Recurrent somatic mutations in regulatory regions of human cancer genomes

Collin Melton1,2, Jason A Reuter1, Damek V Spacek1 & Michael Snyder1

Aberrant regulation of gene expression in cancer can promote survival and proliferation of cancer cells. Here we integrate whole-genome sequencing data from The Cancer Genome Atlas (TCGA) for 436 patients from 8 cancer subtypes with ENCODE and other regulatory annotations to identify point mutations in regulatory regions. We find evidence for positive selection of mutations in transcription factor binding sites, consistent with these sites regulating important cancer cell functions.

Using a new method that adjusts for sample- and genomic locus-specific mutation rates, we identify recurrently mutated sites across individuals with cancer. Mutated regulatory sites include known sites in the TERT promoter and many new sites, including a subset in proximity to cancer-related genes. In reporter assays, two new sites display decreased enhancer activity upon mutation. These data demonstrate that many regulatory regions contain mutations under selective pressure and suggest a greater role for regulatory mutations in cancer than previously appreciated.

Cancer is the second leading cause of death in the United States1. As of 2009, approximately 40% of Americans will develop cancer in their lifetime, and approximately 50% of these individuals will die of their disease2. Despite substantial advances in understanding of the genetic causes of cancer, many therapeutic challenges remain. The complexity of cancer etiology and therapy stems from the fact that no two individuals’ cancers are identical, as cancer arises from the selection of specific point mutations, structural variants and epigenetic alterations from a large pool of such variation.

To better understand the genetic causes of cancer, large-scale projects such as TCGA have performed comprehensive omics profiling of cancer and normal paired samples from thousands of individuals with diverse cancer types. These endeavors have focused primarily on exome sequencing, with more recent efforts involving whole-genome sequencing. Analysis of pan-cancer variation from exome sequencing identified shared sets of mutated genes and pathways among groups of cancer types. Furthermore, these studies have identified mutations in coding genes known as driver mutations that undergo positive selection in cancer3.

Although the majority of sequencing studies in cancer have focused on protein-coding sequences, only a small fraction of the genome encodes protein. Of the remaining genomic sequence, a large portion contains regulatory elements4. It is possible that driver mutations in regulatory elements exist that dysregulate oncogenes and tumor suppressors. Recently, an example of a regulatory mutation in cancer was identified in the regulatory region upstream of the TERT gene (encoding telomerase reverse transcriptase)5,6. Furthermore, Weinhold et al.7 identified recurrent regulatory mutations modulating the expression of PLEKHS1, WDR74 and SDHD in an analysis of mutations in promoters and enhancers. Fredriksson et al.8 identified recurrent mutations in proximity to gene transcriptional start sites (TSSs), although only TERT mutations were significantly associated with altered mRNA transcript levels. Thus far, a genome-wide analysis of potential recurrent mutations in all annotated regulatory regions has yet to be performed.

The Encyclopedia of DNA Elements (ENCODE) Project is a National Human Genome Research Institute (NHGRI)-funded project with the goal of identifying all the functional elements in the human genome. As of 2012, this project assayed up to 12 histone modifications in 46 cell types and 119 different DNA-binding proteins across 72 cell types4. Additional data from this project include DNase I–hypersensitive sites, formaldehyde-assisted isolation of regulatory elements (FAIRE)-identified sequences, DNA methylation, chromosome-interacting regions and RNA transcription. These data and additional genome-wide data sets, including ones from the recent Roadmap Epigenomics Mapping Consortium (REMC)9, have been combined into database resources. One such resource, RegulomeDB10, provides regulatory annotations for any given position in the human genome, enabling facile annotation of regulatory features for potential disease-causing variants.

In this study, we analyze TCGA whole-genome sequencing data to define sets of point mutations for 436 cancer samples from 8 cancer types. We annotate the mutations with regulatory information and implement a statistical framework to define significantly mutated regulatory regions. We identify the previously observed mutations in the TERT promoter and numerous new mutated regulatory sites. This study indicates that there is a far greater role for regulatory region mutations in cancer than has been previously appreciated.

