectal cancer cell line showed an excess of mutations compared to flanking regions in a sample of 213 colorectal cancers, while colorectal cancers deficient in Polymerase ε proofreading showed a depletion of mutations [23]. This effect is amplified in late-replicating regions of the genome [23]. Given the importance of repair in terms of the relationship between replication timing and somatic mutation rate [22] as well as evidence showing that the interplay of replication and repair plays a role in germline substitution rates [24] the relationship between CBSs, POLE deficiency, and mutation rates may have important implications for the mechanism by which somatic mutations occur. CBSs also display substantial overlap with recurrent mutations in regulatory regions [23], which could be due either to positive selection of these sites or to the underlying CBS-associated mutational process.

The correlation between chromatin features and somatic mutation rate is strongest for chromatin features measured in the cell type from which a cancer originated, suggesting that this relationship is cell type specific [25]. Gene expression is also correlated with somatic mutation rate, with regions of high expression having lower mutation frequencies compared to regions of low expression [15, 17, 26]. Mutations are more frequent on the untranscribed strand of genes compared to the transcribed strand, leading to the hypothesis that the relationship between mutation rate and expression is due to transcription coupled repair [7, 26]. A final genomic factor that is known to affect somatic mutation rate is replication timing. Areas of the genome which replicate late during DNA replication have higher mutation rates compared to regions which replicate early [26, 27]. Additionally, regions in close physical proximately to late replicating areas also have elevated mutation rates [27]. Mutations may also display clustering, a phenomenon known as kataegis. These mutations tend to be C>T or C>G mutations in a TpC context and often are associated with genomic rearrangements, and indicate a role of APOBEC family enzymes [7, 28]. Explicitly controlling for the influence of genomic factors which effect mutation rate when identifying driver genes lowers the number of significantly mutated genes identified and eliminates many genes that are highly mutated but for which a role in cancer is biologically implausible, such as olfactory receptors [26].

As methods for the identification of driver genes are extended to larger number of samples and to other regions of the genome, it will be critical to apply strategies that can increase the power to detect true drivers and decrease the rate of false positive discoveries. In particular for strategies that seek to identify recurrently mutated genes, appropriately modelling the background mutation rate can reduce the rate of false positive genes discovered [26]. This is important as some tumours / regions of the genome have a higher mutation rate as a result of enhanced exposure to certain mutational processes, rather than selection for mutations within the region (Figure 1).

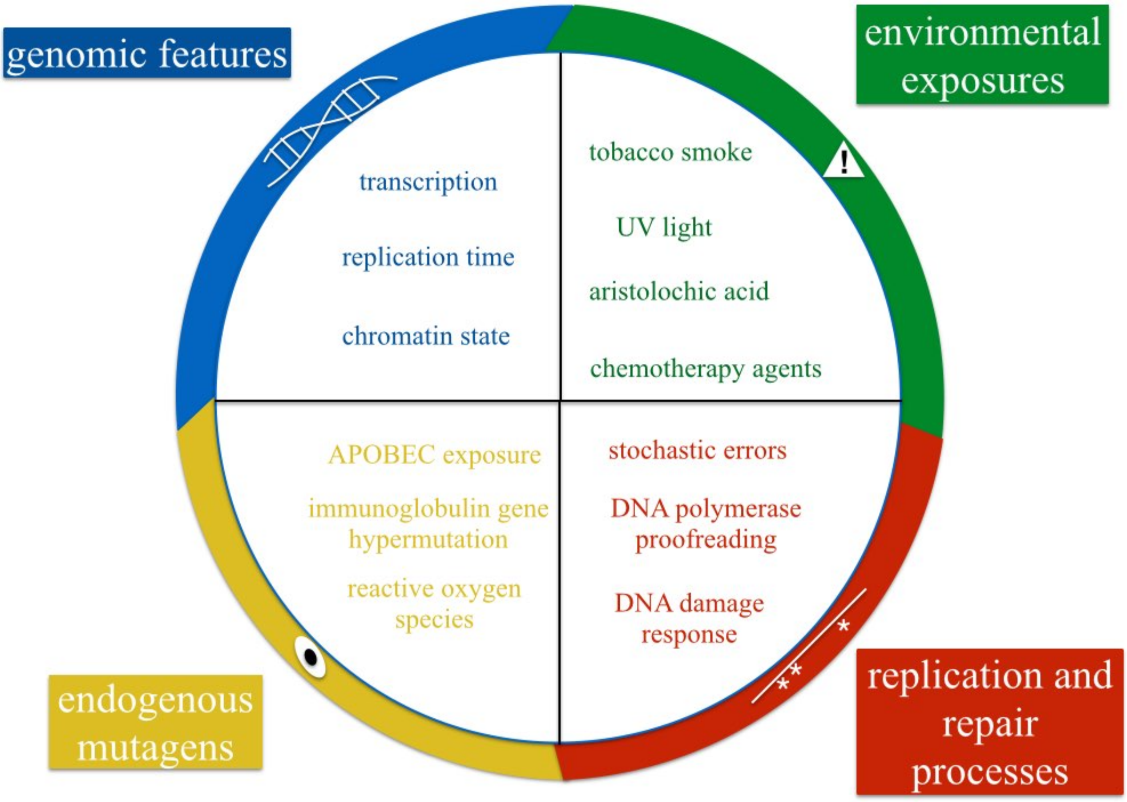

Figure 1: Many factors contribute to mutation rate variation both within and between genomes. I conceptually divide the determinations of mutation rate into four categories. Genomic features such as chromatin state and replication time are major determinants of mutation rate variation across regions of the genome. Many highly mutated genomes result from DNA repair deficiencies, either sporadic or inherited, or because of intense environmental exposures. These factors also leave unique mutational patterns within individual genomes. Normal somatic cells also naturally accumulate mutations, both because of stochastic replication and proofreading errors, as well as endogenous mutation process, such as reactive oxygen species production.

4.3 Identification of driver genes

The heterogeneity in mutational rates and processes described above can obfuscate attempts to identify driver genes or regions in cancer. Many methods have been devised to identify driver genes using either whole exome sequencing (WES) or whole genome sequencing (WGS) [29]. Several studies have applied multiple driver identification methods simultaneously to the same set of cancer sequences to comprehensively identify large sets of driver genes in pan-cancer datasets [30, 31]. These studies used several strategies to identify drivers, including searching for genes with large numbers of somatic mutations [26, 32], identifying genes with significant clustering of mutations along the gene's linear sequence [30, 33], and identifying genes that are enriched for various classes of functional mutation [30, 34, 35].

Mutation rate affects the rate of false negatives in cancer genomic studies in addition to the rate of false positives. Saturation analysis using data from 21 tumour types indicates that while most driver genes that are mutated in greater than 20% of tumours can be detected with currently available sample sizes, many novel drivers that are mutated in less than 20% of samples will remain undiscovered using current sample sizes and computational methods [30]. This problem of identifying genes mutated at low frequencies has been described by analogy of genes to mountains and hills. "Mountains" describe genes that are mutated in a large fraction of tumours, whereas "hills" describe genes that are implicated in cancer, but are mutated at much lower frequency compared to "mountains" [5, 36].

Most studies aimed at identifying driver mutations have focussed on non-synonymous (amino acid altering) mutations in protein coding genes as these are more likely to have a significant functional impact upon the cell. Numerous cancer genes are known to be recurrently targeted by non-syonymous mutations including *BRAF* [37], *APC* [5, 38], *TP53* and *KRAS* [5]. Oncogenes and tumour suppressor genes display different patterns of nonsynonymous mutation. Tumour suppressor genes have an excess of inactivating mutations, such as premature stop mutations, while oncogenes show clustering of mutations near specific amino acid residues [5, 39] which may indicate targeting of functional sites, such as ATP or GTP binding sites [40]. However, the large numbers of somatic mutations that do not alter the amino acid sequences of proteins are not generally tested for the potential to function as driver mutations. Several recent studies have challenged this interpretation and have elucidated roles for synonymous mutations in tumours [10, 11]. In addition, many mutation types, including synonymous, nonsynonymous, and intronic mutations can alter mRNA splicing. Therefore, it may be vital to integrate information on both coding and non-coding mutations to gain a full understanding of the role of somatic mutations in a tumour (Figure 2).

4.4 Non-coding mutations

Early sequencing projects suggested that, like coding mutations, the vast amounts of non-coding mutations are likely passenger mutations [17]. Despite this, mutations with the potential to disrupt binding motifs in the flanking regions of genes were identified, raising the possibility of regulatory driver mutations [17]. The discovery of driver mutations within the promoter of the *TERT* gene [41-43] has generated interest in the role that non-coding mutations play in driving cancer. Telomerase reverse transcriptase (TERT) is part of a complex which catalyzes the lengthening of telomeres, and is generally transcriptionally repressed in noncancerous somatic cells [44]. *TERT* promoter mutations have been observed to generate binding sequences for ETS transcription factors, upregulating *TERT* expression [41] and are highly recurrent across multiple tumour types, suggesting that they are driver mutations in these tissues [9, 12, 43].

In addition to this high frequency in many cancer types, somatic mutations in the *TERT* promoter are also associated with clinical outcomes. Recently, Schulze *et al.* found *TERT* promoter mutations in hepatocellular carcinoma (HCC) were

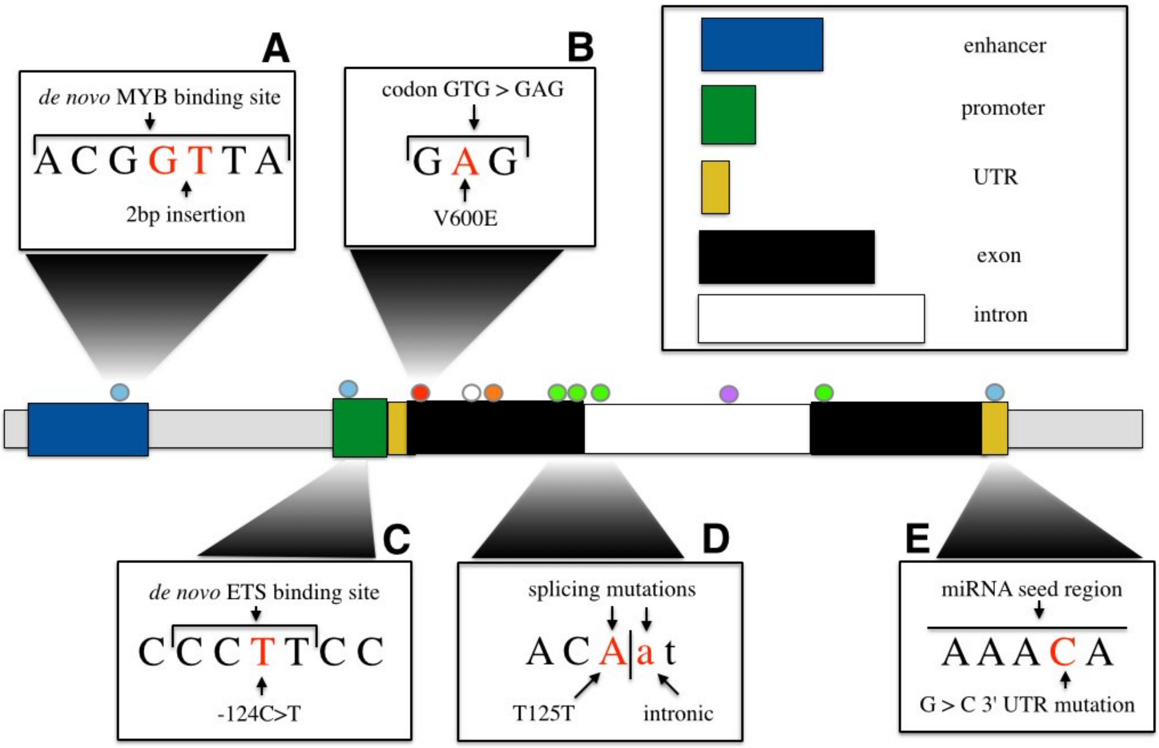

Figure 2: I illustrate the potential consequnces of mutations using a hypothetical gene. Mutations may have diverse functional consequences, producing various mechanisms by which somatic mutations may driver cancer. Mutations may affect distal regulatory elements (**A**) such as enhancers, either creating or destroying transcription factor binding sites. Mutations may result in single amino acid subsitutions (**B**) as critical residues of a protein, critically altering function. Similar to mutations in distal elements, promoter mutations (**C**) may also alter transcription factor binding. Various types of mutations, including nonsynonymous, synonymous, and intronic mutations (**D**) may effect splicing, particularly at exon-intron borders. UTR mutations (**E**) may have various functional consequences, such as altering miRNA binding or other regulatory functions. For each of these categories, I give an example of previously reported recurrent mutations.

enriched in alcohol-related HCCs and were early events in tumour progression in cirrhotic livers [45]. Poorer survival has been observed among patients with *TERT* promoter mutations in clear cell renal carcinoma [46] bladder cancer [47], thyroid carcinomas [12] and gliomas [48]. This relationship also appears to be affected by patient genotypes at the *TERT* promoter SNP rs2853669 [46, 47, 49]. In a panel of 23 urothelial cancer (UC) cell lines, *TERT* promoter mutation was associated with increased *TERT* mRNA levels, protein levels, and telomerase enzymatic activity, as wells as telomere length [50]. Analysis of gene expression data from two separate cohorts of UC patients revealed that *TERT* mRNA expression is associated with poor disease-specific survival in UC, despite previous reports that *TERT* promoter mutation status is not associated with clinical features such as stage and grade in UC [51, 52]. This may be because mRNA expression levels are a stronger prognostic factor in UC [50] or that the impact of *TERT* promoter mutations on survival tend to be independent of other clinical features such as stage.

A recent study by Katainen *et al.* underlines the role for mutations in binding sites in the non-coding genome [53]. The authors integrated whole genome sequencing of >200 colorectal tumours with chromatin immunoprecipitation sequencing to identify frequent mutations in CTCF/cohesin-binding sites in microsatellite stable tumours. The mutations were significantly associated with a particular mutational signature idenified by Alexandrov *et al.* [7], however the consequences of the mutations remain to be elucidated.

Recurrent non-coding mutations have also been identified within the enhancer of *TAL1* in T-cell acute lymphoblastic leukaemia (T-ALL) [54]. Small insertions have been observed near the *TAL1* locus, which create transcription factor binding sites. These mutations also show increased transcription of the target gene in luciferase assays. This increase in transcription is not observed in *MYB* knockout cells or human embryonic kidney cells, suggesting that the impact of *TAL1* enhancer mutations are dependent on tissue-specific regulatory factors [54]. ChIP-seq data indicates that *TAL1* enhancer mutations create *de novo* MYB binding sites which are critical for *MYB* binding. This raises the possibility that non-coding mutations may be important not only for their effect on gene expression, but also because they can alter the transcriptional networks within the cell to create novel pathways.

Mutations in UTRs are another important class of non-coding mutations that are likely to contain oncogenic drivers. There is computational evidence that the 3' UTRs of dosage sensitive onocogenes are enriched for somatic mutations in cancer suggesting that UTR mutations are an important class of mutations that remain under-investigated in cancer genomic studies [10]. In addition, 3' UTR mutations of CD274 in gastric cancer patients have been shown to disrupt miRNA-mediated degradation of the mRNA transcript, resulting in overexpression of CD274 [55], and recurrent 5' UTR mutations in the gene RPS27 have also been identified in melanoma [56].

Recently, several studies have focused on regulatory regions in an attempt to comprehensively identify regulatory regions that contain driver mutations [9, 12]. Given the diversity of regulatory elements present within the non-coding regions of the genome, studies that extend this work are likely to produce valuable insights about the evolution of cancer genomes, but will also face unique challenges that are distinct from the study of coding regions.

4.5 Challenges in the identification of non-coding drivers

Although procedures for identifying coding driver regions can be applied to non-coding regions, the analysis of non-coding regions is more challenging. While WES is sufficient for studies focusing on coding mutations, the requirement for WGS to examine non-coding mutations comprehensively significantly limits the number of samples available for analysis. Given that current studies aimed at coding driver genes are already potentially underpowered for genes mutated at lower frequencies in patients, studies of non-coding mutations with currently available WGS samples will likely face severe power limitations, especially if non-coding drivers tend to be lower in frequency compared to coding drivers (more hill-like than mountain-like) or if mutiple correction penalties are more severe when considering non-coding regions.

The investigation of non-coding mutations also faces several analytical challenges. While genes generally form well-defined genomic regions, the characterization of functional non-coding regions is still in its comparative infancy. Data that have been employed to identify non-coding regions of interest include Ensembl gene annotations [9, 57] as well as data from the ENCODE project [12, 58]. While a more unbiased approach would look at the entire non-coding genome, more restrictive methods could employ stringent criteria to define regulatory regions, and to declare only regulatory regions near known cancer genes to be the regions of interest. A recent study of non-coding mutations in cancer used various methods of defining regions of interest, and it is apparent that more restrictive methods, using stringent criteria to define regulatory regions, increase the power to detect drivers that are among these regions [12].

Once regions of interest have been defined, they must be prioritised based on the strength of positive selection in tumours (Figure 3). A recent study by Weinhold *et al*. used three independent methods to identify putative non-coding driver mutations and identified [9]. The recurrence and clustering (hotspot) based methods applied in this study are similar to methods used for coding mutations. Both methods highlighted the *TERT* promoter as recurrently mutated, and identified statistically significant putative driver non-coding regions in the promoters of the *PLEKHS1* and *WDR74* genes. Notably, the recurrence based method employed two procedures for estimating the background mutation rate, referred to as the local and global methods. LARVA [59] is another recurrence-based method that attempts to improve upon other methods by accounting for heterogeneity in mutation rate in different genomic regions by modelling mutation counts as beta-binomial rather than binomial. Comparison to a simple binomial model showed a better fit of the beta-binomial to observed counts, as well as a decreased number of statistically significant under the beta-binomial model. Downstream analysis of the outputs from LARVA [59] also identified enrichment of mutations in non-coding regions such as BCL3 and CTCF binding sites.