RESULTS
Identification of somatic mutations in cancer
To identify somatic cancer variants that reside in regulatory regions, we established a data processing workflow (Fig. 1a). We subjected...
whole-genome sequencing data generated from cancer and normal tissues collected from 436 patients to rigorous analysis to identify single-nucleotide variants (SNVs) using 2 different algorithms. To increase our power to detect recurrent variants, we analyzed all available patient data from eight different types of cancer (Fig. 1a). We performed additional filtering after mutation calling to remove mutations that were likely falsely called owing to mapping error (Online Methods and Supplementary Fig. 1). This filtering was performed using a heuristic method that searches for homologous genomic regions where a called variant is present in the reference sequence. Lastly, to aid in downstream statistical analyses, we divided our cancer samples into test and validation sets. These two sets were generated to have similar numbers of samples and similar distributions of number of mutations per sample (Supplementary Fig. 1).

Summarizing mutation count data for the samples showed substantial variability between samples within cancer types and between cancer types (Fig. 1b), with a higher mutational load in the lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) samples, consistent with published data11. Mutation allele fractions differed slightly between cancer types, and this might reflect a difference in sample purity and/or clonogenicity (Supplementary Fig. 2a). Mutation count was not correlated with sequencing depth, suggesting that the whole-genome samples were sequenced at sufficient coverage (median of 34-fold for normal and 50-fold for cancer tissues) (Fig. 2a). Supplementary Fig. 1 suggests that the whole-genome samples were sequenced at sufficient coverage (median of 34-fold for normal and 50-fold for cancer tissues) (Fig. 2a). In parallel to our analysis of regulatory sites, we performed enrichment analysis of mutations in coding exons, coding introns, noncoding exons and noncoding introns (Fig. 2c.e). We similarly observed no consistent enrichment or depletion of mutations in coding exons or any other transcript category. This result is consistent with a previous study by Ostrow et al., which examined selective pressure in coding regions through the analysis of synonymous versus nonsynonymous mutation rates. Ostrow et al. found that, in comparison to germline variation, somatic variation in cancer is governed less by purifying selection and more by positive selection13.

**Analysis of categories of regulatory mutations based on RegulomeDB**

Analysis of categories of regulatory mutations based on RegulomeDB score identified some enriched categories and cancer type combinations; however, there was no consistent trend across all cancer types (Fig. 2d). Regulatory mutations with scores 2a, 2c and 3b overlapping with peaks in transcription factor chromatin immunoprecipitation and sequencing (ChiP-seq) identified binding sites with matching motifs. We subcategorized these mutations by transcription factor and computed the enrichment in actual compared to simulated mutations. All sites that passed significance (false discovery rate (FDR) < 0.001) in the combined analysis of all cancer types in both the

![Annotation of cancer mutations](image)

*To investigate the role of regulatory mutations, we annotated our mutation set with gene and regulatory information from GENCODE and RegulomeDB. RegulomeDB is a searchable database of publicly available regulatory data that includes a wealth of experimental data for transcription factors, epigenetic marks, motifs and DNA accessibility10. We used this database because it provides uniformly processed, high-quality annotations of known regulatory information for any given location in the human genome. For any given site, RegulomeDB provides a score between 1 and 7 in order of decreasing regulatory evidence. In addition to regulatory annotation, we performed transcript annotation using GENCODE transcripts. Of note, the potential false positive mutations due to mapping error filtered out of our analysis were highly enriched in non-regulatory regions (Supplementary Fig. 1), presumably owing to the fact that the excluded sites often fell in difficult-to-align genomic regions that are also filtered out in the ENCODE peak calling pipeline employed by RegulomeDB (Supplementary Fig. 1). Overall, the annotation procedures showed that mutations in coding exons and potential regulatory regions represent approximately 0.036–0.056% and 31–39% of the called mutations for each cancer type, respectively (Fig. 2a,b).*