A third method employed by Weinhold *et al*. [9] searched for regions that were enriched for mutations that either create or disrupt ETS binding sites. Among other significant regions, this method identified mutations in the *SDHD* promoter which showed significant association with both mRNA expression and patient survival using melanoma exome sequencing data. *SDHD* promoter mutations have been evaluated in an independent melanoma sample where they were observed in a smaller proportion of patients (4% in Scholz et al. vs 10% in Weinhold et al.) and without a significant association with survival. This example shows the power of using well chosen functional metrics to discover novel regions that can be investigate by downstream analyses and in further independent analyses. Recently, Melton *et al.* used regulatory information from the RegulomeDB resource [60] to search for recurrent somatic mutations in regulatory regions, confirming the *TERT* promoter and identifying several novel regulatory regions in genes implicated in cancer [61]. Other genomic variables which relate to function such as evolutionary conservation could also be used, which will reveal regions with other interesting mutational features. The method FunSeq [62] and it's extension FunSeq2 [63] use human polymorphism data to annotate non-coding regions that are under purifying selection [62]. Application of FunSeq to somatic variants in cancer identified mutations that occur in regions that are subject to purifying selection at the population level, including *WDR74* promoter mutations, which were identified as recurrent by Weinhold *et al.* [9]. FunSeq2, as well as other methods of annotating non-coding variants, has been applied to promoter mutations in a melanoma cell line and was found to be moderately predictive of mutational impact in terms of transcriptional effect in reporter assays [64].

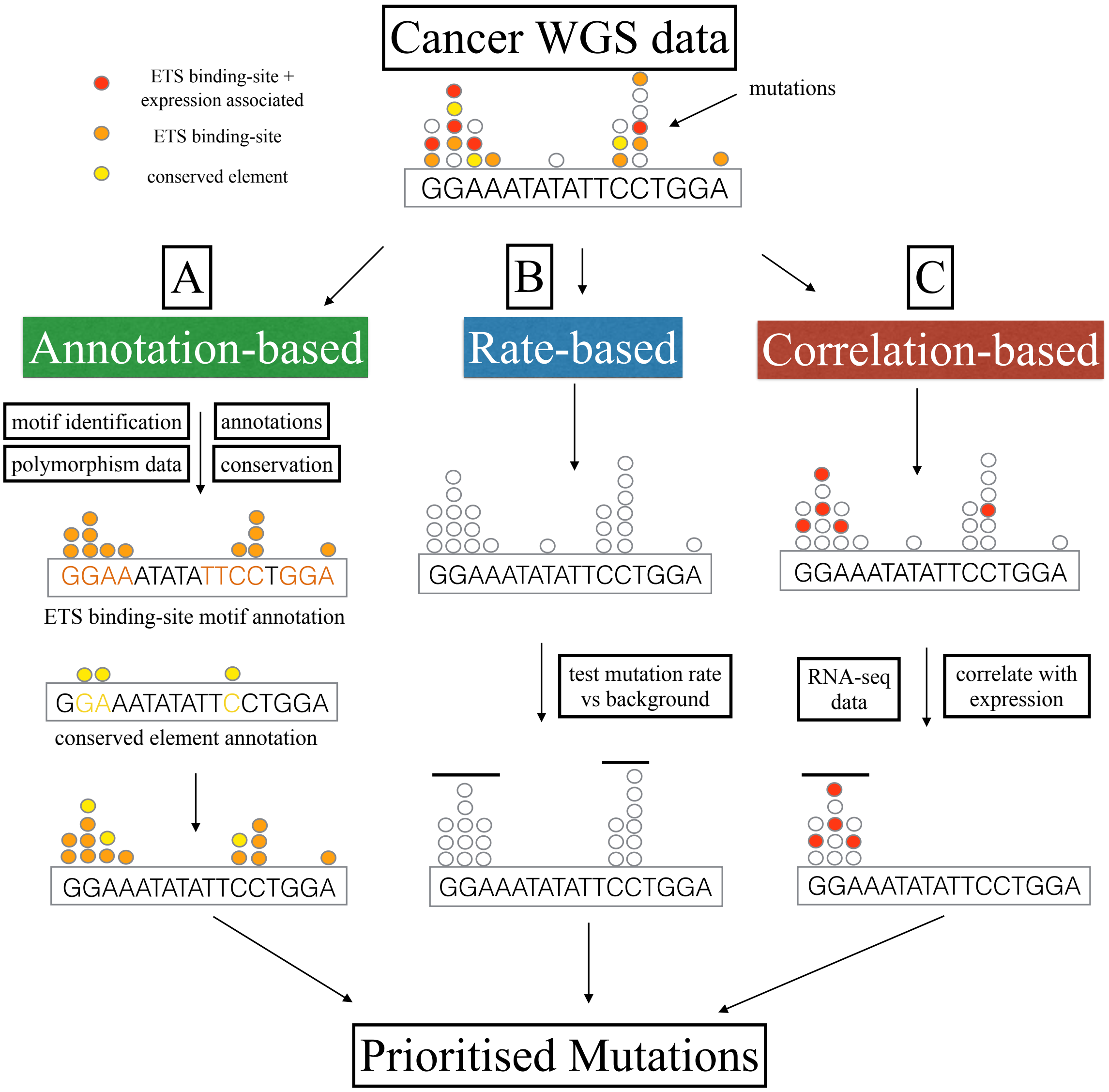

Figure 3: Many methods have been devised to interrogate non-coding mutations in cancer. I divide these methods into three categories: annotation-based methods (**A**) rely on using annotations or other information, such as information about regulatroy motifs or evolutionary conservation, rate-based methods focus on identifying regions mutated more often than expected under some background model of mutation rate, and correlation-based methods (**C**) seek to identify correlations between tumour level mutation status within a region and some other variable, such as RNA expression or clinical variables.

Putatively functional promoter mutations were relatively frequent in this analysis (4 putative functional promoter mutations in a single genome, 17% of analysed promoter mutations) although only 1 was recurrent in TCGA melanoma data, suggesting that even mutations with altered transcriptional activity may be passenger mutations [64].

A local background mutational rate method was used in the development of the software SASE-hunter in order to identify putative positively selected promoter regions, some of which are associated with gene expression or clinical features [65]. The authors searched for signatures of accelerated somatic evolution in non-coding regions and found that lymphoma patients with mutations in the *MYC* promoter were significantly younger compared to others and that patients with mutations in the *BCL2* promoter were significantly older. In addition, the mutational signatures in the promoters of *BCL2, TCL1A*, or *BTG2* were associated with less favourable clinical outcome. In melanoma, patients with mutations in the *RBM5* promoter had significantly shorter survival and were more likely to have distant or lymph node metastasis.

Fredriksson *et al*. [12] used a method which correlated mRNA expression of genes with non-coding mutations in the region surrounding the gene. This analysis revealed that *TERT* promoter mutations are unique in the strength of association between the presence of mutations and mRNA expression. Associations between mutations surrounding other genes do not show as significant signals of association with mRNA expression and it is unclear why mRNA-mutation associations outside of *TERT* promoter mutations, if they exist, are so difficult to discover. In part, this lack of association may be due to sample size considerations. Even truly recurrent mutations may be rare or confined to certain cancer types, and may therefore require larger sample sizes to achieve genome-wide significance. The mutation-expression correlation analysis in Melton *et al.* [66] even failed to identify *TERT*, suggesting that small sample sizes or lack of sample diversity (e.g. the tumour types included) may impact this analysis. Fredriksson *et al.* [12] implemented several strategies that reduce the number of regions considered, which can help limit multiple testing penalities. These strategies may be useful in future work. However, several simulations in Fredriksson *et al.* [12] show that their method is sufficiently powered to detect mutations with effects weaker than *TERT*, indicating that factors other than sample size may contribute to lack of association.An interesting observation made by Fredriksson *et al*. is the apparent mutual exclusivity or co-occurrence that exists between some non-coding mutations and coding mutations known to be implicated in cancer. Accounting for non-coding mutations will be necessary to fully understand the heterogeneity that exists both within and between tumours. Furthermore, the identification of driver mutations may be more effective if both coding and non-coding mutations are studied together in the context of known biological pathways. For example, a recent analysis of non-coding mutations in B-cell lymphoma [67] used pathway analysis to identify pathways enriched for genes with promoter mutations, and employed a combined analysis of both coding and regulatory mutations to identify frequently altered genes and to assess the effect of these alterations on gene expression of the target gene, as well as network neighbours.

Each method for the identification of non-coding drivers has distinct benefits and disadvantages. Recurrence based methods are very similar to methods that have already been applied to coding regions, but may need to be adapted to work optimally in non-coding regions. The high mutation rate and repetitive nature of the non-coding genome requires that issues such as high background mutation rate and potential mapping errors are controlled. Using information other than simple recurrence, such as evolutionary, functional genomics, network, and motif information may help to highlight interesting variants or regions. FunSeq2 [63] is an example of a method that uses the vast array of available genomic data, such as ENCODE [58] Roadmap Epigenomics [68] position weight matrices, and network information to attempt to prioritise variants. Conversely, the requirement for additional data may limit the samples that can be used for analysis. For example, correlating mutations with RNA-seq based expression limits analysis only to samples with available RNA-seq data.

4.6 Objective of this study

My objective was to develop a scoring system that can prioritise regions of the genome, particularly non-coding regions, in terms of their potential to act as driver mutations in cancer. To achieve this, I decided to use whole genome somatic mutation data to identify recurrently mutated regions, as well as regions that have mutations at conserved nucleotides. When developing this scoring method, I focussed on designing a recurrence score that is able to account for the variability in mutation rate across the genome. To do this, I chose to use a method that normalizes mutation rates in regions of interest by observed mutation rates in flanking regions. I then validated the scores that I developed on exonic regions, using information on genes that are known to be frequently mutated in cancer. In addition to a pan-cancer analysis, I also sought to apply my scoring system in a cancer type specific manner. I applied my scoring system to somatic mutations from over 1300 whole cancer genomes to identify known drivers, both coding and non-coding, as well as some promising novel candidates.

5 Methods

In order to identify recurrently mutated non-coding regions that are potential targets of somatic selection during the development of cancer, I devised a scoring system to prioritise regions of the genome based on signatures that are indicative of selection. In the context of coding mutations, driver genes are known to be recurrently mutated above background mutation rates and also show a pattern of enrichment for functional mutations (e.g. stop-gain, non-synoymous) compared to mutations that are less likely to be function (e.g. synonymous mutations). Applying these same principles to non-coding regions, I developed two scores, one that is designed to detect regions that are recurrently mutated, and a second designed to detect regions that have mutations at conserved bases, working on the hypothesis that conserved positions are more likely to be functional. I then applied these scores, as well as a combined score, to a set of over 1300 cancer whole genomes.

5.1 Whole genome mutation data

I assembled a set of pre-called somatic mutations from three sources: release 18 of ICGC [69], data from Alexandrov *et al.* [7], and the supplemental materials of *Wang et al.* [70]. Some of these sources contain data from both whole exome and whole genome sequencing. I only analyzed mutations annotated as coming from whole genome sequencing. To avoid the possibility of duplicated samples, in cases where the same tumour type was included in ICGC and the data from Alexandrov *et al.* I included data from only one source. The distribution of samples across tumour types and data sources is summarized in Table 1. After filtering out samples lacking sufficient numbers of mutations, I was left with a total of 1349 samples for my final analysis.

Table 1: The number of samples with 1000 or more valid mutations included in my final analysis, as well as information about tumour type and original publication for each sample. For the ICGC samples I give ICGC project codes and use this to categorise tumour type throughout this work. Although some project codes imply the same tumour type (e.g. LICA-FR and LINC-JP are both liver cancers) I treat these separately in case these cohorts might have different properties, either technical or biological.

Source	Cancer type	Cancer cohort	Number of samples
Alexandrov *et al.*	Acute lymphoblastic leukemia (ALL)	Acute lymphoblastic leukemia	1
ICGC	Bone	BOCA-FR	3
Alexandrov *et al.*	Breast	Breast	116
ICGC	Chronic lymphocytic leukemia (CLL)	Chronic lymphocytic leukemia	21
ICGC	Prostate	EOPC-DE	9
ICGC	Esophageal	ESAD-UK	97
Wang *et al.*	Gastric	Gastric	98
ICGC	Liver	LICA-FR	5
ICGC	Liver	LINC-JP	31
ICGC	Liver	LIRI-JP	238
Alexandrov et al.	Lung	Lung	24
ICGC	Lymphoma	MALY-DE	44
Alexandrov *et al.*	Medulloblastoma	Medulloblastoma	42
ICGC	Ovarian	OV-AU	75
ICGC	Pancreatic	PACA-AU	148
ICGC	Pancreatic	PACA-CA	151
ICGC	Pancreatic	PACA-IT	29
ICGC	Pancreatic	PAEN-AU	37
ICGC	Prostate	PRAD-CA	89
ICGC	Renal	RECA-EU	88
ICGC	Thyroid	THCA-SA	3

5.2 Annotation data

I used the UCSC genome browser [71, 72] to obtain various annotation files, including dbSNP and COSMIC variants, information on gene models, conservation, mappability, and epigenetic data.

5.3 Software

I processed genomic data using bedtools v2.25.0 [73] and conducted statistical analysis and data manipulation in R 3.2.3 [74].

5.4 Processing mutation data

I mapped all data to hg19. Preliminary analysis revealed several frequent mutations that overlap known germline SNPs, suggestive of the possibility that these mutations are not truly somatic. I removed from consideration mutations that occur at the same genomic coordinate as a known dbSNP entry, unless that genomic position was also annotated as mutated in COSMIC (cancer.sanger.ac.uk) [75]. After filtering out known dbSNP entries, I also excluded mutations from individual tumour samples with fewer than 1000 total mutations. For dbSNP variants, I used build 142 of dbSNP. dbSNP and COSMIC variant locations were obtained in bed format from the UCSC Table Browser [71].

5.5 Annotating and filtering genomic regions

I divided the reference hg19 genome into 50bp, non-overlapping windows using the bedtools makewindows command. I mapped mutations to each window, and calculated the mean 100-way PhyloP score as well as the mean 35bp uniqueness (a measure of sequence mappability) across mutations that fell within the window. I excluded from further consideration any window that had a mean mappability of its overlapping mutations that was less than 0.5, as well as any window that was mutated in fewer than 3 patients (because these regions lack sufficient mutations to be considered recurrent).

5.6 Calculation of recurrence score

Selection for driver mutations may cause genomic regions to have large numbers of mutations, as a result for selection for these mutations. To detect such recurrently mutated regions, I developed a recurrence score which quantifies regional mutation frequencies. For each 50bp region that met my filtering criteria (candidate regions), I calculated a recurrence score representing the level of enrichment of the region with mutations compared to the mutation rate within the region of the genome flanking the region under consideration. For each candidate region, I formed a flanking region (Figure 4), which included the region of the genome that was within 0.5 Mb of the 50bp candidate region on either side, truncated at chromosome ends. I removed bases within the flanking region that had mappability less than 0.5. I calculated a flanking mutation rate for each candidate region by dividing the number of mutations in my set of whole genomes that overlap valid flanking base positions by the number of valid bases within the flanking region. I calculated a raw mutation score (Equation 1) by dividing the rate (mutations per nucleotide) in the candidate region by the flanking mutation rate. I normalized this raw mutation score by subtracting the median score from all candidate regions and dividing each score by the median absolute deviation (mad) over all candidates (Equation 2). I initially planned to perform the normalization by flanking mutation rate separately for each tumour sample, but this was not feasible due to the sparsity of mutations in some samples. Equations for the raw and normalized recurrence scores are:

$$raw\ score = \frac{T/T_0}{(L+R)/(L_0+R_0)} \qquad \text{Equation 1}$$

Where T is the number of mutations observed in the target region, T_0 is the length of the target region, L and R are the number of mutations in the left and right flanking regions of the target region, and L_0 and R_0 are the lengths of the left and right flanking regions.

$$normalized\ score = \frac{raw\ score - median(raw\ score)}{mad(raw\ score)} \qquad \text{Equation 2}$$

5.7 Calculation of conservation score

For each candidate region, I also calculated a conservation score. My strategy was to use a basepair level measure of conservation, and average across mutations to score a region based on conservation. I chose the PhyloP score [76] calculated on a 100-way species tree, which is available from the UCSC genome browser. PhyloP scores as implemented in the UCSC Genome Browser are negative log base 10 p-values for a likelihood ratio test against the null hypothesis of neutral evolution. These scores are calculated at the nucleotide level and are assigned a sign based on the direction of observed depature from a neutral (unselected) evolutionary model. The scores are positive when the test indicates that the nucleotide evolves more slowly (i.e. is conserved) and negative in the case that it evolves more quickly (acceleration). For each mutation, I mapped PhyloP scores of the base position at which the mutation occurred. Within each 50bp candidate region, I took the mean of the PhyloP scores for each mutation within the region as a raw conservation score. Similar to my recurrence score, I normalized this raw conservation score by subtracting the median score and dividing by the median absolute deviation.

5.8 Calculation of combined score

For each candiate region, I calculated the combined score as the simple average of normalized recurrence and conservation scores.

5.9 Statistical analysis

For comparison of scores in different classes of regions, I used Mann-Whitney tests, as implemented in R. I also performed simulations to compare the median scores of known driver regions to non-driver exonic regions. I repeated sampled with replacement 10,000 samples of non-driver regions with size equal to the number of candidate regions overlapping known driver regions, took the median score for each sample, and compared to the observed median for known driver genes.

5.10 Identification of known driver genes

Driver genes were identified in humans by combining gene lists from two previously published lists of driver genes from Vogelstein *et al.* and Lawrence *et al.* [5, 30]. Gene names were taken from table S2A of Vogelstein *et al.* [5] and from supplemental table 2 from Lawrence *et al.* [30]. These gene names were entered into the UCSC Table Browser [71] to obtain hg19 coordinates for the coding exons of these genes, which were mapped to mutations using bedtools [73]. I considered a region to be a known driver if it overlapped a coding exon of a gene listed in either publication. In total, I constructed a set of 308 driver genes by this method.