**Recurrent mutation across regulatory element binding sites**

Analysis of categories of regulatory mutations based on RegulomeDB score identified some enriched categories and cancer type combinations; however, there was no consistent trend across all cancer types (Fig. 2d). Regulatory mutations with scores 2a, 2c and 3b overlapping with peaks in transcription factor chromatin immunoprecipitation and sequencing (ChiP-seq) identified binding sites with matching motifs. We subcategorized these mutations by transcription factor and computed the enrichment in actual compared to simulated mutations. All sites that passed significance (false discovery rate (FDR) < 0.001) in the combined analysis of all cancer types in both the
test and validation sets are shown in Figure 2g. There were numerous categories of mutations that were enriched, indicating positive selection in cancer, and few categories that were depleted and thus negatively selected. Clustering of transcription factors by the extent of shared binding sites identified clusters of factors with a higher number of shared sites and similar enrichment patterns (Supplementary Fig. 3b). These clustered transcription factors were often related family members. Mutations in the binding sites of CEBP factors were highly enriched and significant across all cancer types ($P < 1 \times 10^{-10}$, two-sided Fisher’s exact test). CEBP factors are involved in both gene transcription and a specific residue of the motif. This selection might more effectively inhibit transcription factor binding, increase affinity of the transcription factor for the motif or allow the creation of a new motif. We favor the possibility that mutations in CEBP sites modulate but do not destroy the specificity of these sites, as many of the mutations did not decrease the matching of the motif to PWMs from the literature and previous reports showed that a CEBP binding site matching the mutant version identified here drove transcription in luciferase reporter assays.

Analysis of the motif match scores for the mutant sequences of all transcription factors in comparison to reference sequences showed that, for the majority of transcription factors, the mutated sites significantly reduced the match score (Fig. 3c). Comparison of scores to those obtained for randomly mutated motifs showed that six of the transcription factors (GATA3, GATA6, MAFF, POS, AP-1 and NFE2) had mutated sites with match scores significantly worse than with randomly generated mutated sites ($P < 0.001$, two-sided paired Wilcoxon rank-sum test; Fig. 3d). For these factors, the distributions of the deviance for the real and random mutations in these transcription factor sites are shown in Figure 3c, and the identity of these transcription factors and the mean deviance of their scores from those for reference and random sites are shown in Figure 3f. Mutations in the binding sites for these factors tended to be concentrated at one or a few bases of the motif, suggesting selective pressure for these specific mutations (Supplementary Fig. 4f). All six of these transcription factors have known roles in cancer. GATA3 and GATA6 belong to
the GATA family of transcription factors, which are often lost during advanced carcinogenesis16. MAFK belongs to the small MAF protein family and can form heterodimers with CNC transcription factors, including NFE2, and their expression is often lower in cancer17. FOS transcription factors as well as MAF family proteins are components of various AP-1 transcription factor complexes, which are canonically thought to promote proliferation and survival in cancer but in certain contexts may also have tumor-suppressor activity18. Overall, these results indicate recurrent patterns of mutation in specific transcription factor sites that may either enhance or repress gene expression.

Given the uneven distribution of mutations within the binding sites of a number of transcription factors, we investigated the possibility that specific mutations within these sites might create de novo binding sites for other factors. We focused our attention on the highly significant enrichment of mutations in sites of the CEBP transcription factors and screened known binding site motifs for ones having significantly higher match scores with the mutant CEBP sites than with the reference sites. We found that the mutated sites had higher match scores than the reference sites for SOX17, DBP, SOX5, FOS, CMAF, PARP, EN1, BCL6 and CAP PWMs (Supplementary Fig. 4a).

Alignment of the mutant sequences identified a consensus motif shown in Supplementary Fig. 4a, with the most common instances of this motif shown in Supplementary Fig. 4b, c, d, and e. Expression of the transcription factors corresponding to the high-scoring motifs was not significantly different between samples with and without mutations in these sites (Supplementary Fig. 4c). These results suggest that selective pressure exists in cancer to convert CEBP sites into a common oncogenic

Figure 3 Effects of mutations on transcription factor binding sites. (a) A schematic is shown of the methodology of aligning sequences and generating match scores for mutations contained within transcription factor (TF) binding sites. (b) Reference binding regions at mutated sites for each transcription factor were aligned to the factor’s PWM, and sequence logos were generated. Sequence logos for the literature motif, the aligned reference and the aligned mutant sequences as well as the mutation counts at each position are shown for two representative transcription factors (CEBP and SPI1). (c) For each transcription factor, the match score to the transcription factor PWM was determined for all the factor’s mutated sites. Plotted for each transcription factor (x axis) is the mean difference in the match scores for the mutated and reference sites (y axis). Red indicates FDR < 0.05. P value for difference in mean match score was computed by two-sided paired Wilcoxon rank-sum test. (d) Breakdown of transcription factors into those having mutated sites with match scores significantly different from the reference match score and/or significantly different from the match scores of random mutations. (e) A histogram of pooled match scores for transcription factor sites with match scores for real mutations significantly worse than both the corresponding reference and random mutation match scores. (f) For each transcription factor with mutant binding sites having significantly worse match scores than the reference and random mutation match scores, the mean difference between the match scores for the mutant sites and the reference (x axis) is plotted against the mean difference in the match scores for the mutant sites and the random mutant sites (y axis). The color of the text scales with the −log10 (P value) corresponding to the significance of the difference between the real and random scores, and point size scales with the number of mutant sites.
probability (−log10) of repeated mutation of site is found mutated (are then subjected to an FDR cutoff of 0.05. Post-analysis filtering is performed to limit the number type, transcript annotation and sample ID. Post-conditioned on replication timing, base-pair model estimates the probability of mutation according to a logistic regression model. This each mutation in each cancer sample is assigned is approximated by a Poisson binomial model. more cancers containing a mutation in the region regulatory site or window, the probability of (mutation of individual sites in regulatory regions by) Shown is the probability (−log10) of repeated mutation of 10-bp windows in regulatory regions plotted against the number of times the region is found mutated. Arrows point to two known regulatory mutations in the TERT promoter.