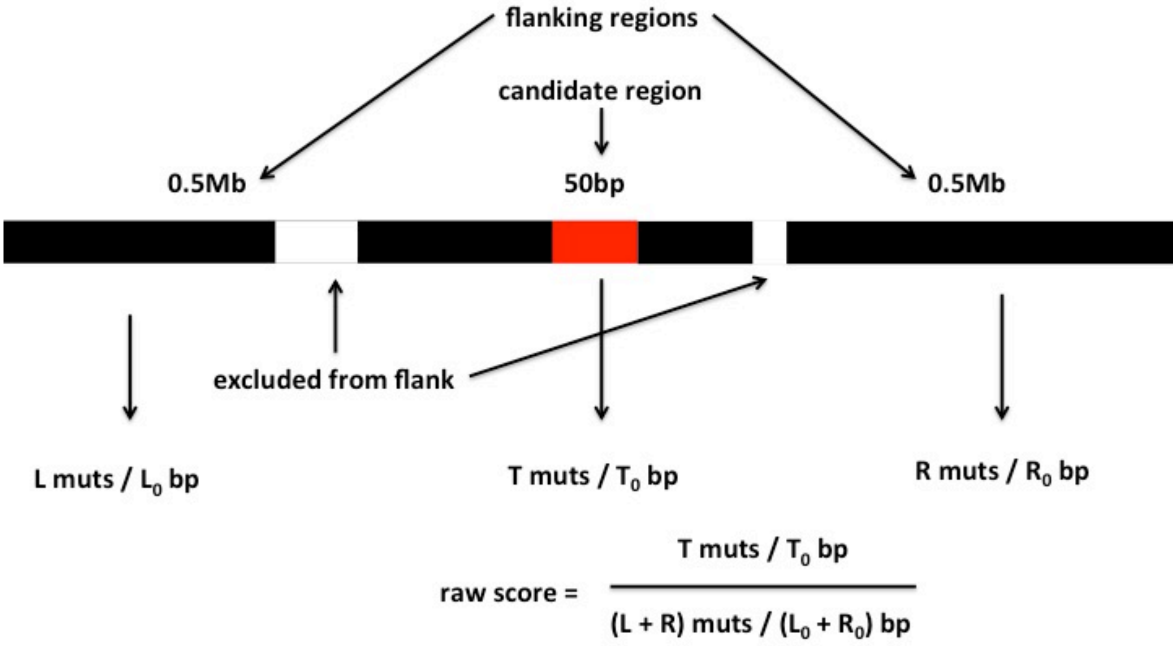

Figure 4: For each candidate region (red) I identified a flanking region (black) that encompassed 0.5 Mb on either side, excluding regions of low mappability (white). I calculated the mutation rate (mutations per base) seperately for the candidate region and the flanking region. To calculate the rates, I obtained mutation counts L, R, and T of mutations overlaping upstream flanks, downstream flanks, and the target region, respectively. I then divided the mutation counts by the lengths in basepairs L_0, R_0, and T_0. The raw score is then the ratio of candidate rate to flanking rate, implying that recurrent regions should have a high raw score.

6 Results

I have developed a set of scores, described in sections 5.6-5.8, that identify regions of the genome that are more frequently mutated compared to flanking regions (recurrence score) and that have mutations at bases that are more highly conserved (conservation score). I have calculated these scores based on 1349 whole cancer genomes from a variety of cancer types for 50bp windows spanning the entire human genome. Unlike previous efforts aimed at identifying non-coding driver mutations, which have usally focussed on a limited set of non-coding regions (e.g. promoters, DNase I hypersensitive sites) I have applied my method in an unbiased manner to the entire genome, with the sole exception of regions where mappability is a concern. Here, I examine the characteristics and performance of my scores, as well as highlight some promising top scoring regions.

6.1 Mutational processes in cancer whole genomes

My objective was to identify regions of the non-coding genome that are under positive selection during tumourogenesis. I searched for regions of the genome that are recurrently somatically mutated in cancer, a signal of positive selection. Although recurrent mutation may be a result of selection, it may also result from mutational processes acting on cancer genomes. There is considerable heterogeneity in mutation rates between different regions of the genome as well as between different tumours (Figure 5). To discover regions that are mutated more than would be expected from simple mutational processes, I implemented a score that normalized for the mutation rate in flanking regions. This method can account for mutational processes that are constant over large portions of the genome, but may falsely identify portions of the genome that are particularly susceptible to mutation within a focussed region. Because of the possibility that such focal mutational processes might contaminate regions identified by my scoring method, I additionally sought to understand mutational processes acting

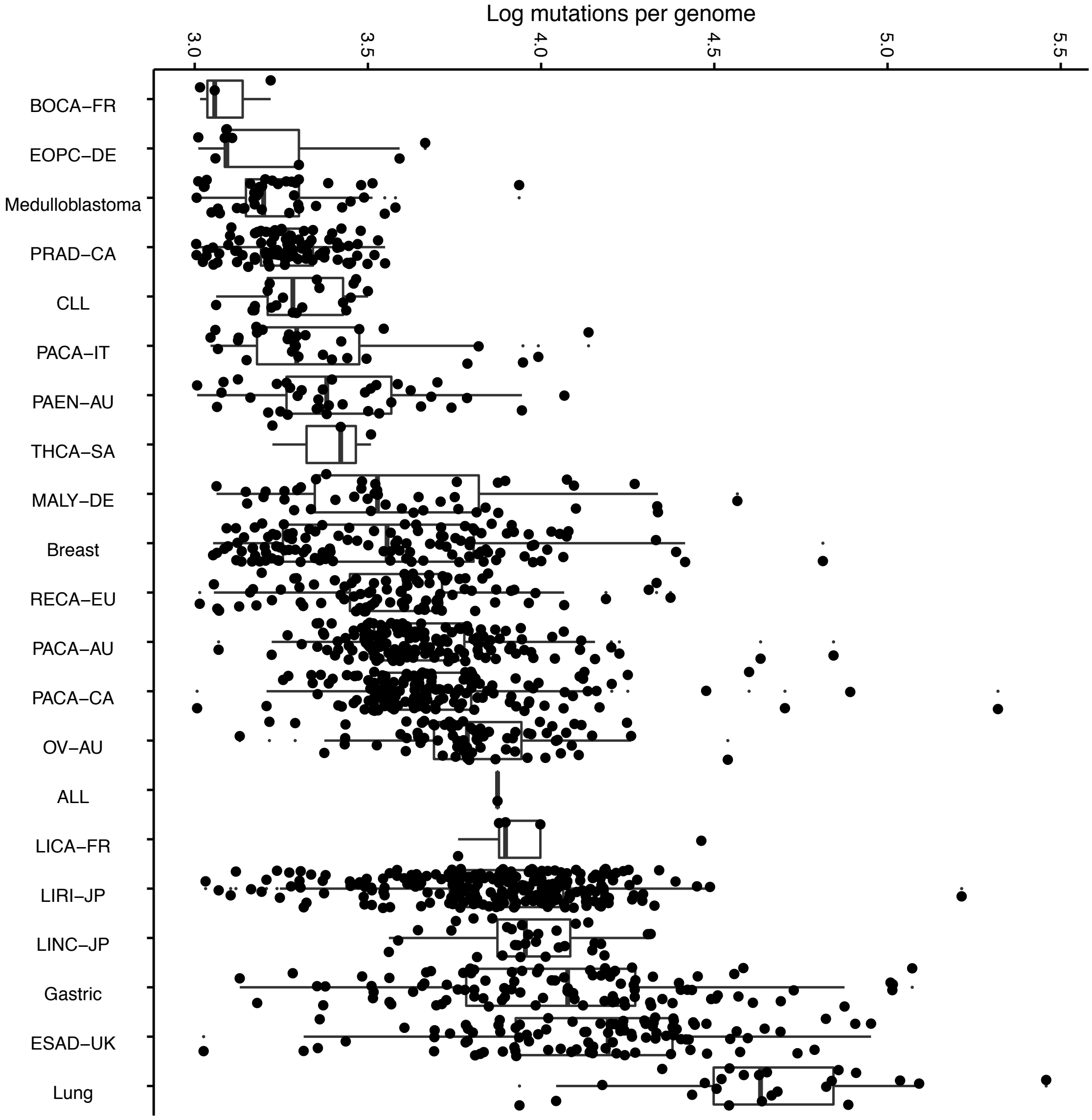

Figure 5: Log10 of total mutations per genome, ordered by median mutations within each tumour type. There is considerable variation both within and between cancer types. The most highly mutated cancer types are generally associated with intense exposure to known mutational process, such as tobacco smoke in lung cancer and DNA repair defects in gastric cancer.

on whole cancer genomes for the purpose of flagging regions that are potentially false positives. To distinquish between the potential causes of recurrent mutation, I refer to regions as “putatively hypermutated” to suggest that they may be mutated due to exposure to mutational processes, as opposed to selection.

6.2 Identification of putative hypermutated regions

I reasoned that regions of the genome with unusually high exposure to mutational processes would be expected to have a consistently elevated likelihood of mutation, whereas selection is expected to diminish once a driver has already been mutated. For example, gain of function mutations in oncogenes generally only need to occur once to confer driver activity, and often display mutual exclusivity with other mutations that have the same effects or that target the same pathway. This can be explained by a substantial decrease in selection pressure once an activating mutation has already occurred. Tumour suppressor genes are an exception, where two mutations may be required to confer driver activity. Thus, regions that are recurrently mutated due to mutational processes are more likely to sustain repeated mutations within the same region in the same tumour, while regions that are recurrently mutated due to selection are more likely to be mutated only once per tumour. In order to identify regions that may be recurrently mutated due to mutational processes rather than selection, I calculated the average number of mutations per patient for each region under consideration. I considered a region to be potentially hypermutated when the region had an average of 1.2 mutations per patient or greater. I examined the prevalence of mutations within these putative hypermutated regions across tumour types. Several tumour types have an excess of mutations from hypermutated regions (Figure 6) such as lymphomas (“MALY-DE”) and renal cancers (“RECA-EU”). Several of the regions that I have identified as being hypermutated by this methods lie in promoter regions and are primarily mutated in lymphoma, potentially suggesting that these regions are targets of somatic hypermutation rather than selection. Several of these regions such as the promoter regions of *BCL2* and *MYC* have been identified as putative targets of selection in a previous analysis [65]. Analysis of mutational signatures within the putatively hypermutated regions that I identified did not identify any specific mutation process that could expain the pattern of base subsitutions in these regions (Figure 7), although it is possible that this mutational pattern is partially due to a process identified in CLL and lymphoma that is associated with somatic hyper mutation [7].

6.3 Mutational processes at CTCF binding sites

In addition to the putatively hypermutated regions that I identified, I also observed that many recurrently mutated regions overlap regions with ChIP-seq evidence of CTCF binding (Figure 8 panel A, CTCF binding vs other regions p = 3.8 x 10^{-18}, CTCF DNase I hypersensitive vs other regions p = 2.08 x 10^{-263}, CTCF binding vs CTCF DNase I hypersensitive p = 1.24 x 10^{-46}). A recent analysis also identified an association between CTCF binding and recurrent mutation [59] potentially suggesting selection of these mutations, while other evidence from colorectal cancer by Katainen *et al.* suggests that CTCF binding sites may be subject to a unique mutational process which displays an excess of T>G (A>C) and T>C (A>G) mutations [23]. To discern whether the observed recurrence at CTCF binding sites in my dataset could result from a mutational process rather than selection, I compared the mutations at CTCF binding sites with the signature observed in Katainen *et al.* [23]. While CTCF binding sites in general do not show a signature similar to the one in [23] CTCF binding sites that I aso identified as recurrent in my analysis display an excess of T>G and T>C mutations (Figure 9). When I examined specific recurrently mutated CTCF binding site that was also identified in [23] I found that the same bases within the binding site were recurrently mutated (Figure 10, compare to Figure 3 in [23]). This suggests that the recurrently mutated CTCF binding sites identified by my analysis are likely the result of the same process implicated in Katainen *et al.* [23]. Recent analyses [77, 78] have shown that transcription factor bound regions of the genome are subject to unique mutational processes and these mutations often preferentially target certain bases (e.g. G/C bases). The recurrence score correlates slightly with GC context (rank correlation 0.113) perhaps due to coding driver genes having high

GC% (Figure 8, panel B). Regions with recurrence score > 10 have comparable GC% to regions with score < 10 (Wilcoxon rank sum p-value = 0.81). Mutations can also be more prevalent at some dinucleotide combinations. In particular, C>T (G>A) mutations are prevalent at CpG dinucleotides [17] which produces at detectable mutational signature [7]. Some regions may appear to have large numbers of mutations due to this mutation process, rather than selection. Although I can not completely rule out this possibility, Figure 7, which plots mutational contributions by 5' and 3' base, does not suggest an excess of CpG mutations, including within hypermutated and recurrent hypermutated regions.

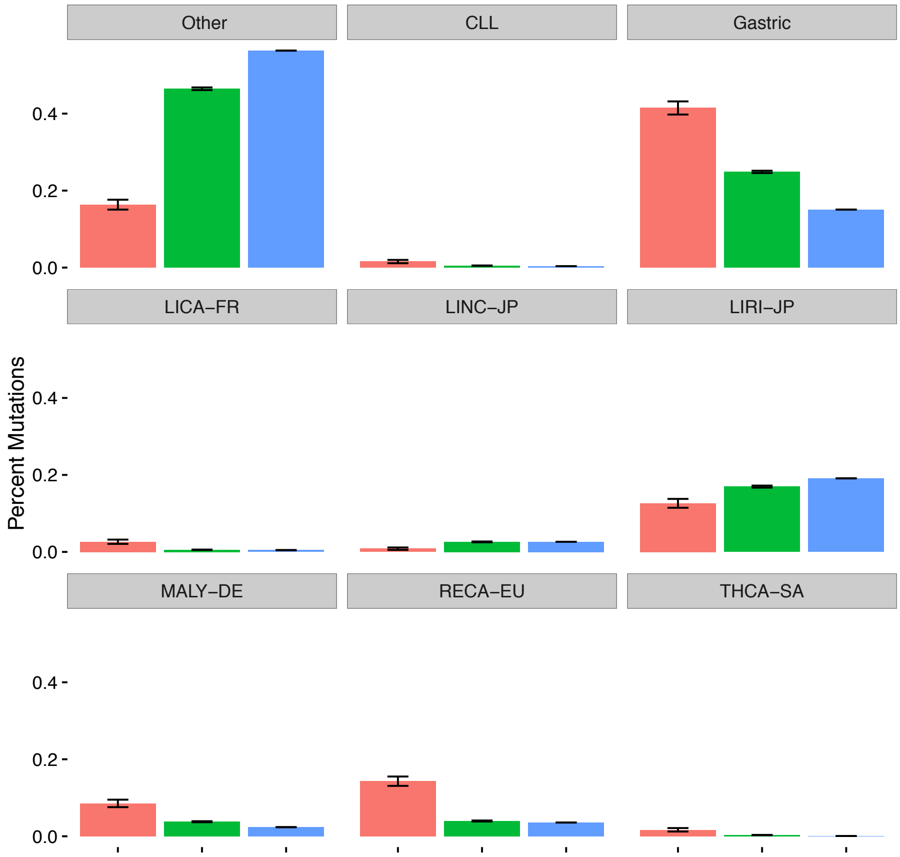

Figure 6: For each of three categories: recurrent and hyper mutated regions (red, 832 total mutations), non-recurrent hypermutated regions (green, 20958 total mutations), and other regions (blue, 10713694 total mutations), I give the percent of mutations within region that belong to different cancer types. Malignant lymphoma has a disproportionate share of hypermutated regions, suggesting that my method of identifying hypermutated regions is capturing some regions that are targets of somatic hypermutation in this cohort (this is also consistent with my observation that many apparently recurrent regions are only mutated in this cohort). Other cancer types also display an excess of hypermutation including renal cancer (RECA-EU) and gastric cancer, suggesting focal mutational processes in these cancer types. I define a region to be hypermutated when it has > 1.2 mutations per tumour, and to be recurrently mutated when it has a recurrence score greater than 10.

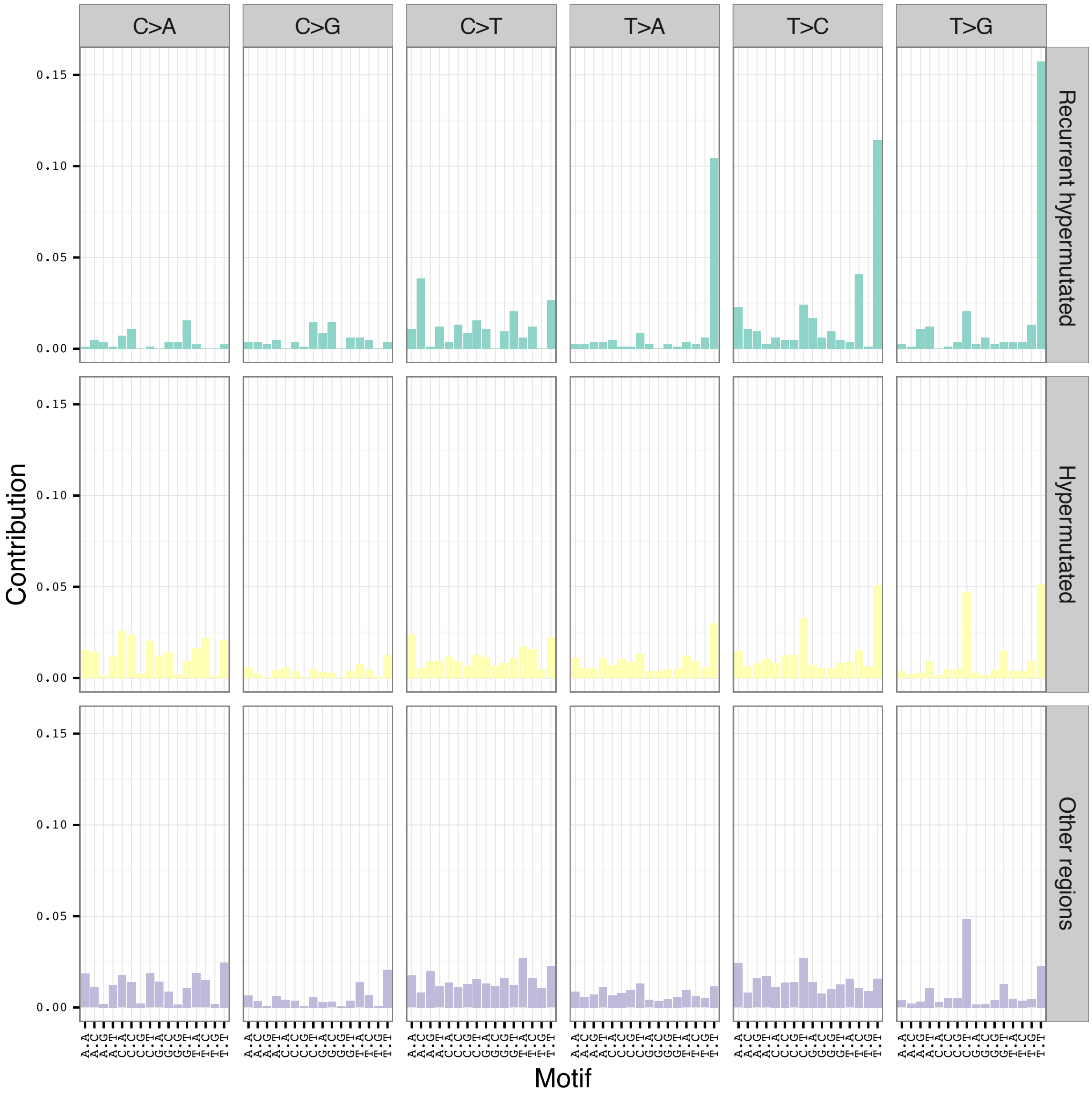

Figure 7: Observed mutational spectra within recurrent hypermutated, non-recurrent hypermutated, and non-hypermutated regions. Each column represents a particular category of mutation, defined by the base change, as well as the bases that flank the mutated nucleotide, both 5' and 3'. The height of each bar is proportional to the frequency of the mutational category within each region type.

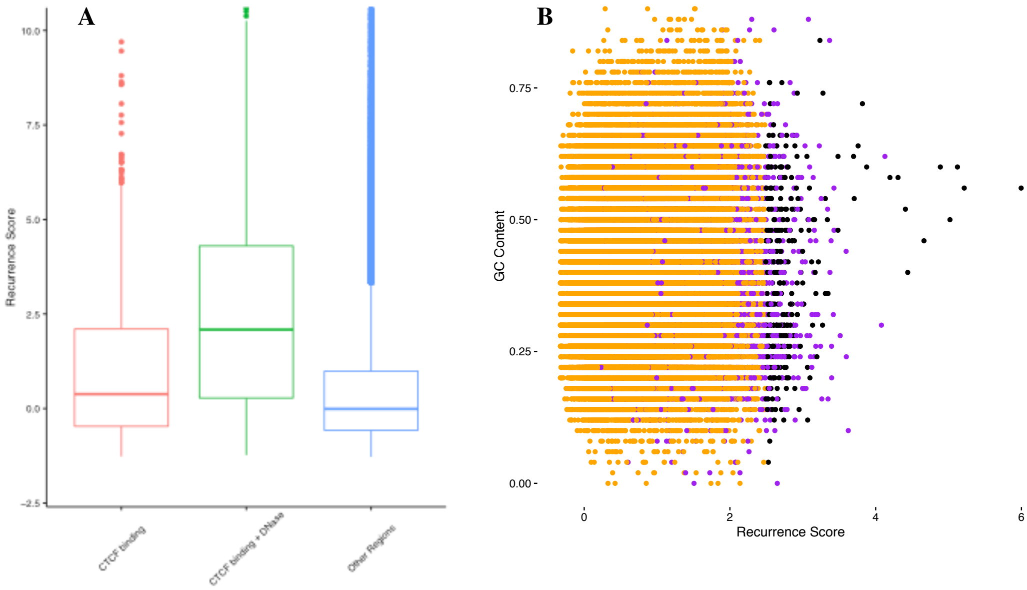

Figure 8: In panel A, I show CTCF binding sites that overlap Dnase I hypersensitive sites (green) as well as non-hypersensitive CTCF binding sites (red) show a higher recurrence score compared to non-CTCF binding regions (blue). In panel B, I show recurrence score (plotted as log(score + 2)) plotted against GC content. Regions with mutations per patient > 1.2 are in orange, with recurrence score > 10 and mutations per patient <= 1.2 in black, and all others in purple.

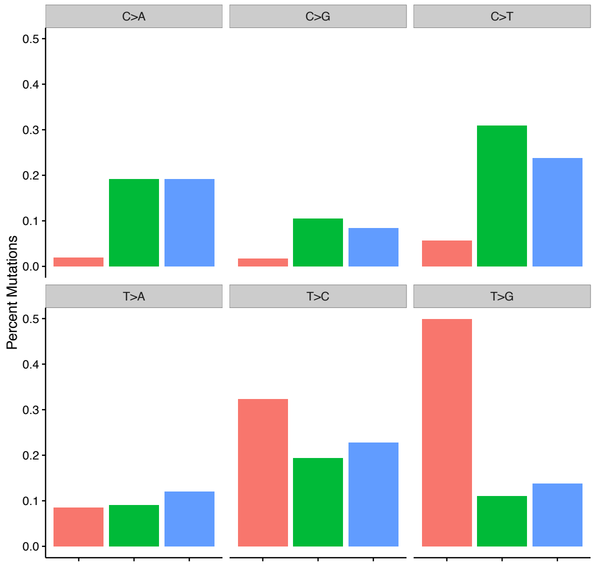

Figure 9: I classified mutations as coming from recurrent CTCF binding sites (red) non-recurrent CTCF binding sites (green) and non-CTCF binding sites (blue). For each of these three categories, I give percentages indicating how many mutations from each category exhibit each of the six possible base changes. Although non-recurrent CTCF binding sites appear similar to non-CTCF binding sites in terms base change patterns, recurrent CTCF binding sites in my data show a base change pattern that matches a signature previously observed at CTCF binding sites in colorectal cancer. I define a CTCF binding site are recurrent when it has a recurrence score greater than 10.

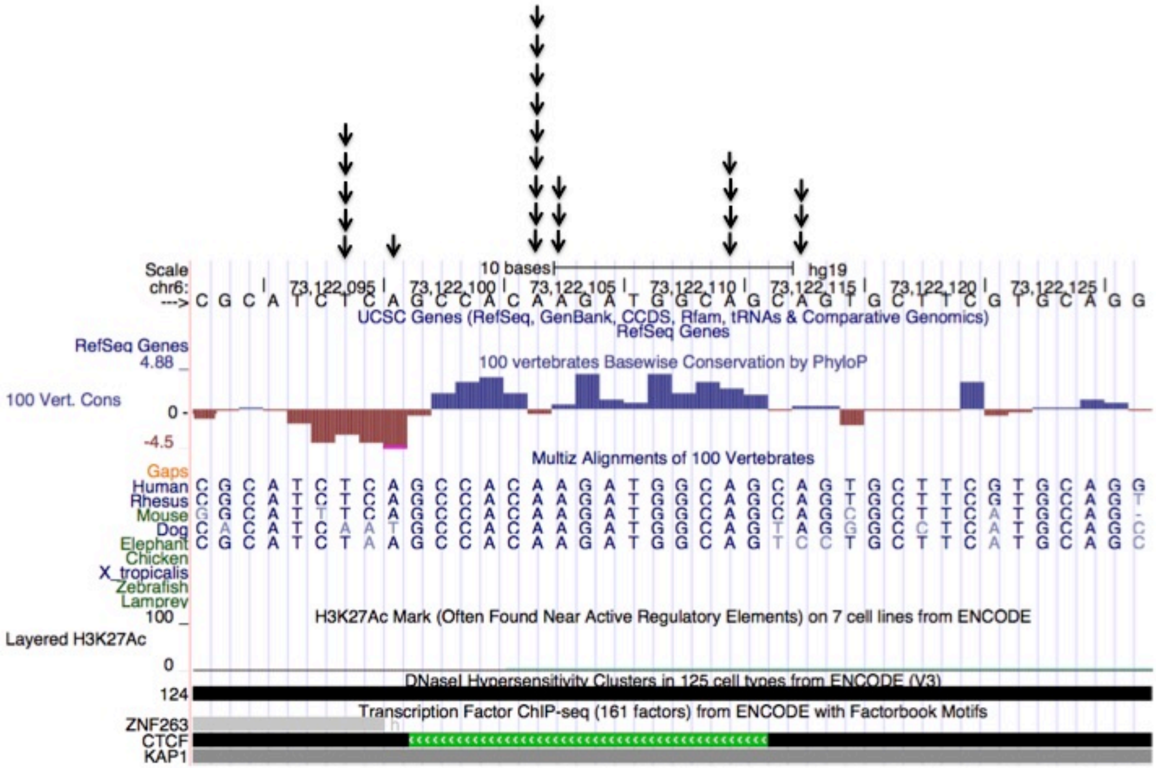
Scale
chr6:
10 bases
hg19
73,122,095
73,122,100
73,122,105
73,122,110
73,122,115
73,122,120
73,122,125
---> C G C A T C T C A G C C A C A A G A T G G C A G C A G T G C T T C G T G C A G G
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
RefSeq Genes
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
124
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
ZNF263
CTCF
KAP1

Figure 10: For comparision, I show the location of mutations (black arrows) within a recurrent CTCF binding site that was highlighted in a previous analysis [23]. The mutation positions and nucleotide changes observed in my sample match those identified in this previous analysis, despite the fact that my dataset lacks any colorectal cancer samples.

6.4 Pan-cancer prioritisation of non-coding mutations

Having identified CTCF binding sites and regions with >1.2 mutation per tumour as regions that might be enriched for false positives, I next sought to identify regions that were likely to be under selection. I validated my prioritisation scores by considering exonic regions within my sample, because many large analyses have already identified known driver genes in protein coding regions. My recurrence score ($p = 3.8 \times 10^{-27}$), conservation score ($p = 1.32 \times 10^{-19}$), and combined score ($p = 3.22 \times 10^{-30}$) were able to discriminate known driver genes within the set of all exonic regions (Figure 11), suggesting that my method has reasonable effectiveness within this subset of the genome, despite the fact that I did not take advantage of annotations that are available for coding mutations (e.g. non-synonymous vs synonymous mutations). I confirmed this by direct comparison of scores between driver and non-driver regions, as well as by simulation. To compare the known driver regions to a set of non-drivers of equal size, I resampled the non-driver exonic regions 10,000 times for each score, and compared the median score of the sampled non-drivers to the observed median of the known drivers. For all three scores, none of the 10000 samples exceeded the median driver score (Figure 11). Several of the top scoring coding regions overlap well known driver genes such as *TP53* and *KRAS*. To check whether the inclusion of coding sequence within flanking regions had an impact on the regions identified, I also rescored each candidate region, this time excluding coding regions from the calculation of the flanking mutation rate. The regions identified were largely similar, with 94% of top regions in common between the two scoring methods. In order to assess whether the mutational counts are dominated by hypermutated samples, I recalculated the number of mutations in each 50bp window, excluding samples that are two standard deviations above the mean number of mutations. These counts are highly correlated ($r = 0.88$, $p < 0.0001$) and this correlation is maintained when considering only regions that have greater than 5 mutations in the full dataset ($r = 0.937$, $p < 0.00001$).

In addition to identifying known coding drivers, I also identified recurrently mutated non-coding regions, including both previously identifed regions as well as novel regions (Figure 12, Tables 2-5). I identifed *TERT* (Figure 13) and *PLEKHS1* (Figure 14) [9] promoters as being recurrently mutated, consistent with previous analyses. *TERT* appears in the top 50 regions genome-wide by recurrence (Table 2) but not when ranked by the combined score (Table 4). One explanation for this is that in a genome-wide context, adding conservation will tend prioritise coding regions more highly, given the higher conservation of coding compared to non-coding regions. In support of this interpretation, Table 4 appears to be enriched for coding drivers relative to Table 2, while comparison of the top ten non-coding, non-hypermutated regions based on recurrence (Table 3) and combined score (Table 5) are highly similar. Despite the similarity of these lists, adding conservation does bring some interesting regions into the top ten, including an intronic region that shows high conservation, as well as a conserved region of a miRNA. I discuss both of these regions in more detail in the next section.

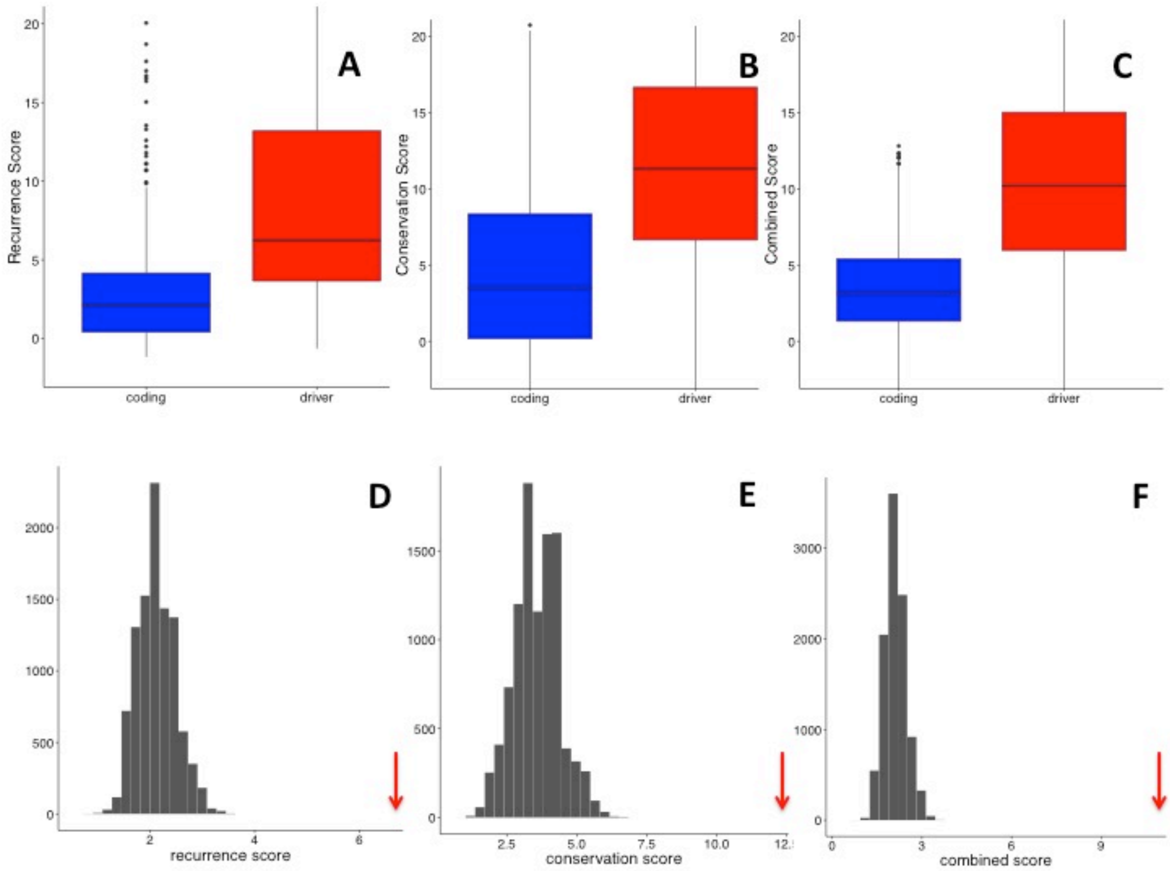

Figure 11: For exonic regions, known driver genes score significantly higher in terms of recurrence (**A**,**D**) conservation (**B**,**E**) and combined scores (**C**,**F**). I also compare the observed medians scores for drivers (red arrows) to control medians generated by resampling non-driver regions (grey bars, **D-E**).

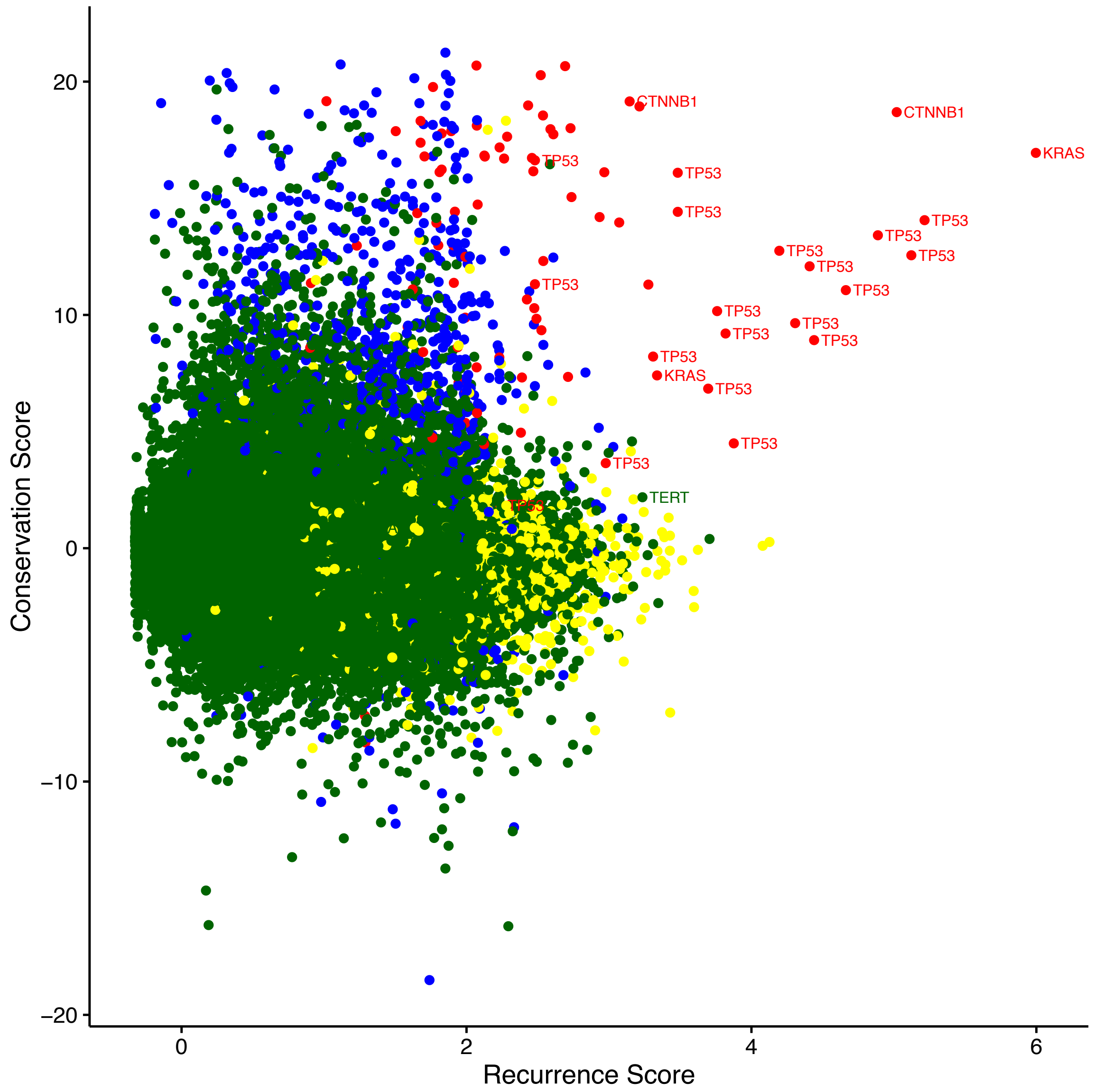

Figure 12: Scatterplot of all regions mutated in more than two patients with conservation score on the vertical axis and Log(recurrence score + 2) on the horizontal axis. The points are colored based on a classification of each region into one of four categories: coding, non-driver regions (blue), coding driver regions (red), non-coding, hypermutated regions (yellow), and non-coding non-hypermutated regions (green). Several known driver regions are also labelled.