Recurrencyt mutated specific regulatory sites We next sought to further investigate the possibility that a subset of recurrent mutations exist owing to positive selection in cancer. Previous evidence for this phenomenon comes from the identification of somatic mutations upstream of the TERT gene encoding telomerase reverse transcriptase⁶,⁶ and somatic mutations in regulatory regions of PLEKHS1, WDR74 and SDHD⁷, as well as examples of germline mutations such as those found in the CCND1 enhancer¹⁹. Previous methodologies for whole-genome analysis of recurrent mutations did not take into account sample-specific probabilities dependent on replication timing, base-pair composition and presence of annotated transcripts—covariates suggested by previous work to affect background mutation rate and model performance¹¹. Our algorithm employed logistic regression to determine sample-specific and covariate-corrected background mutation probabilities followed by a Poisson binomial model to account for patient-specific probabilities (Fig. 4a). We trained our logistic regression model on all mutations from all samples and used the resulting model to estimate the probability of mutation for any particular sample given any combination of replication timing, base-pair type and transcript region. Box plots of these predicted probabilities for all samples and all combinations of covariates showed that base-pair composition and replication timing but not transcript region strongly influence the predicted mutation probabilities (Supplementary Fig. 5a). We evaluated all mutated genomic loci for recurrent mutation of both single nucleotides and 10-bp windows around identified mutations. Mutations that were potential false positives from mapping errors or represented common SNPs were removed from the analysis, and multiple-hypothesis testing was conducted by applying a 5% FDR cutoff with the Benjamini-Hochberg procedure²⁰. The total number of regions tested for the FDR cutoff was estimated as the average total number of positions in the genome with high read coverage.

Results are plotted for both single nucleotides (Fig. 4b) and 10-bp windows (Fig. 4c). Selection of driver mutations in coding genes has been studied extensively, and we replicated a number of known findings, including the identification of selected mutations in the coding regions of TP53, AKT1, PIK3CA, PTEN, EGFR, CDKN2A and KRAS (Fig. 5). In addition to previously identified TERT promoter mutations, we found eight new mutated loci potentially regulating known cancer-associated genes, including GNAS, INPP4B, MAP2K2, BCL11B, NEDD4L, ANKRD11, TRPM2 and P2RY8.
GNAS, NEDD4L, BCL11B and ANKRD11 are recurrently mutated in cancers. TRPM2 and INPP4B are in the tumor-suppressor database. Additionally, GNAS, MAP2K2, BCL11B and P2RY8 are in the Cancer Gene Census.

Our search for recurrent regulatory mutations is related to mutations of TERT (two sites) and PLEKHS1 (one site). Of the remaining 14 mutations, 12 were in cancer types we did not analyze and 2 were present but below our statistical threshold. If we relaxed the criteria so that mutations were called by both MuTect and VarScan 2, then these two additional sites would also be called. Notably, our method was much more sensitive than the one employing a five-mutation threshold, allowing us to identify many more significantly mutated sites.

In comparison to results from Weinhold et al., our sites overlapped with mutational hotspot regions near C1orf159, TERT and PLEKHS1. Weinhold et al. used a larger set of tumors and found 193 hotspot regions in promoters and enhancers. We found mutations in our samples in 102 of these regions. Thirty-five of the regions were completely removed from our analysis when accounting for potential mapping errors, and seven additional regions had at least one mutation that was likely a mapping error. Twelve of the 102 regions had counts high enough that they might have been called significant in our analysis if all the mutations had been within 10 bp of each other; however, all but the PLEKHS1 hotspot region had a size greater than 10 bp, and thus our analysis with single nucleotides and 10-bp windows did not result in identification of these sites. In comparison to the method of Weinhold et al., we found an additional 210 significant single-base recurrently mutated sites that were not in annotated coding regions. These data suggest that our methodology provides substantial gains in both sensitivity and specificity over the method of Weinhold et al.

We additionally searched for correlations between 116 mutated regulatory loci and their associated genes, for which matched RNA sequencing (RNA-seq) data were available both for all cancers and each cancer type individually. After multiple-hypothesis correction, one site in proximity to the PLCXD1 gene was significantly correlated with RNA expression (Supplementary Fig. 5c). This analysis is severely limited by small sample size at any given site and failed to show a correlation even for TERT. In the case of TERT, the trend for increased expression with mutation was present but failed to reach significance after multiple-hypothesis correction. This result is likely due to our limited sample size and the fact that increased TERT expression is driven by mechanisms in addition to promoter mutation in cancer.

Somatic mutation alters enhancer activity

We chose to investigate the functional activity of nine new candidate regulatory mutations as well as four additional regions that were mutated in at least three samples but did not reach significance. One TERT promoter mutation was included as a positive control. We generated luciferase reporter constructs to test the enhancer activity of wild-type and mutated regulatory regions (Fig. 5a) and screened them in a lung cancer cell line. We found two regions in addition to TERT that had substantial reporter activity relative to the empty vector, and both of these regions showed decreased activity with the mutant versions (Supplementary Fig. 6). We investigated these two regions further with the lung cell line and two additional lines (bladder and esophageal). Both mutations in the chromosome 14 region decreased activity in all three cell lines, and the chromosome 19 mutation decreased activity in all but the esophageal cell line. In contrast, the mutant TERT promoter element increased activity in all cell lines (Fig. 5b). The chromosome 19 mutation is in an intron of the GP6 gene, which currently does not have a well-documented role in cancer. However, the chromosome 14 mutations are situated in a potential enhancer between SETD3 and BCL11B. BCL11B is implicated as a haploinsufficient tumor suppressor in T cell acute lymphoblastic leukemia; thus, mutations in this element may decrease the activity of BCL11B and contribute to oncogenesis. The GP6 mutated locus can bind ETS1, GAPBP1, POLR2A, SIN3A and EGR1, and the mutated site falls on the edge of an ESR1 motif. The BCL11B locus can bind IRF1, NFXA, NFXY, SP1, TBP, EGR1, MAX, FOS and POLR2A. It remains unclear which factors may have disrupted binding as a result of the mutations.

DISCUSSION

Noncoding DNA elements comprise the majority of the genome, and the data in this study highlight the importance of these regions and regulatory regions in particular in cancer. Our study shows that mutations in specific categories of regulatory sites undergo positive selection. We observe patterns of mutation that suggest that cancers select for specific mutations that destroy and/or possibly create new binding sites for certain transcription factors. Furthermore, we find evidence for the recurrent mutation of specific regulatory sites, including many in proximity to known cancer-related genes. Two of these new regulatory regions display reduced regulatory activity upon mutation.

We find that, for most cancers, somatic mutations do not globally experience purifying selection in regulatory regions. This finding is in contrast to recent evidence that regulatory regions experience purifying selection at the human population level. A major difference between selection in cancer and selection at the human population level is that, in cancer, only a single cell type is subject to selection whereas, in an organism, selection can occur via dysfunction of any cell type. Furthermore, in cancer, damaging mutations may be more tolerated owing to dysfunction in the normal apoptotic process. Previous studies, including those of Weinhold et al., show that the rate of mutation in intergenic regions is greater than in coding and regulatory regions. Our analysis suggests that the observed differences in mutation rate can be explained largely, if not entirely, by potential false positive mutations from mapping errors and by differences in mutation rate relating to base-pair type and replication timing. Nevertheless, our study is not without important caveats. Our results were generated using regulatory information collected from many cell types. Because only a fraction of annotated regulatory regions are active in any given cell type, the mixture of regulatory and non-regulatory regions may be skewing our results. Indeed, recent data from Polak et al. show that the intrachromosomal mutation rate in cancer is inversely related to the open chromatin state of the cell type of origin for each cancer type but not for other cell types. This finding is consistent with purifying selection occurring only in the subset of regulatory regions active in the cell type of origin for the cancer.