Table 2: Top 50 regions in terms of recurrence score identified by my method. I give the position of the region, number of genomes that are mutated within the region, the recurrence score, and a classification of the region based annotations and my method of identifying hypermutated regions. I also manually annotated each region by viewing in the UCSC genome browser.

rank	chr	Start	end	Mutated samples	score	Automated annotation	Manual annotation
1	chr12	25398250	25398300	256	399.9	Driver	*KRAS* exon
2	chr17	7577100	7577150	68	182.1	Driver	*TP53* exon
3	chr17	7577500	7577550	62	165.7	Driver	*TP53* exon
4	chr3	41266100	41266150	65	149.3	Driver	*CTNNB1* exon
5	chr17	7578400	7578450	50	130.6	Driver	*TP53* exon
6	chr17	7577550	7577600	41	103.9	Driver	*TP53* exon
7	chr17	7578200	7578250	32	82.8	Driver	*TP53* exon
8	chr17	7578250	7578300	31	80.1	Driver	*TP53* exon
9	chr17	7577050	7577100	29	72.2	Driver	*TP53* exon
10	chr17	7578500	7578550	26	64.4	driver	*TP53* exon
11	chr10	96652800	96652850	14	60.0	hotspot	non-coding
12	chr12	6899300	6899350	3	57.1	hotspot	*CD4* intron
13	chr17	7574000	7574050	19	46.2	driver	*TP53* exon
14	chr17	7578450	7578500	18	43.5	driver	*TP53* exon
15	chr17	7578350	7578400	17	40.9	driver	*TP53* exon
16	chr3	195892250	195892300	18	38.7	non-coding	non-coding
17	chr17	7577000	7577050	14	38.3	driver	*TP53* exon
18	chr12	64749950	64750000	7	35.5	hotspot	*C12orf56* intron
19	chr13	50016900	50016950	8	34.5	hotspot	*CAB39L* intron
20	chr11	63881800	63881850	9	34.4	hotspot	*FLRT1* intron
21	chr15	64857000	64857050	9	31.6	hotspot	*ZNF609* intron
22	chr17	7578150	7578200	13	30.6	driver	*TP53* exon
23	chr17	7578550	7578600	13	30.5	driver	*TP53* splice site
24	chr16	88383450	88383500	7	28.9	hotspot	Non-coding / TF binding
25	chr14	24895100	24895150	11	28.8	hotspot	Non-coding / TF binding
26	chr17	79389900	79389950	9	28.8	hotspot	*BAHCC1* intron
27	chr17	17424850	17424900	7	28.5	hotspot	*PEMT* intron
28	chr22	46697350	46697400	5	27.8	hotspot	*GTSE1* intron
29	chr8	30717550	30717600	7	27.8	hotspot	*TEX15* exon-intron border
30	chr7	76949650	76949700	6	27.6	hotspot	*GSAP* intron
31	chr14	74239050	74239100	8	27.2	hotspot	*ELMSAN1* intron
32	chr4	819750	819800	6	27.0	hotspot	*CPLX1* intron
33	chr16	81908550	81908600	7	26.4	hotspot	*PLCG2* intron
34	chr4	39684550	39684600	10	26.4	non-coding	non-coding
35	chr22	39962000	39962050	6	26.2	hotspot	non-coding
36	chr12	25380250	25380300	20	26.1	driver	*KRAS* exon
37	chr3	43746400	43746450	11	25.4	non-coding	*ABHD5* intron
38	chr17	7579300	7579350	10	25.4	driver	*TP53* exon
39	chr9	21971100	21971150	12	24.5	driver	*CDKN2A* exon
40	chr8	9921850	9921900	12	24.3	non-coding	*MRSA* intron
41	chr11	70764100	70764150	6	24.1	hotspot	*SHANK2* intron
42	chr19	12597300	12597350	9	23.8	hotspot	*ZNF709* intron
43	chr17	49455750	49455800	10	23.6	hotspot	non-coding
44	chr5	1295200	1295250	14	23.4	non-coding	*TERT* promoter
45	chr7	151591800	151591850	6	23.2	hotspot	non-coding
46	chr21	44524450	44524500	9	22.9	driver	*U2AF1* exon
47	chr1	45914900	45914950	7	22.7	hotspot	*TESK2* intron
48	chr8	29901300	29901350	9	22.4	non-coding	non-coding
49	chr7	606050	606100	7	22.0	hotspot	*PRKAR1B* intron
50	chr2	49173750	49173800	27	22.0	non-coding	CTCF binding

Table 3: Top ten non-coding, non-hypermutated regions in terms of recurrence score.

rank	chr	start	end	samples mutated	score	manual annotation
1	chr3	195892250	195892300	18	38.7	non-coding
2	chr4	39684550	39684600	10	26.4	non-coding
3	chr3	43746400	43746450	11	25.4	*ABHD5* intron
4	chr8	9921850	9921900	12	24.3	*MSRA* intron
5	chr5	1295200	1295250	14	23.4	*TERT* promoter
6	chr8	29901300	29901350	9	22.4	non-coding
7	chr2	49173750	49173800	27	22.0	CTCF binding
8	chr8	70576150	70576200	21	21.8	CTCF binding
9	chr19	893450	893500	9	21.6	*MED16* promoter
10	chr2	47359300	47359350	8	21.0	*C2orf61* intron

Table 4: Top 50 regions in terms of combined score identified by my method. I give the position of the region, number of genomes that are mutated within the region, the combined score, and a classification of the region based annotations and my method of identifying hypermutated regions. I also manually annotated each region by viewing in the UCSC genome browser.

rank	chr	Start	End	Mutated samples	Score	Automated annotation	Manual annotation
1	chr12	25398250	25398300	256	208.4	driver	*KRAS* exon
2	chr17	7577100	7577150	68	98.1	driver	*TP53* exon
3	chr17	7577500	7577550	62	89.1	driver	*TP53* exon
4	chr3	41266100	41266150	65	84.0	driver	*CTNNB1* exon
5	chr17	7578400	7578450	50	72.0	driver	*TP53* exon
6	chr17	7577550	7577600	41	57.5	driver	*TP53* exon
7	chr17	7578250	7578300	31	46.1	driver	*TP53* exon
8	chr17	7578200	7578250	32	45.8	driver	*TP53* exon
9	chr17	7577050	7577100	29	40.9	driver	*TP53* exon
10	chr17	7578500	7578550	26	38.6	driver	*TP53* exon
11	chr10	96652800	96652850	14	30.1	hotspot	Non-coding
12	chr12	6899300	6899350	3	28.6	hotspot	*CD4* intron
13	chr17	7578450	7578500	18	26.4	driver	*TP53* exon
14	chr17	7578350	7578400	17	25.5	driver	*TP53* exon
15	chr17	7574000	7574050	19	25.4	driver	*TP53* exon
16	chr17	7578550	7578600	13	23.3	driver	*TP53* exon
17	chr17	7577000	7577050	14	22.6	driver	*TP53* exon
18	chr17	7578150	7578200	13	22.5	driver	*TP53* exon
19	chr21	44524450	44524500	9	20.9	driver	*TP53* exon
20	chr3	41266050	41266100	10	20.2	driver	*CTNNB1* exon
21	chr3	195892250	195892300	18	19.5	non-coding	Non-coding
22	chr9	21971100	21971150	12	17.9	driver	*CDKN2A* exon
23	chr12	64749950	64750000	7	17.7	hotspot	*C12orf56* intron
24	chr17	7579300	7579350	10	16.8	driver	*TP53* exon
25	chr2	198266800	198266850	9	16.8	driver	*SF3B1* exon
26	chr12	25380250	25380300	20	16.8	driver	*KRAS* exon
27	chr18	48591900	48591950	11	16.8	driver	*SMAD4* exon
28	chr3	178936050	178936100	9	16.7	driver	*PIK3CA* exon
29	chr11	63881800	63881850	9	16.3	hotspot	*FLRT1* intron
30	chr13	50016900	50016950	8	16.0	hotspot	*CAB39L* intron
31	chr19	11134250	11134300	6	15.7	driver	*SMARCA4* exon
32	chr15	64857000	64857050	9	15.5	hotspot	*ZNF609* intron
33	chr20	57484400	57484450	13	15.5	driver	*GNAS* exon
34	chr16	3786700	3786750	5	15.4	driver	*CREBBP* exon
35	chr17	17424850	17424900	7	14.9	hotspot	*PEMT* intron
36	chr14	24895100	24895150	11	14.7	hotspot	Non-coding / TF binding
37	chr18	48575150	48575200	7	14.7	driver	*SMAD4* exon
38	chr18	48604750	48604800	7	14.6	driver	*SMAD4* exon
39	chr19	11132500	11132550	5	14.6	driver	*SMARCA4* exon
40	chr17	79389900	79389950	9	14.3	hotspot	*BAHCC1* exon
41	chr18	48591800	48591850	8	14.2	driver	*SMAD4* exon
42	chr3	178952050	178952100	7	14.2	driver	*PIK3CA* exon
43	chr7	76949650	76949700	6	14.0	hotspot	*GSAP* intron
44	chr14	74239050	74239100	8	13.9	hotspot	*ELMSAN1* intron
45	chr17	56408600	56408650	5	13.9	non-coding	*MIR142* non-coding
46	chr22	46697350	46697400	5	13.6	hotspot	*GTSE1* intron
47	chr8	30717550	30717600	7	13.4	hotspot	*TEX15* exon-intron border
48	chr10	89692900	89692950	3	13.3	driver	*PTEN* exon
49	chr17	7577600	7577650	5	13.3	driver	*TP53* splice site
50	chr4	819750	819800	6	13.2	hotspot	*CPLX1* intron

Table 5: Top ten non-coding, non-hypermutated regions in terms of combined score.

rank	**chr**	**start**	**end**	**samples mutated**	**score**	**manual annotation**
1	chr3	195892250	195892300	18	38.7	non-coding
2	chr4	39684550	39684600	10	26.4	non-coding
3	chr3	43746400	43746450	11	25.4	*ABHD5* intron
4	chr8	9921850	9921900	12	24.3	*MSRA* intron
5	chr5	1295200	1295250	14	23.4	*TERT* promoter
6	chr8	29901300	29901350	9	22.4	non-coding
7	chr2	49173750	49173800	27	22.0	CTCF binding
8	chr19	893450	893500	9	21.6	*MED16* promoter
9	chr6	142706200	142706250	9	18.0	*GPR126* intron
10	chr17	56408600	56408650	5	11.3	*MIR142*

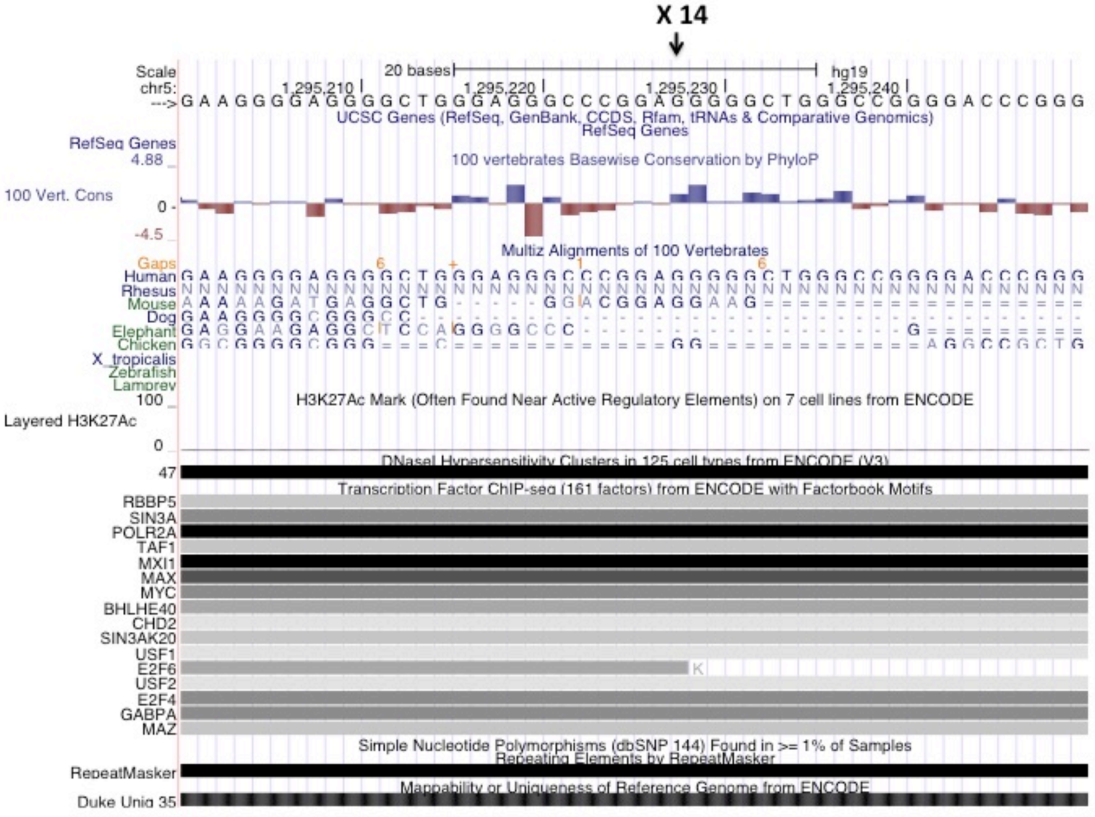
X 14
Scale
chr5:
20 bases
hg19
1,295,210
1,295,220
1,295,230
1,295,240
---> G A A G G G G A G G G G C T G G G A G G G C C C G G A G G G G G C T G G G C C G G G G A C C C G G G
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
RefSeq Genes
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
47
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
RBBP5
SIN3A
POLR2A
TAF1
MXI1
MAX
MYC
BHLHE40
CHD2
SIN3AK20
USF1
E2F6
USF2
E2F4
GABPA
MAZ
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
RepeatMasker
Mappability or Uniqueness of Reference Genome from ENCODE
Duke Uniq 35

Figure 13: Recurrent *TERT* promoter mutations identified in my data set. The mutations occur at one of the previously identified bases, generating a *de novo* ETS binding site.

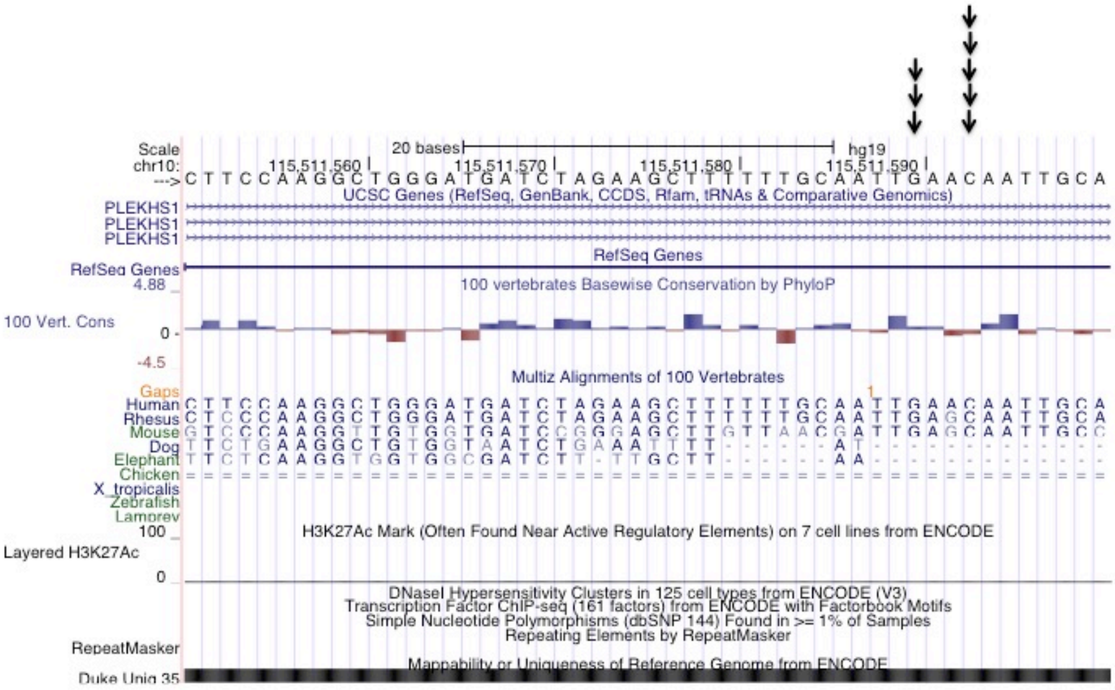
Scale
20 bases
hg19
chr10:
115,511,560
115,511,570
115,511,580
115,511,590
---> C T T C C A A G G C T G G G A T G A T C T A G A A G C T T T T T T G C A A T T G A A C A A T T G C A
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PLEKHS1
PLEKHS1
PLEKHS1
RefSeq Genes
RefSeq Genes
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
RepeatMasker
Mappability or Uniqueness of Reference Genome from ENCODE
Duke Uniq 35

Figure 14: PLEKHS1 recurrently mutated region that has previously been identified. I identifiy mutations at the same base position as previous analyses.

6.5 Novel recurrent non-coding mutations

My method has highlighted several novel non-coding regions that may be selected. Many highly recurrent regions are either known coding drivers or are regions that I have identified as hypermutated. Although a region can be both hypermutated and selected, I focus on highlighting regions that are less likely to hypermutated. To demonstrate the types of novel regions identified by my analysis, I examined two regions that scored among the top regions in terms of both recurrence and conservation scores in my pan-cancer analysis.