Further subclassification of regulatory mutations by transcription factor binding site identified a subset of regulatory regions that are selectively mutated likely to either destroy the site or increase the affinity of the site for transcription factor binding. Presumably, such mutations could promote cancer by repressing tumor suppressors or by activating oncogenes. Mutations could inactivate tumor-suppressor genes by removing activating sites or adding repressive sites and could activate oncogenes by removing repressive sites or adding activating sites.
Our study also highlights the potential importance of regulatory driver mutations at specific genomic loci in cancer progression. We find that, in addition to the well-known TERT promoter mutations, there exist a number of regulatory mutations that are positively selected in cancer. Our study implicates a number of regulatory regions in proximity to known cancer-associated genes. Recurrent mutations occur near GNASE, INPP4B, MAP2K2, BCL11B, NEDD4L, ANKRDL1, TRPM2 and P2RY8, which are known to be involved in cancer, as well as near genes newly implicated in cancer, such as G6P. These analyses mark an advance toward identifying recurrent functional regulatory region mutations. In the future, incorporation of additional cancer samples from a greater number of cancer types will enable a more comprehensive analysis of shared and cancer type–specific regulatory driver mutations.

It has been suggested that there are as many or more genomic regions responsible for gene regulation than genes themselves4. As such, a large number of these regions could be mutated in cancer. This regulatory redundancy may protect against selection for deleterious mutations in the regulatory regions of some tumor suppressors but should also increase the amount of regulatory sequence that could have a functional impact if mutated. The presence of multiple regions upstream of TERT that are implicated in cancer and the fact that any one gene can have numerous different enhancers26 support this hypothesis. Overall, we expect that many regulatory regions will prove to have important roles in cancer, and the approaches and information employed in this study thus represent a significant advance in the analysis of such regions.

URLs. ENCODE 50-mer alignability tracks, http://genome.ucsc. edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability; ENCODE Replication Timing, http://hgdownload.cse.ucsc.edu/goldenPath/ hg19/encodeDCC/wgEncodeUwRepSeq/; The Cancer Genome Atlas (TCGA), http://cancergenome.nih.gov/; GitHub site for RecurrentMutationStats code, https://github.com/collinnelmont/ RecurrentMutationStats/; Cancer Genomics Hub (CGHub), https:// cghub.ucsc.edu/; dbSNP, http://www.ncbi.nlm.nih.gov/SNP/; UCSC Genome Browser Table Browser, https://genome.ucsc.edu/cgi-bin/ hgTables; RegulomeDB, http://regulome.stanford.edu/; R Matching package, http://cran.r-project.org/web/packages/Matching/index.html.

METHODS

Methods and any associated references are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank members of the Snyder laboratory for critical reading of the manuscript. We would also like to thank A. Boyle for helpful discussion regarding the use of RegulomeDB. Finally, we would like to thank members of the ENCODE Consortium for their helpful scientific feedback during the course of this work. C.M. was supported by the Stanford Biomedical Informatics Training program and funds from the US National Institutes of Health (US NIH; 1K99CA191093). D.V.S. was supported by US NIH/National Human Genome Research Institute grant T32HG000044 and the Genentech Graduate Fellowship. This work was supported by funds to C.M. (US NIH, 1K99CA191093-01) and M.S. (US NIH, SU54HG06996-04).

AUTHOR CONTRIBUTIONS

J.A.R. and D.V.S. contributed to experimental design, execution and analysis for the experiments in Figures 5 and Supplementary Figure 6. C.M. contributed to all other figures. C.M. and M.S. conceived the experiments, analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/ reprints/index.html.

1. Hoyert, D.L. & Xu, J. Deaths: preliminary data for 2011. Natl. Vital Stat. Rep. 61, 1–51 (2012).
2. Howlander, N. et al. SEER Cancer Statistics Review, 1975–2010 http://seer.cancer. gov/csr/1975_2010/ (2013).
3. Tamborero, D. et al. Comprehensive identification of mutational cancer driver genes in 12 tumor types. Sci. Rep. 3, 2650 (2013).
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
5. Huang, F.W. et al. Highly recurrent TERT promoter mutations in human melanoma. Nature 465, 1160–1165 (2012).
6. Vinagre, J. et al. Frequency of TERT promoter mutations in human cancers. Nat. Commun. 4, 2185 (2013).
7. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. 46, 741–749 (2014).
8. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat. Genet. 46, 1258–1303 (2014).
9. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
10. Boyle, A.P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22, 1790–1797 (2012).
11. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).
12. Lee, W. et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465, 473–477 (2010).
13. Ostrow, S.L., Barshir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer evolution is associated with pervasive positive selection on globally expressed genes. PLoS Genet. 10, e1004239 (2014).
14. Huber, R., Pietzsch, D., Panterotti, T. & Brand, K. Regulation of C/EBPβ and resulting functions in cells of the monocytic lineage. Cell. Signal. 24, 1287–1296 (2012).
15. Miyazawa, K., Mori, A., Yamamoto, K. & Okudaira, H. Transcriptional roles of CCAT/enhancer binding protein-β, nuclear factor-κB, and C-promoter binding factor 1 in interleukin IL-1β-induced IL-6 synthesis by human rheumatoid fibroblast-like synoviocytes. J. Biol. Chem. 13, 7620–7627 (1998).
16. Zheng, R. & Blobel, G.A. GATA transcription factors and cancer. Genes Cancer 1, 1178–1184 (2010).
17. Kannan, M.B., Solovieva, V. & Blank, V. The small MAF transcription factors MAFF, MAFK and MAFG and MAFK: current knowledge and perspectives. Biochim. Biophys. Acta 1823, 1841–1846 (2012).
18. Shaulian, E. AP-1—the Jun proteins: oncogenes or tumor suppressors in disguise? Cell. Signal. 22, 894–899 (2010).
19. Schödel, J. et al. Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. Nat. Genet. 44, 420–425 (2012).
20. Benjamin, X. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B 57, 289–300 (1995).
21. Zhao, M., Sun, J. & Zhao, Z. TSGene: a web resource for tumor suppressor genes. Nucleic Acids Res. 41, D90–D97 (2013).
22. Futreal, P.A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).
23. Van Vlierbergh, P. & Ferrando, A. The molecular basis of T cell acute lymphoblastic leukemia. J. Clin. Invest. 122, 3398–3406 (2012).
24. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342, 1235587 (2013).
25. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518, 360–364 (2015).
26. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98 (2012).
ONLINE METHODS

Mutation calling. Access to cancer and normal whole-genome sequencing data was obtained from the TCGA Project via the database of Genotypes and Phenotypes (dbGaP), and raw data were downloaded prealigned in BAM file format from the Cancer Genomics Hub (CGHub). All aligned BAM files were subjected to paired local realignment via RealignerTargetCreator and IndelRealigner in GATK. Base recalibration in the realigned BAM files was then performed using BaseRecalibrator in GATK. Point mutations were called with default settings using MuTect and the VarScan 2 Somatic caller. The intersection of the calls from MuTect and VarScan 2 was used for further analysis.

Regulatory annotations. Mutations were assigned regulatory annotations via RegulomeDB. Regulatory mutations were considered to be any mutation in RegulomeDB categories 1–5, which include all regions that minimally contain a transcription factor binding peak or a DNase I binding peak identified from experimental data.

Transcription factor binding site analysis. PWMs for transcription factors were obtained from Jaspar and Transfac. R code was written to align experimental data.

RegulomeDB categories 1–5, which include all regions that minimally contain a transcription factor binding site, were considered. Random mutations were obtained from Jaspar and Transfac. R code was written to align transcription factor binding site analysis. A score between 0 and 100 was given on the basis of the length of the genomic region matched the called variant allele. A 201-bp window was chosen because it should be sufficiently long to cover all potential overlapping reads, as mapped read sizes are typically smaller than 100 bp. For the analysis using 10-bp windows, a regional score was generated by averaging the scores of all the 10-bp windows, a regional score was generated by averaging the scores of all the

Replication timing. ENCODE replication timing data were downloaded from the UCSC Genome Browser. The average waveform from the following cell types was used to associate genomic regions with replication timing: BJ, GM06990, GM12801, GM12812, GM12813, GM12878, HeLa-S3, HepG2, HUVEC, IMR-90, K-562, MCF-7, NHEK and SK-N-SH. These averages were then assigned to bins from 1 to 100 to discretize the data.