The first region that I examined (Figure 15) lies between the protein coding gene *MED16* and the small nuclear RNA *RNU6-2*. This regions lies within a Dnase I hypersentitivity site and shows heavy transcription factor binding, suggestive of promoter activity or some other regulatory function. Each mutation (black arrows) within the region lies within a conserved sub-region of the window. No mutations fall within the unconserved regions surrounding this sub-region or within the nearby RNA gene, despite the fact that these latter regions make up the majority of the window. Driver mutations often displaying clustering within specific functional regions. The pattern observed in this region, with mutations clustered within a single conserved element, is potentially suggestive of driver activity. Given the evidence for transcription factor binding in this region, one possibility is that this conserved sub-region is a motif associated with protein binding. Athough mutations at this locus are focussed within this conserved sub-region, the mutations are spread throughout the sub-region, not focussed at any single nucleotide, and do not always show consistent base changes in the cases where the mutations do occur at the same nucleotide. Assuming that these mutations are in fact targeting some kind of binding motif, the relatively even distribution of mutations without consistent base changes possibly suggests that these mutations are disrupting a binding motif as opposed to a creating a novel motif.

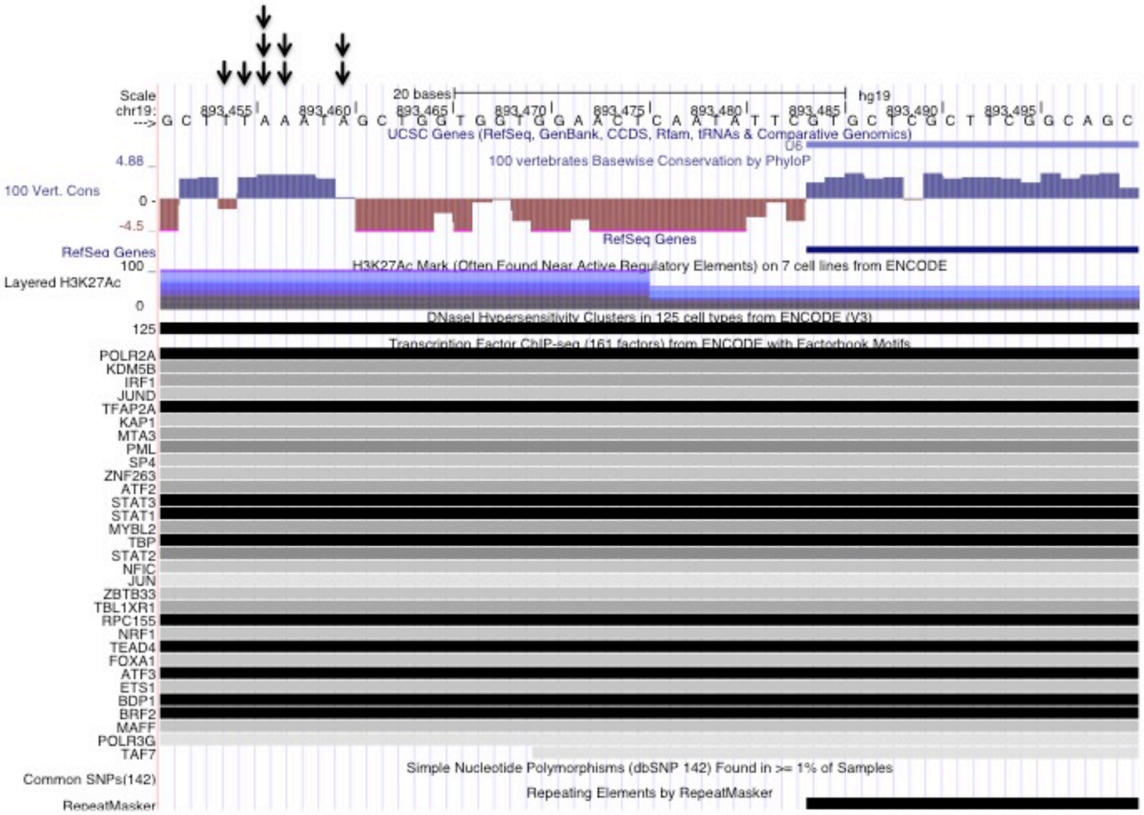
Scale
20 bases
hg19
chr19:
893,455
893,460
893,465
893,470
893,475
893,480
893,485
893,490
893,495
---> G C T T T A A A T A G C T G G T G G T G G A A C T C A A T A T T C G T G C T C G C T T C G G C A G C
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
D6
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
RefSeq Genes
RefSeq Genes
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
125
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
POLR2A
KDM5B
IRF1
JUND
TFAP2A
KAP1
MTA3
PML
SP4
ZNF263
ATF2
STAT3
STAT1
MYBL2
TBP
STAT2
NFIC
JUN
ZBTB33
TBL1XR1
RPC155
NRF1
TEAD4
FOXA1
ATF3
ETS1
BDP1
BRF2
MAFF
POLR3G
TAF7
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 15: UCSC browser image depiciting a recurrently mutated region identified by my method. Mutations are depicted by black arrows. This region is flanked on the left by the gene *MED16.* The mutations observed in this region are focussed within a conserved region overlapping an region of the genome with ENCODE evidence of transcription factor binding, possibly indicating selection.

The second region that I highlight (Figure 16) is deep within the intron of the gene *GPR126.* This region shows high levels of conservation, and the mutations observed is region occur exclusively at two base positions. All mutations within this region are entirely mutually exclusive, and there are no other mutations within thi region other than at these two positions. This pattern of mutation is similar to what was initially observed at mutations in the *TERT* promoter, and is suggestive of driver activity. These mutations also occur at the same positions within a motif (GAAC) as mutations in the *PLEKHS1* promoter, potentially suggesting a common process is occuring at these two loci. These mutations lie far from any exon-intron boundaries, ruling out the possibility that they affect donor or acceptor sites. This regions overlaps a DNase I hypersensitive site, potentially suggesting that this region contains on intronic regulatory elements.

I additionally identified recurrent mutations at highly conserved positions overlapping the miRNA *MIR142* (Figure 17). These mutations are spread throughout the region, and occur exclusively in lymphoma samples, suggesting that this region may be a target of somatic hypermutation. *Puente et al.* [79] also identified recurrent mutations near *MIR142* in CLL, which they attribute to somatic hypermutation. Despite the fact that this region may be a target of hypermutation rather than selection, the appearance of this region within the top ten non-coding, non-hypermutated regions in terms of combined score (Table 5) but not recurrence score (Table 3) suggests that conservation can highlight regions that are highly conserved but have lower recurrence. As a result, it may be useful to use both scores seperately to nominate regions with different characteristics. For example, the region I highlight in Figure 16 also appears in Table 5, but not Table 3.

Finally, I highlight a recurrently mutated region in an intron if the gene *MSRA* (Figure 18). Similar to several of the other regions highlighted, this region is mutated predominantly at two base positions, which in this case occur at neighboring positions.

6.6 Cancer type specific analysis

So far, I have focussed on regions that are mutated in multiple cancer types. It may also be that some non-coding driver mutations are mutated primarily in one or a few cancer types only. I therefore attempted to identify recurrently non-coding mutations in a cancer type specific manner my applying my scoring method independently to each cancer type in the dataset with more than 75 whole genomes. Consistent with my pan-cancer analysis, when I applied my method to the exonic regions of specific cancer types, I again identified many known cancer genes that scored highly within this sample (Figure 19). Several of the genes that I considered seem to be particularly prominent in cancer types where they are known to be highly mutated, such as *VHL* in renal cancer, *PIK3CA* in breast cancer, *TP53* in ovarian cancer, *SMAD4* in esophageal and gastric cancer, and *KRAS* in pancreatic cancer.

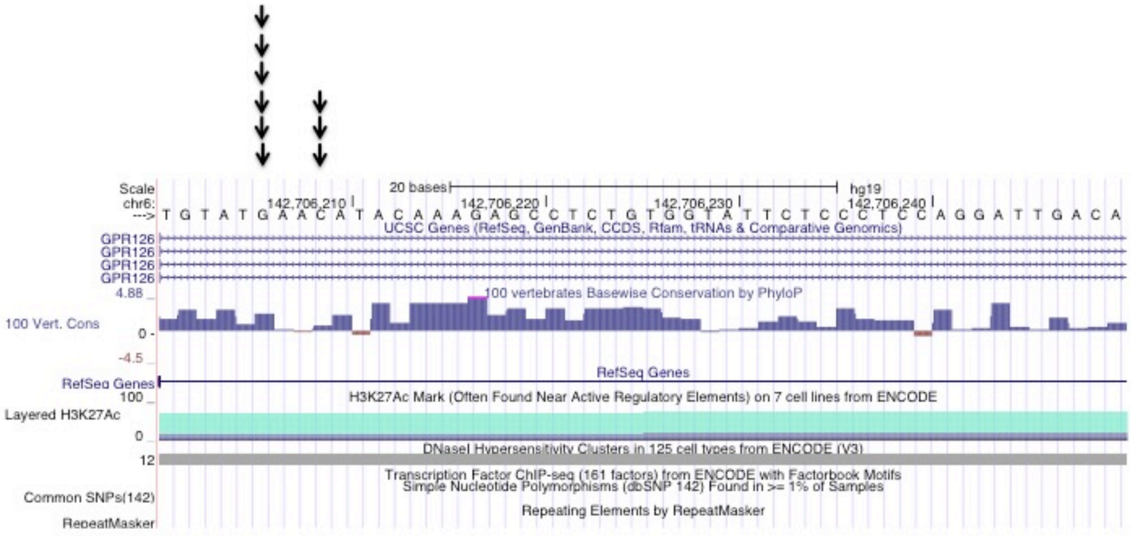
Scale
20 bases
hg19
chr6:
142,706,210
142,706,220
142,706,230
142,706,240
---> T G T A T G A A C A T A C A A A G A G C C T C T G T G G T A T T C T C C C T C C A G G A T T G A C A
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
GPR126
GPR126
GPR126
GPR126
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
RefSeq Genes
RefSeq Genes
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
12
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 16: UCSC browser image of a second recurrently mutated region identified by my method. Mutations are depicted by black arrows. This region overlaps a resonably conserved intron of the gene *GPR126.* The mutations within this region occur exclusively at two nucleotides in a wholly mutually-exclusive manner.

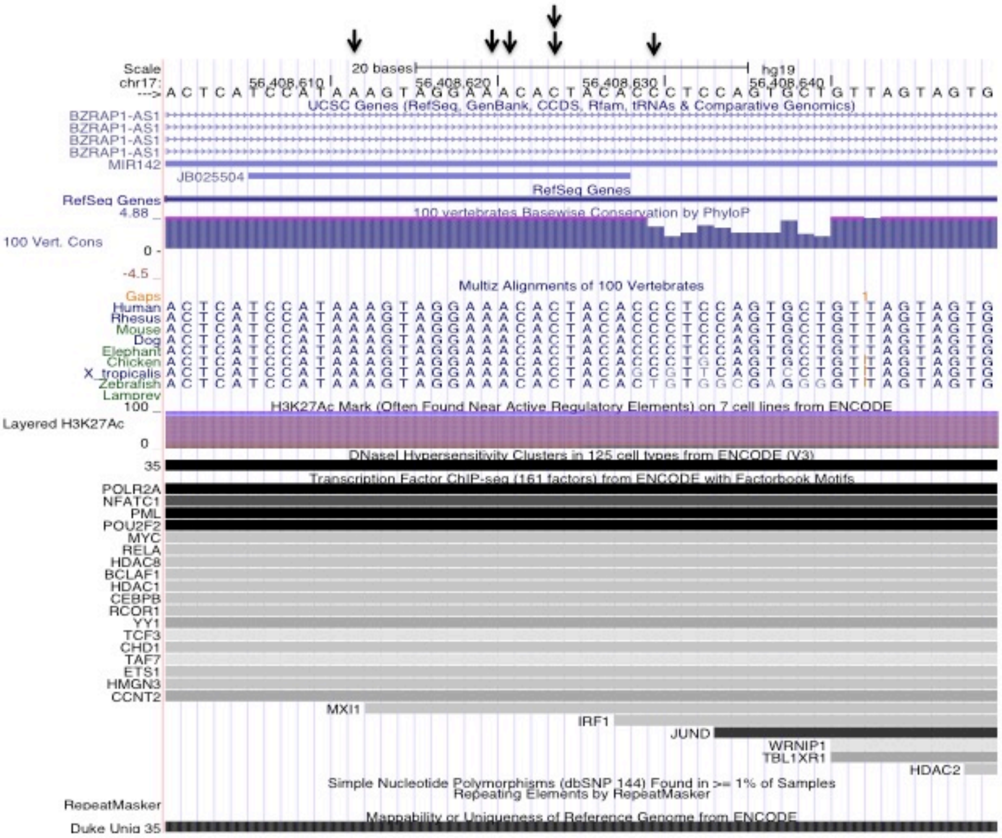
Scale
20 bases
hg19
chr17:
56,408,610
56,408,620
56,408,630
56,408,640
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
BZRAP1-AS1
BZRAP1-AS1
BZRAP1-AS1
BZRAP1-AS1
MIR142
JB025504
RefSeq Genes
RefSeq Genes
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
35
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
POLR2A
NFATC1
PML
POU2F2
MYC
RELA
HDAC8
BCLAF1
HDAC1
CEBPB
RCOR1
YY1
TCF3
CHD1
TAF7
ETS1
HMGN3
CCNT2
MXI1
IRF1
JUND
WRNIP1
TBL1XR1
HDAC2
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
RepeatMasker
Mappability or Uniqueness of Reference Genome from ENCODE
Duke Uniq 35

Figure 17: Recurrently mutated region overlapping the miRNA *MIR142*. The region is highly conserved, as suggested by its inclusion among the top non-coding regions based on combined score. The mutations are spread throughout the region and occur exclusively in lymphoma samples, suggesting that this region may be a target of somatic hypermutation, not nessesarily selection.

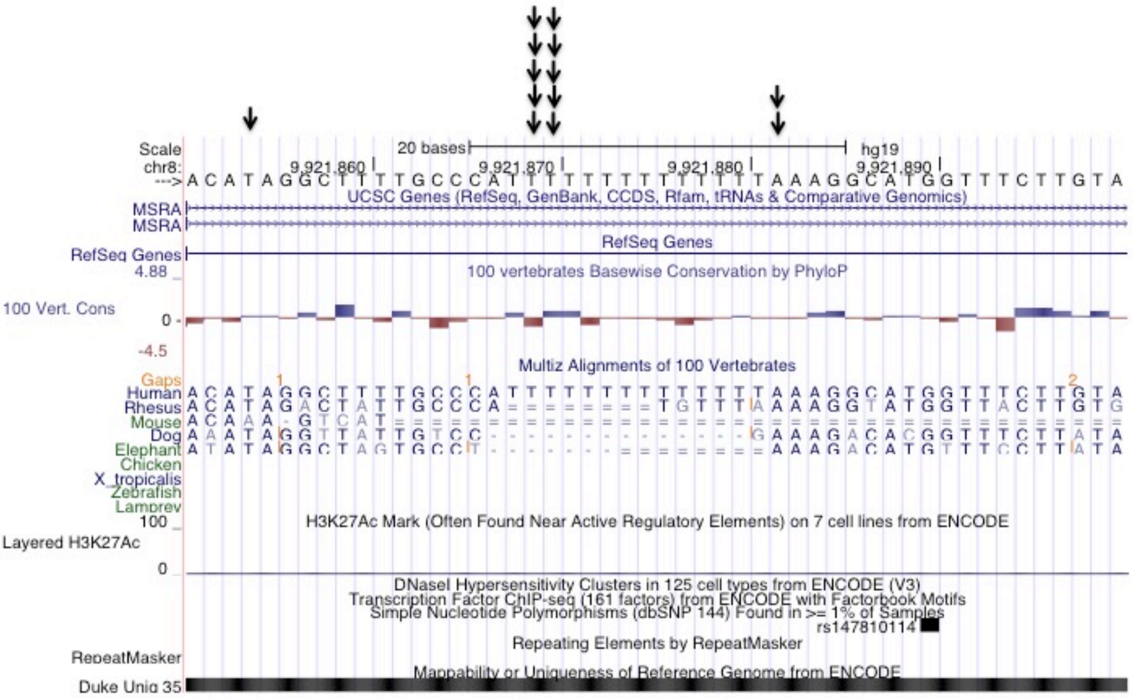
Scale
20 bases
hg19
chr8:
9,921,860
9,921,870
9,921,880
9,921,890
---> A C A T A G G C T T T T G C C C A T T T T T T T T T T T T T T A A A G G C A T G G T T T C T T G T A
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
MSRA
MSRA
RefSeq Genes
RefSeq Genes
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
rs147810114
Repeating Elements by RepeatMasker
RepeatMasker
Mappability or Uniqueness of Reference Genome from ENCODE
Duke Uniq 35

Figure 18: Recurrently mutation overlapping an intron of the gene MSRA*.* The mutations occur primarily at two neighboring bases.