Statistical modeling of recurrent mutations. All unique chromosomal mutations were tested for statistical significance. Let n be the total number of individuals and k be the number of individuals with a mutation in a particular region. Given a region of interest, we sought to estimate the probability of k or more mutations in n individuals. We used a Poisson binomial model, which differs from the binomial model in that each tumor is assigned its own probability of mutation. Formally, the probability we calculate is as follows:

\[
\Pr(K \geq k) = \sum_{l=k}^{n} \prod_{j=1}^{l} (1 - p_j)
\]

where \( F_l \) is the set of all subsets of k integers that can be selected from \{1, 2, ..., n\}, n is the number of tumors, k is the number of tumors with the region mutated, \( p_j \) is the probability that site i is mutated, A is a single set of k integers that can be selected from \{1, 2, ..., n\} and \( A^\prime \) is the complement of this set.

In practice, we used an approximation for the Poisson binomial in the poibin R package. Each individual’s unique mutation probability for a given region was determined by fitting a logistic regression model to all data with the following features: patient ID, replication timing bin, base pair (G•C versus A•T), and presence and type (none, coding exon, noncoding exon, coding intron or noncoding intron) of annotated transcripts in the region. Interactions between each patient ID and replication timing, as well as between patient ID and base-pair type, were included in the model. These select interactions were included to increase the fit of the model. For regions, as opposed to single sites, the regional probability was determined as follows:

\[
\Pr(\text{site is mutated}) = 1 - \prod_{i=1}^{l} (1 - p_i)
\]

where \( i \) is the base position within a site of length \( l \) and \( p_i \) is the probability that base \( i \) is mutated.

Filtering out false positives from mapping errors and SNPs. SNPs from dbSNP Build 141 were downloaded from the UCSC Genome Browser Table Browser. Called mutations that had the same chromosomal position and variant allele as a common SNP were filtered out. Predicted mapping errors were determined by querying BLAT with a 201-bp region centered on the genomic position of the variant. Notably, the variant allele was used in place of the reference allele for this analysis. A score between 0 and 100 was given on the basis of the length of the longest aligned region for a given BLAT result that included a match of up to 100 bp in length to the reference genome such that the reference allele for the matched genomic region matched the called variant allele. A 201-bp window was chosen because it should be sufficiently long to cover all potential overlapping reads, as mapped read sizes are typically smaller than 100 bp. For the analysis using 10-bp windows, a regional score was generated by averaging the scores of all the mutations contained within the region. Regions with an average score of greater than 50 were filtered out as potential false positives. The 1000 Genomes Project phase37/35 reference was used for BLAT searches. The analysis in Supplementary Figure 1 of overlap between the filtered-out regions and difficult-to-align regions of the genome was performed using 50-mer alignability tracks. Any mutation with a score of 0.5 or less was considered difficult to align.

Luciferase reporter assays. To examine the effects of recurrent somatic mutations on transcription, wild-type and mutant regions of 201 bp in length and centered on the position of each mutation were synthesized and cloned into the KpnI and Nhel sites of the PGL4.23[luc2p/min] luciferase reporter construct (Promega). Lung adenocarcinoma (NCI-H1437), esophageal adenocarcinoma

doi:10.1038/ng.3332
KYSE-450 and bladder carcinoma (Ku-19-19) cells, growing in 96-well plates, were transfected in quadruplicate with 200 ng of the pGL4.23 reporter construct and 40 ng of the pRL-CMV Renilla control plasmid (Promega). Forty-eight hours after transfection, luciferase activity was measured using the Dual-Glo Luciferase Assay System (Promega). Statistically significant differences in relative luciferase activity between wild-type and mutant regions were determined using a two-sided Student's t test, assuming equal variance. Visually, the variances appeared equal in the tested data and there was no biological reason that the variance should be different. Cell lines were obtained from the American Type Culture Collection (ATCC) and tested negative for mycoplasma.

**Code availability.** Code for the identification of recurrently mutated genomic sites is available through GitHub (see URLs).

27. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

28. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

29. Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).

30. Maltys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).

31. Kel, A.E. et al. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**, 3576–3579 (2003).

32. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

33. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).

34. Hansen, R.S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139–144 (2010).

35. Hong, Y. On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.* **59**, 41–51 (2013).

36. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).