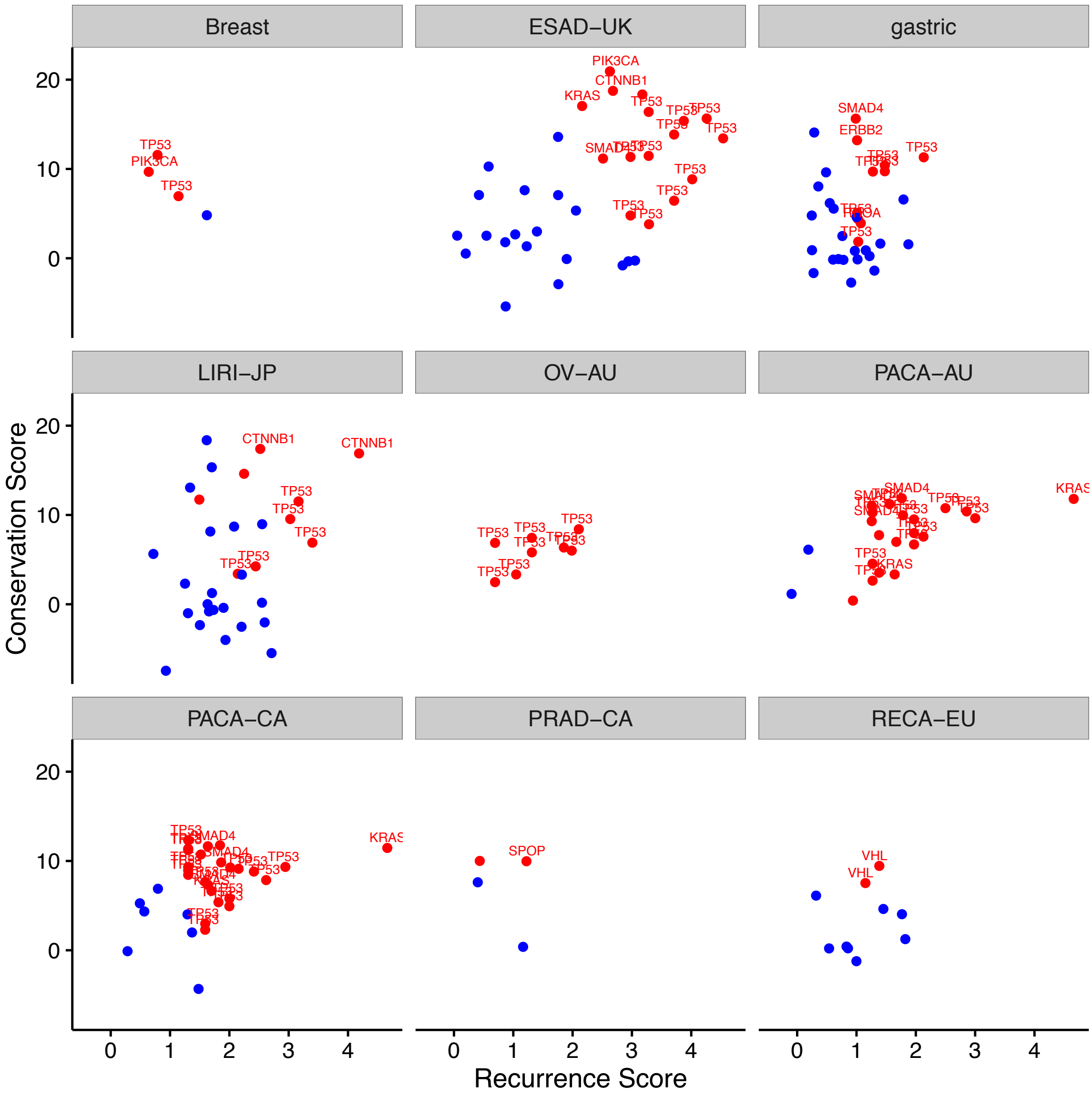

Figure 19: Scatterplots of exonic regions with more than 2 patients mutated within each cancer type. For each scatterplot, I plot regions mutated in three or more samples from a cancer type based on scores calculated only within each cancer type. Regions overlapping known driver genes are depicted in red, while other coding regions are depicted in blue. Several known driver genes are labelled in each plot. Several known cancer type specific trends are apparent, such as the recurrence of *VHL* specifically in renal cancer, as well as prominent *KRAS* mutations in pancreatic cancer and *SMAD4* mutations in gastric and esophageal cancer.

6.7 Cancer type specific non-coding mutations

In additon to the regions identified in my pan-cancer analysis, I also identified some non-coding regions that are recurrently mutated in individual cancer types (Tables 6 and 7). I identified recurrent mutations within an intron of the *PRIM2* gene (Figure 20) specifically in renal cancer. These mutations occurred at two bases in a whole mutally exclusive manner, and exclusively in renal cancer samples. I also identifed recurrent mutations within an intron of *RAD51B* in several breast cancer samples (Figure 21). *RAD51B* is a DNA repair gene invloved in homologous recombination [80]. Given the importance of this repair pathway in breast cancer, this region may be worth further study in this cancer type. Within the regions prioritised by the combined score, I also identified several extremely highly conserved regions that are recurrently mutated in the LIRI-JP cohort, including non-coding regions of the genes *BCL11A*, *BCL6*, and *PAX5* (Table 7).

Table 6: Top ten non-coding, non-hypermutated regions in terms of recurrence score within each cancer type.

chr	start	End	mutated samples	score	Cohort	Annotation
chr10	75457700	75457750	4	1.4405	Breast	*AGAP5* intron
chr10	115511550	115511600	5	1.6216	Breast	*PLEKHS1* promoter
chr11	296200	296250	3	0.9419	Breast	Non-coding
chr14	69134600	69134650	3	1.2896	Breast	*RAD51B* intron
chr14	104675450	104675500	4	1.0002	Breast	Non-coding
chr16	88260400	88260450	3	0.9971	Breast	Non-coding
chr16	88463300	88463350	3	1.0136	Breast	Non-coding
chr17	5025550	5025600	3	1.1389	Breast	*ZNF232* intron
chr22	26714000	26714050	5	1.5020	Breast	*SEZ6L* intron
chr4	120322450	120322500	4	2.3094	Breast	Non-coding
chr1	52344450	52344500	4	27.1847	ESAD-UK	*NDR1* promoter
chr10	30966600	30966650	5	28.3009	ESAD-UK	Non-coding
chr11	63777600	63777650	4	27.7663	ESAD-UK	*MACROD1* intron
chr16	89081750	89081800	6	34.0315	ESAD-UK	Non-coding
chr17	521450	521500	5	34.1686	ESAD-UK	*VPS53* intron
chr17	552650	552700	6	42.0077	ESAD-UK	*VPS53* intron
chr17	78287250	78287300	6	47.7300	ESAD-UK	*RNF213* intron
chr19	49172650	49172700	4	27.2390	ESAD-UK	*NTN5* intron
chr7	151591850	151591900	5	33.9840	ESAD-UK	Non-coding
chr9	42858850	42858900	4	27.6646	ESAD-UK	LOC286297 ncRNA
chr1	46385450	46385500	6	4.1470	Gastric	*MAST2* intron
chr10	129059650	129059700	11	4.6119	Gastric	*DOCK1* intron
chr16	4039350	4039400	6	4.0935	Gastric	*ADCY9* intron
chr17	62640700	62640750	5	3.9717	Gastric	*SMURF2* intron
chr19	19088350	19088400	7	4.5256	Gastric	Non-coding
chr2	25089450	25089500	6	3.8879	Gastric	*ADCY3* intron
chr3	195892250	195892300	18	11.3319	Gastric	Non-coding
chr4	169312800	169312850	8	6.5470	Gastric	*DDX60L* intron
chr4	169313950	169314000	5	3.7960	Gastric	*DDX60L* intron
chr5	179868450	179868500	6	3.9871	Gastric	non-coding TF binding
chr10	89346150	89346200	5	9.4418	LIRI-JP	non-coding TF binding
chr16	1709700	1709750	4	11.7738	LIRI-JP	*CRAMP1L* intron
chr19	893450	893500	4	17.4535	LIRI-JP	*MED16* promoter
chr20	796800	796850	5	12.8574	LIRI-JP	CTCF binding
chr20	10151100	10151150	6	7.6185	LIRI-JP	*SNAL25-AS1* ncRNA
chr21	16350350	16350400	7	9.6832	LIRI-JP	*NRIP1* intron
chr3	43746350	43746400	7	10.0986	LIRI-JP	*ABHD5* intron
chr4	119394200	119394250	6	7.9062	LIRI-JP	Non-coding
chr4	142289250	142289300	7	7.3136	LIRI-JP	Non-coding
chr7	65090800	65090850	6	8.1183	LIRI-JP	Non-coding
chr1	169975450	169975500	3	-0.4557	OV-AU	*KIFAP3* intron
chr1	216051800	216051850	3	-0.4633	OV-AU	*USH2A* intron
chr12	88776700	88776750	3	0.0648	OV-AU	Non-coding
chr12	101545800	101545850	3	-0.1503	OV-AU	Non-coding
chr18	32342800	32342850	3	-0.2974	OV-AU	*DTNA* intron
chr2	69530800	69530850	3	1.0857	OV-AU	Non-coding
chr2	166337300	166337350	3	0.2182	OV-AU	*CSRNP3* intron
chr5	97091250	97091300	3	-0.0193	OV-AU	non-coding
chr5	131366350	131366400	3	0.3409	OV-AU	non-coding

chr8	117337100	117337150	4	0.3649	OV-AU	LINC00536 / TF binding
chr1	51589650	51589700	3	2.5931	PACA-AU	*C1orf185* intron
chr10	118803350	118803400	4	2.0360	PACA-AU	*KIAA15998* intron
chr14	74083000	74083050	3	3.4429	PACA-AU	non-coding
chr16	48483950	48484000	5	4.2544	PACA-AU	*MIR5095* non-coding
chr16	69602250	69602300	3	2.6323	PACA-AU	*NFAT5* intron
chr17	1600600	1600650	3	3.4882	PACA-AU	non-coding
chr2	191157850	191157900	3	2.0794	PACA-AU	*HIBCH* intron
chr2	203851600	203851650	3	2.2701	PACA-AU	non-coding
chr20	14680700	14680750	4	2.2964	PACA-AU	*MACROD2* intron
chr6	107807150	107807200	3	3.2244	PACA-AU	non-coding
chr1	21793400	21793450	3	4.0058	PACA-CA	*NBPF3* intron
chr1	27503450	27503500	3	4.3162	PACA-CA	non-coding
chr1	206519250	206519300	5	6.8580	PACA-CA	*SRGAP2* intron
chr10	89005600	89005650	3	3.5191	PACA-CA	*NUTM2A-AS1* ncRNA
chr15	67334650	67334700	4	4.1283	PACA-CA	non-coding TF binding
chr16	70039200	70039250	4	3.7022	PACA-CA	*PDXDC2P* ncRNA
chr5	141582500	141582550	4	4.5125	PACA-CA	non-coding
chr9	123950	124000	5	6.0043	PACA-CA	*CBWD1* intron
chr9	42540200	42540250	3	15.9414	PACA-CA	non-coding
chr9	70996500	70996550	4	3.9799	PACA-CA	*PGM5* intron
chr1	228032600	228032650	3	1.5023	PRAD-CA	*PRSS38* intron
chr15	41726800	41726850	3	1.4188	PRAD-CA	*RTF1* intron
chr15	73669700	73669750	4	2.2093	PRAD-CA	non-coding
chr16	173850	173900	3	1.7730	PRAD-CA	*NPRL3* intron
chr16	68244200	68244250	4	2.9728	PRAD-CA	*NFATC3* intron
chr19	49815100	49815150	3	1.3712	PRAD-CA	*SLC6A16* intron
chr2	131394050	131394100	4	2.2702	PRAD-CA	*POTEJ* intron
chr4	39684550	39684600	6	3.5422	PRAD-CA	non-coding
chr7	141323150	141323200	4	1.5960	PRAD-CA	*AGK* intron
chr9	66313600	66313650	3	2.4021	PRAD-CA	non-coding
chr1	46226150	46226200	6	2.0896	RECA-EU	non-coding
chr11	44336600	44336650	5	1.1582	RECA-EU	non-coding
chr12	655150	655200	6	1.3493	RECA-EU	*B4GALNT3* intron
chr12	122822000	122822050	5	1.4043	RECA-EU	*CLIP1* intron
chr2	84825300	84825350	7	1.3525	RECA-EU	*DNAH6* intron
chr3	46746150	46746200	6	3.1945	RECA-EU	*TMIE* intron
chr3	56204600	56204650	5	1.8347	RECA-EU	*ERC2* intron
chr6	57343100	57343150	12	4.5837	RECA-EU	*PRIM2* intron
chr7	28744950	28745000	6	1.6668	RECA-EU	*CREB5* intron
chr8	29901300	29901350	9	2.7300	RECA-EU	non-coding

Table 7: Top ten non-coding, non-hypermutated regions in terms of combined score within each cancer type.

chr	start	End	mutated samples	score	Cohort	Annotation
chr10	75457700	75457750	4	0.5052	Breast	*APGAP5* intron
chr10	115511550	115511600	5	0.9404	Breast	*PLEKHS1* promoter
chr13	23615650	23615700	3	0.4879	Breast	non-coding
chr14	69134600	69134650	3	0.8344	Breast	*RAD51B* intron
chr16	10746650	10746700	4	0.5450	Breast	*TEXT5* intron / TF binding
chr16	88260400	88260450	3	0.6453	Breast	non-coding
chr16	88463300	88463350	3	0.6663	Breast	non-coding
chr19	42466450	42466500	3	0.9510	Breast	non-coding
chr4	120322450	120322500	4	0.6950	Breast	non-coding
chr6	168637600	168637650	3	0.9844	Breast	non-coding
chr1	52344450	52344500	4	13.6021	ESAD-UK	*NRD1* promoter
chr10	30966600	30966650	5	12.6044	ESAD-UK	non-coding
chr16	89081750	89081800	6	15.8425	ESAD-UK	non-coding
chr17	521450	521500	5	16.9202	ESAD-UK	*VPS53* intron
chr17	552650	552700	6	20.3809	ESAD-UK	*VPS53* intron
chr17	78287250	78287300	6	23.4776	ESAD-UK	*RNF213* intron
chr6	38461900	38461950	5	13.9464	ESAD-UK	*BTBD9* intron / CTCF binding
chr7	127898750	127898800	4	13.2956	ESAD-UK	non-coding
chr7	151591850	151591900	5	17.0298	ESAD-UK	non-coding
chr9	42858850	42858900	4	13.6816	ESAD-UK	LOC286297 ncRNA
chr11	31150650	31150700	3	2.2649	gastric	*DCDC1* intron
chr17	31038650	31038700	4	2.3389	gastric	*MYO1D* intron
chr17	59465950	59466000	3	4.6770	gastric	*BCAS3* intron
chr2	143949750	143949800	3	2.5988	gastric	*ARHGAP15* intron
chr3	195892250	195892300	18	5.8723	gastric	non-coding
chr4	169312800	169312850	8	3.6715	gastric	*DDX60L* intron
chr4	169313950	169314000	5	2.3140	gastric	*DDX60L* intron
chr6	43041350	43041400	5	2.3349	gastric	*KLC4* intron
chr6	50570100	50570150	3	2.8894	gastric	CTCF binding
chr8	65519150	65519200	3	2.7909	gastric	*CYP7B1* intron
chr16	1709700	1709750	4	5.8069	LIRI-JP	*CRAMP1* intron
chr16	52531300	52531350	3	9.9815	LIRI-JP	*TOX3* intron
chr19	893450	893500	4	11.9033	LIRI-JP	*MED16* promoter
chr2	7342150	7342200	3	7.7298	LIRI-JP	non-coding
chr2	60684450	60684500	3	5.9503	LIRI-JP	*BCL11A*
chr20	796800	796850	5	6.1129	LIRI-JP	CTCF binding
chr21	16350350	16350400	7	5.6199	LIRI-JP	*NRIP1* intron
chr3	43746350	43746400	7	5.4083	LIRI-JP	*ABHD5* intron
chr3	187439750	187439800	3	8.5100	LIRI-JP	*BCL6* intron
chr9	36940450	36940500	3	5.2957	LIRI-JP	*PAX5* intron
chr1	169975450	169975500	3	-0.0932	OV-AU	*KIFAP3* intron
chr1	216051800	216051850	3	-0.2351	OV-AU	*USH2A* intron
chr12	88776700	88776750	3	0.0358	OV-AU	non-coding
chr12	101545800	101545850	3	-0.0795	OV-AU	non-coding
chr18	32342800	32342850	3	-0.0308	OV-AU	*DTNA* intron
chr2	69530800	69530850	3	0.0374	OV-AU	non-coding
chr2	166337300	166337350	3	0.0700	OV-AU	*CSRNP3* intron
chr5	97091250	97091300	3	-0.3319	OV-AU	non-coding
chr5	131366350	131366400	3	0.2291	OV-AU	non-coding
chr8	117337100	117337150	4	0.6292	OV-AU	LINC00536 ncRNA /

						TF binding
chr14	74083000	74083050	3	0.9026	PACA-AU	non-coding
chr16	48483950	48484000	5	2.1310	PACA-AU	*MIR5059* ncRNA
chr16	69602250	69602300	3	1.9837	PACA-AU	*NFAT5* intron
chr17	1600600	1600650	3	1.7198	PACA-AU	non-coding
chr19	5250450	5250500	3	1.4586	PACA-AU	*PTPRS* intron
chr2	203851600	203851650	3	1.7673	PACA-AU	non-coding
chr20	14680700	14680750	4	0.9078	PACA-AU	*MACROD2* intron
chr6	107807150	107807200	3	1.7200	PACA-AU	non-coding
chr7	122981650	122981700	3	0.9430	PACA-AU	non-coding
chr8	120286150	120286200	3	0.9145	PACA-AU	non-coding
chr1	21793400	21793450	3	2.0401	PACA-CA	*NBPF3* intron
chr1	27503450	27503500	3	4.8274	PACA-CA	non-coding
chr1	206519250	206519300	5	2.8859	PACA-CA	*SRGAP2* intron
chr10	89005600	89005650	3	2.0901	PACA-CA	*NUTM2A-AS1* ncRNA
chr18	44002100	44002150	3	2.0086	PACA-CA	*RNF165* intron
chr5	99390600	99390650	3	3.9516	PACA-CA	non-coding
chr9	123950	124000	5	3.3592	PACA-CA	*CBWD1* intron
chr9	35357550	35357600	3	1.8179	PACA-CA	*UNC13B* intron
chr9	42540200	42540250	3	8.0164	PACA-CA	non-coding
chr9	70996500	70996550	4	1.8112	PACA-CA	*PGM5* intron
chr10	16330150	16330200	3	0.4259	PRAD-CA	non-coding
chr14	75148450	75148500	3	0.5223	PRAD-CA	*AREL1* intron
chr15	73669700	73669750	4	1.0125	PRAD-CA	non-coding
chr16	173850	173900	3	0.9023	PRAD-CA	*NPRL3* intron
chr16	68244200	68244250	4	0.3793	PRAD-CA	*NFATC3* intron
chr17	29476300	29476350	3	0.5816	PRAD-CA	*NF1* intron
chr19	7152400	7152450	3	0.4746	PRAD-CA	*INSR* intrron
chr2	203520450	203520500	3	0.4251	PRAD-CA	*FAM117B* intron
chr4	39684550	39684600	6	1.0123	PRAD-CA	non-coding
chr7	143666450	143666500	3	0.6287	PRAD-CA	non-coding
chr1	46226150	46226200	6	1.2094	RECA-EU	non-coding
chr12	69755800	69755850	4	1.0794	RECA-EU	*YEATS4* intron
chr12	122822000	122822050	5	1.0534	RECA-EU	*CLIP1* intron
chr16	18820800	18820850	4	3.3245	RECA-EU	*SMG1* UTR
chr18	57125350	57125400	4	1.0331	RECA-EU	*CCBE1* intron
chr3	46746150	46746200	6	2.4872	RECA-EU	*TMIE* intron
chr3	56204600	56204650	5	1.2752	RECA-EU	*ERC2* intron
chr6	57343100	57343150	12	2.0394	RECA-EU	*PRIM2* intron
chr7	28744950	28745000	6	1.0070	RECA-EU	*CREB5* intron
chr8	29901300	29901350	9	1.5436	RECA-EU	non-coding

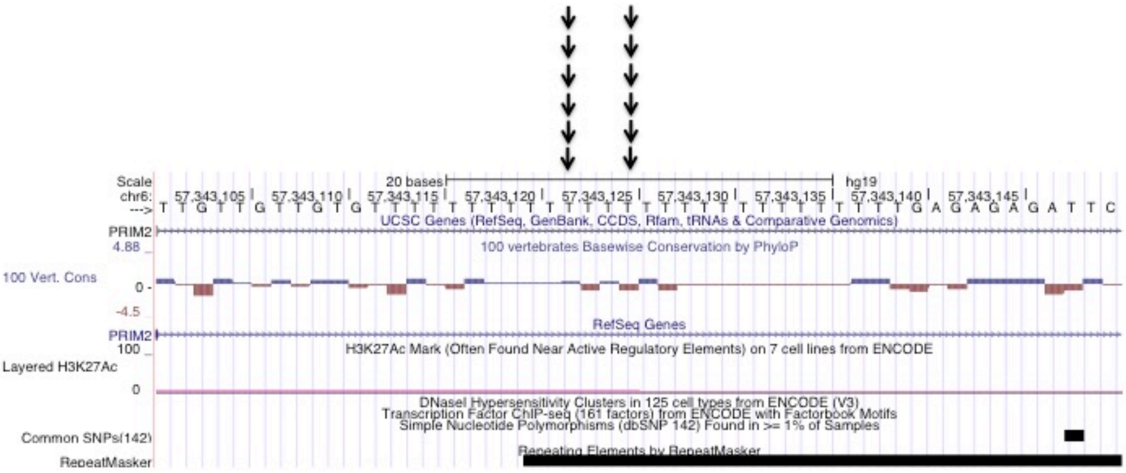
Scale
20 bases
hg19
chr6:
57,343,105
57,343,110
57,343,115
57,343,120
57,343,125
57,343,130
57,343,135
57,343,140
57,343,145
--->
T T G T T G T T G T G T T T T T T T T T T T T T T T T T T T T T T T T T T G A G A G A G A T T C
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PRIM2
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
RefSeq Genes
PRIM2
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 20: UCSC browser image of a recurrently mutated region overlapping an intron of the gene *PRIM2*. Mutations are depicted as black arrows. All mutations within the region occur at one of two nucleotides and are mutually exclusive.

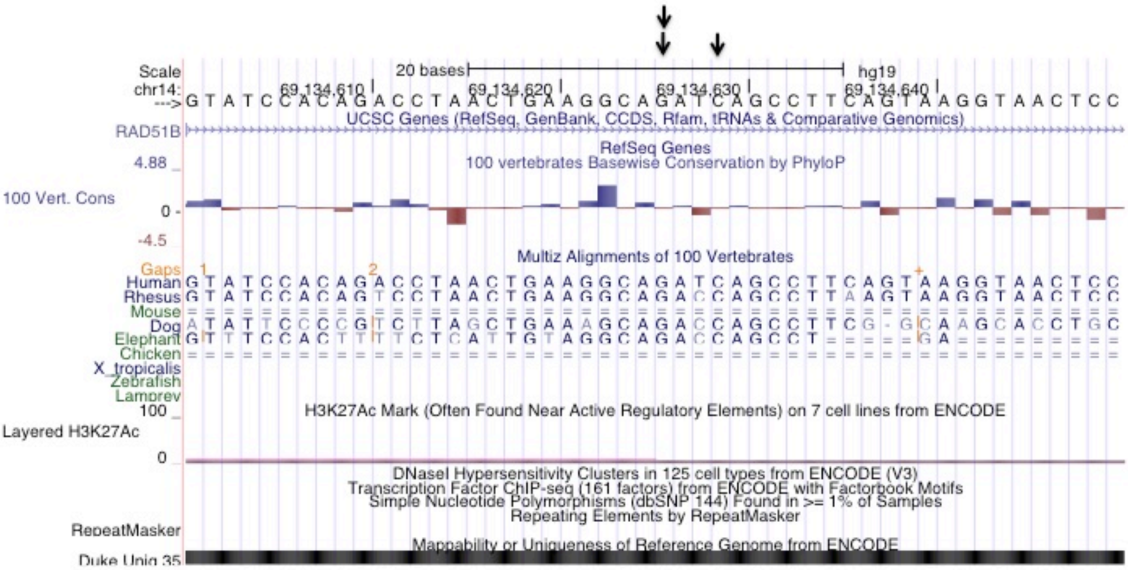
Scale
20 bases
hg19
chr14:
69,134,610
69,134,620
69,134,630
69,134,640
---> G T A T C C A C A G A C C T A A C T G A A G G C A G A T C A G C C T T C A G T A A G G T A A C T C C
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
RAD51B
RefSeq Genes
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
RepeatMasker
Mappability or Uniqueness of Reference Genome from ENCODE
Duke Uniq 35

Figure 21: UCSC browser image depiciting a recurrently mutated region in an intron of the DNA repair gene *RAD51B*. This region is mutated specifically in breast cancer.

7 Discussion

As is the case in the analysis of coding mutations, I have found that mutational heterogeneity is a critical factor that impacts the identification of non-coding driver regions in cancer. My initial analysis revealed that several promising candidate regions, some of which have been suggested in the literature as potential driver regions, may actually be recurrently mutated primarily due to focal mutational processes rather than selection. I highlight AID induced somatic hypermutation as well as a recenty identified process [23] which targets CTCF binding sites as a prominent local mutational process. I also propose methods for identifying and filtering out these putatively hypermutated regions, allowing me to focus on regions for which I believe the evidence favoring positive selection is stronger.

Using the exome to validate my scoring system, I showed that all three scores can differentiate known drivers from other coding regions. I also identified several known driver genes that display a mutation pattern across cancer types consistent with expectations. For example, I scored *VHL* highly specifically in renal cancer, consistent with the known high mutation rate of this gene in this tumour type. Several well known driver genes are are known to be mutated in many cancer types were scored highly in several cancer types indendently, such as *KRAS* and *TP53*. While these genes scored highly across many cancer types, they also appear more prominently in certain cancer types where they are known to be particularly highly mutated, such as *TP53* in ovarian cancer and *KRAS* in pancreatic cancer.

In addition to using recurrence as previous studies often have, I also included conservation as part of my prioritization scores. I show that my conservation score can separate known coding drivers from non-drivers. Conservation may also be useful in the analysis of non-coding mutations, both to increase confidence that recurrent non-coding mutations have to potential to impact function, as well as to highlight non-coding regions that may have lower recurrence but strong driver potential due to strong conservation. The combined score also appears to outperform the recurrence score alone in terms of distinguishing known driver regions from other exonic regions, suggesting that conservation provides valuable information in addition to recurrence, although this may be more difficult to interpret within the context of non-coding mutations, given that non-coding regions are generally less well conserved as a whole compared to coding regions. On the one hand, the generally low conservation present in non-coding regions sugggests that functional non-coding mutations might not necessarily always occur at conserved positions. Thus, it is useful to consider recurrent mutations, even if they are not at highly conserved positions. On the other hand, high conservation implies a higher likelihood that a base has functional importance, so incorporating conservation as a complement to recurrence can help strengthen the case that a region has driver potential. Using a measure such as the combined score may also highlight regions that have moderate recurrence but which are highly conserved. These regions would be good candidates for more “hill-like” drivers. As a result, I believe that using both recurrence and a combined score that incorporates both recurrence and conservation to prioritise regions that may have different properties is a promising strategy. It is also worth noting that more complex ways of combining these scores might yield additional benefits. I have simply averaged the scores, after normalizing to make the scores roughly comparable, but other transformations might also produce insights.

Within these genomes, I also identified several novel recurrently mutated regions, some of which may operate through novel mechanisms. In addition to the novel recurrent regions I identified in a pan-cancer analysis, I aso identifed several novel non-coding regions that appear to be cancer type specific, some of which have high frequencies in the cancer types in which they occur. These regions, as well as other regions that score highly within my framework, may be good targets for future analyses of non-coding somatic mutations in cancer. Although the methods used here can not definitively establish a mutation as a driver, further investigation of non-coding mutations using these and other methods may reveal new non-coding driver mutations. These drivers may have important implications for cancer therapy if they are directly targetable by drugs or involved in the regulation of pathways that are targetable. For example, my analysis identified mutations within an intron of the gene *GPR126*, a G-protein coupled receptor that has been associated with Adolescent idiopathic scoliosis [81]. Although I have not definitively shown the mutations within this region to be functional, the fact that this gene is already thought to play a role in human disease is intriguing. Non-coding mutations such as *TERT* promoter mutations [50] have been associated with clinical outcomes, as have mutational processes in cancer [82-84]. I have highlighted regions that have an excess of mutations in cancer genomes. These regions may lead to important insights that may have clinical implications if they are either under selection or indicative of a unique mutational processes.

8 References

1. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2014.
2. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
3. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
4. Garraway, L.A. and E.S. Lander, *Lessons from the cancer genome.* Cell, 2013. **153**(1): p. 17-37.
5. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
6. Furney, S.J., et al., *Prioritization of candidate cancer genes--an aid to oncogenomic studies.* Nucleic Acids Res, 2008. **36**(18): p. e115.
7. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
8. Schuster-Bockler, B. and B. Lehner, *Chromatin organization is a major influence on regional mutation rates in human cancer cells.* Nature, 2012. **488**(7412): p. 504-+.
9. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-5.
10. Supek, F., et al., *Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers.* Cell, 2014. **156**(6): p. 1324-1335.
11. Gartner, J.J., et al., *Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(33): p. 13481-13486.
12. Fredriksson, N.J., et al., *Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types.* Nat Genet, 2014. **46**(12): p. 1258-63.
13. Pfeifer, G.P., *How the environment shapes cancer genomes.* Curr Opin Oncol, 2015. **27**(1): p. 71-7.
14. Viros, A., et al., *Ultraviolet radiation accelerates BRAF-driven melanomagenesis by targeting TP53.* Nature, 2014. **511**(7510): p. 478-82.
15. Pleasance, E.D., et al., *A comprehensive catalogue of somatic mutations from a human cancer genome.* Nature, 2010. **463**(7278): p. 191-6.
16. Govindan, R., et al., *Genomic landscape of non-small cell lung cancer in smokers and never-smokers.* Cell, 2012. **150**(6): p. 1121-34.
17. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. **463**(7278): p. 184-90.
18. Roberts, S.A. and D.A. Gordenin, *Hypermutation in human cancer genomes: footprints and mechanisms.* Nat Rev Cancer, 2014. **14**(12): p. 786-800.
19. Poon, S.L., et al., *Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool.* Science Translational Medicine, 2013. **5**(197).

20. Tomasetti, C., B. Vogelstein, and G. Parmigiani, *Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation.* Proc Natl Acad Sci U S A, 2013. **110**(6): p. 1999-2004.
21. Polak, P., et al., *Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair.* Nature Biotechnology, 2014. **32**(1): p. 71-+.
22. Supek, F. and B. Lehner, *Differential DNA mismatch repair underlies mutation rate variation across the human genome.* Nature, 2015.
23. Katainen, R., et al., *CTCF/cohesin-binding sites are frequently mutated in cancer.* Nat Genet, 2015.
24. Reijns, M.A., et al., *Lagging-strand replication shapes the mutational landscape of the genome.* Nature, 2015.
25. Polak, P., et al., *Cell-of-origin chromatin organization shapes the mutational landscape of cancer.* Nature, 2015. **518**(7539): p. 360-4.
26. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
27. Liu, L., S. De, and F. Michor, *DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes.* Nat Commun, 2013. **4**: p. 1502.
28. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-93.
29. Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas, *Identification of oncogenic driver mutations.* Experimental Medicine, 2014. **32**.
30. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
31. Tamborero, D., et al., *Comprehensive identification of mutational cancer driver genes across 12 tumor types.* Sci Rep, 2013. **3**: p. 2650.
32. Dees, N.D., et al., *MuSiC: identifying mutational significance in cancer genomes.* Genome Res, 2012. **22**(8): p. 1589-98.
33. Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.* Bioinformatics, 2013. **29**(18): p. 2238-44.
34. Gonzalez-Perez, A. and N. Lopez-Bigas, *Functional impact bias reveals cancer drivers.* Nucleic Acids Res, 2012. **40**(21): p. e169.
35. Reimand, J. and G.D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Mol Syst Biol, 2013. **9**: p. 637.
36. Wood, L.D., et al., *The genomic landscapes of human breast and colorectal cancers.* Science, 2007. **318**(5853): p. 1108-13.
37. Davies, H., et al., *Mutations of the BRAF gene in human cancer.* Nature, 2002. **417**(6892): p. 949-54.
38. Kong-Beltran, M., et al., *Somatic mutations lead to an oncogenic deletion of met in lung cancer.* Cancer Res, 2006. **66**(1): p. 283-9.
39. Davoli, T., et al., *Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome.* Cell, 2013. **155**(4): p. 948-62.
40. Stehr, H., et al., *The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors.* Mol Cancer, 2011. **10**: p. 54.

41. Huang, F.W., et al., *Highly Recurrent TERT Promoter Mutations in Human Melanoma.* Science, 2013. **339**(6122): p. 957-959.
42. Horn, S., et al., *TERT promoter mutations in familial and sporadic melanoma.* Science, 2013. **339**(6122): p. 959-61.
43. Vinagre, J., et al., *Frequency of TERT promoter mutations in human cancers.* Nat Commun, 2013. **4**: p. 2185.
44. Gunes, C., et al., *Expression of the hTERT gene is regulated at the level of transcriptional initiation and repressed by Mad1.* Cancer Research, 2000. **60**(8): p. 2116-2121.
45. Schulze, K., et al., *Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets.* Nat Genet, 2015. **47**(5): p. 505-11.
46. Hosen, I., et al., *TERT promoter mutations in clear cell renal cell carcinoma.* Int J Cancer, 2015. **136**(10): p. 2448-52.
47. Rachakonda, P.S., et al., *TERT promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(43): p. 17426-17431.
48. Eckel-Passow, J.E., et al., *Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors.* N Engl J Med, 2015.
49. Spiegl-Kreinecker, S., et al., *Prognostic quality of activating TERT promoter mutations in glioblastoma: interaction with the rs2853669 polymorphism and patient age at diagnosis.* Neuro Oncol, 2015.
50. Borah, S., et al., *Cancer. TERT promoter mutations and telomerase reactivation in urothelial cancer.* Science, 2015. **347**(6225): p. 1006-10.
51. Allory, Y., et al., *Telomerase reverse transcriptase promoter mutations in bladder cancer: high frequency across stages, detection in urine, and lack of association with outcome.* Eur Urol, 2014. **65**(2): p. 360-6.
52. Hurst, C.D., F.M. Platt, and M.A. Knowles, *Comprehensive mutation analysis of the TERT promoter in bladder cancer and detection of mutations in voided urine.* Eur Urol, 2014. **65**(2): p. 367-9.
53. Katainen, R., et al., *CTCF/cohesin-binding sites are frequently mutated in cancer.* Nat Genet, 2015. **47**(7): p. 818-21.
54. Mansour, M.R., et al., *An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element.* Science, 2014. **346**(6215): p. 1373-1377.
55. Wang, W., et al., *A frequent somatic mutation in CD274 3'-UTR leads to protein over-expression in gastric cancer by disrupting miR-570 binding.* Hum Mutat, 2012. **33**(3): p. 480-4.
56. Dutton-Regester, K., et al., *A highly recurrent RPS27 5'UTR mutation in melanoma.* Oncotarget, 2014. **5**(10): p. 2912-7.
57. Cunningham, F., et al., *Ensembl 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D662-9.
58. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
59. Lochovsky, L., et al., *LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations.* Nucleic Acids Res, 2015. **43**(17): p. 8123-34.
60. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-7.

61. Melton, C., et al., *Recurrent somatic mutations in regulatory regions of human cancer genomes.* Nat Genet, 2015. **47**(7): p. 710-6.
62. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.
63. Fu, Y., et al., *FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer.* Genome Biol, 2014. **15**(10): p. 480.
64. Poulos, R.C., et al., *Systematic Screening of Promoter Regions Pinpoints Functional Cis-Regulatory Mutations in a Cutaneous Melanoma Genome.* Mol Cancer Res, 2015. **13**(8): p. 1218-26.
65. Smith, K.S., et al., *Signatures of accelerated somatic evolution in gene promoters in multiple cancer types.* Nucleic Acids Res, 2015.
66. Melton, C., et al., *Recurrent somatic mutations in regulatory regions of human cancer genomes.* Nat Genet, 2015.
67. Mathelier, A., et al., *Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas.* Genome Biol, 2015. **16**: p. 84.
68. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
69. Zhang, J., et al., *International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data.* Database (Oxford), 2011. **2011**: p. bar026.
70. Wang, K., et al., *Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer.* Nat Genet, 2014. **46**(6): p. 573-82.
71. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.
72. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
73. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
74. Team, R.D.C., *R: A language and environment for statistica computing*. 2010, R Foundation for Statistical Computing: Vienna, Austria.
75. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.
76. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.
77. Perera, D., et al., *Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes.* Nature, 2016. **532**(7598): p. 259-63.
78. Sabarinathan, R., et al., *Nucleotide excision repair is impaired by binding of transcription factors to DNA.* Nature, 2016. **532**(7598): p. 264-7.
79. Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic leukaemia.* Nature, 2015. **526**(7574): p. 519-24.
80. Takata, M., et al., *The Rad51 paralog Rad51B promotes homologous recombinational repair.* Mol Cell Biol, 2000. **20**(17): p. 6476-82.
81. Kou, I., et al., *Genetic variants in GPR126 are associated with adolescent idiopathic scoliosis.* Nat Genet, 2013. **45**(6): p. 676-9.
82. Le, D.T., et al., *PD-1 Blockade in Tumors with Mismatch-Repair Deficiency.* N Engl J Med, 2015. **372**(26): p. 2509-2520.
83. Snyder, A., et al., *Genetic basis for clinical response to CTLA-4 blockade in melanoma.* N Engl J Med, 2014. **371**(23): p. 2189-99.

84. Waddell, N., et al., *Whole genomes redefine the mutational landscape of pancreatic cancer.* Nature, 2015. **518**(7540): p. 495-501.