HUNDREDS OF PUTATIVE NON-CODING CIS-REGULATORY DRIVERS IN CHRONIC LYMPHOCYTIC LEUKAEMIA AND SKIN CANCER

Halit Ongen[1,2,3*], Olivier Delaneau[1,2,3], Michael W. Stevens[1,2,3], Cedric Howald[1,2,3], Emmanouil T. Dermitzakis[1,2,3*]

1 Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
2 Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland
3 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
* Corresponding authors

SUMMARY

Perturbations of the coding genome and their role in cancer development have been studied extensively. However the non-coding genome's contribution is poorly understood (Khurana et al., 2016), due to not only the difficulty of defining the non-coding regulatory regions and the genes they regulate, but also the limited power arising from the regulatory regions' small size. In this study, we try to resolve this issue by defining modules of coordinated non-coding regulatory regions of genes (cis regulatory domains or CRDs) using the correlation between histone modifications, assayed by ChIP-seq, in immortalized B-cells (317 samples) and skin fibroblasts (78 samples). We search for CRDs that accumulate an excess of somatic mutations in chronic lymphocytic leukaemia (CLL) and skin cancer, which affect these cell types. At 5% FDR, we find 149 CRDs with significant excess somatic of mutations in CLL, 92 of which regulate 163 genes, and in skin cancer, 465 significant CRDs 142 of which regulate 187 genes. The genes these CRDs regulate include ones involved in tumorigenesis, and are enriched in pathways already implicated in the respective cancers, like the B-cell receptor signalling pathway in CLL and the Ras/Ref/MAPK signalling pathway in skin cancer. We discover that the somatic mutations in the significant CRDs of CLL are hitting bases more likely to be functional than the mutations in non-significant CRDs, and in both cancers there is a significant deviation from the standard mutational signatures observed in the significant CRDs vs. their null sequences. Both results indicate selection acting on these CRDs during tumorigenesis. Finally, we find that the transcription factor biding sites that are disturbed by the somatic mutations in significant CRDs are enriched for factors known to be involved in cancer development. In conclusion,

we are describing a new powerful approach to discover non-coding regulatory somatic mutations likely driving tumorigenesis in CLL and skin cancer, and our approach could be applied to other cancers to find this class of underexplored drivers.

INTRODUCTION

The Cancer Genome Atlas Consortium (TCGA) (Cancer Genome Atlas Research, 2008) and International Cancer Genome Consortium (ICGC) (International Cancer Genome et al., 2010) have so far been great resources for both the generation and the characterization of genomics data from a variety of human cancer types. By investigating somatic mutations in the protein coding parts of the genome, they have identified many genes as putative driver genes across multiple cancers (Kandoth et al., 2013), as well as many known and novel putative drivers in specific cancers (Cancer Genome Atlas, 2012, 2015a, b; Cancer Genome Atlas Research, 2012, 2013, 2014a, b). However, with the whole genome sequencing (WGS) data generated by these projects, it became evident that the majority of the somatic mutations observed lie in the non-coding portion of the human genome (Khurana et al., 2016). One important question that follows from this observation is whether these non-coding somatic mutations are involved in tumorigenesis, or whether they are simply passenger mutations that are not contributing to tumorigenesis and are therefore not under positive selection in cancer development.

As mentioned, while the coding genome's impact on tumorigenesis is heavily scrutinized by searching for putative driver genes that accumulate an excess of protein altering mutations, the contribution of the non-coding genome to cancer development is less well understood, since interpreting the non-coding genome remains challenging (Khurana et al., 2016). There are examples where the non-coding regulatory mutations are involved in tumorigenesis like the *TERT* promoter in melanoma (Horn et al., 2013; Huang et al., 2013) and in other cancers (Vinagre et al., 2013), and recurrent non-coding mutations in chronic lymphocytic leukaemia (CLL) (Puente et al., 2015) and breast cancer (Rheinbay et al., 2017). However, these studies have each identified a small number of

regulatory regions mainly due to power issues, as regulatory elements are individually studied and tend to be small in size. We have also previously shown that non-coding germline variation could be involved in tumorigenesis, and identified genes with putative somatic regulatory drivers in colorectal cancer, using perturbations in allele specific expression during tumour development (Ongen et al., 2014). However, directly finding non-coding regulatory regions that accumulate an excess of somatic mutations using WGS data is an unsolved problem due to the difficulty in defining the bounds of the non-coding regulatory regions, assessing which genes these regulate, identifying the null genome to compare, and failure to accumulate signal over multiple regulatory regions thus decreasing power in detecting these regions. In this study, we address this by identifying cis regulatory domains (CRDs), sets of coordinated non-coding regulatory regions, using the correlation amongst three histone modifications and finding the genes that are regulated by these CRDs; followed by testing whether these accumulate an excess of somatic mutations by controlling for differential somatic mutation rate across the genome and between open and closed chromatin regions.

RESULTS

Identification of cis regulatory domains

In order to define regions of the non-coding genome that are likely to be regulatory we have used data generated by the SysGenetiX (SGX) project (Delaneau et al., 2017), which comprises genotypes, H3K27ac, H3K4me1, H3K4me3 histone modifications assayed by ChIP-seq, and transcriptomes assayed by RNA-seq from 317 lymphoblastoid cell line (LCL, immortalized B-cells) and 78 skin fibroblast cell lines. With this dataset, we can use the correlation between the chromatin marks in the population to identify likely regulatory regions of the genome in a cell type specific manner. To achieve this, we computed all the pairwise correlations between nearby chromatin peaks, which were grouped into a tree using hierarchical clustering. We defined cis regulatory domains (CDRs) as nodes on the tree where the mean correlation of the leaves the node is double the background correlation and find 21607 and 18261 CRDs in LCLs and skin fibroblasts, respectively (these counts

are higher than the original SGX study, since we are also considering CRDs comprised of overlapping peaks). Furthermore, using the resulting CRD quantification and gene expression quantifications we were able to discover 15161 and 3652 nearby genes (FDR = 5%, 1 Mb cis window) that the CRDs regulate in LCLs and fibroblasts, respectively (**Supplementary methods & Supplementary figure 1**). By finding CRDs and genes that covary in the population, we can estimate the non-coding regulatory regions of genes, which will not only enable identification of non-coding regulatory regions with excess somatic mutations but also link these regions to specific genes.

Cis regulatory domains with excess somatic mutations

Next, we wanted to devise a method to detect excess somatic mutations in CRDs, which would signify positive selection in tumorigenesis and hence indicate driving potential. To do so we first found the regulatory regions of the CRDs identified in SGX project that do not overlap with known exons. Subsequently, we created a set of regions that did not overlap any of the ChIP-seq peaks in SGX or known exons, which make up non-exonic non-regulatory null regions that we call spacers. In order to account for differential somatic mutation rates observed across the genome (Hodgkinson et al., 2012; Liu et al., 2013; Pleasance et al., 2010; Woo and Li, 2012), we took the spacers that are in between the regulatory regions of a CRD and the two flanking ones as the null for a given CRD. We know that different cancers have different mutational signatures, i.e. the types of somatic mutations they accumulate (Alexandrov et al., 2013). In order to incorporate this into our method, we considered the local context of the somatic mutations observed by taking into account the immediate 5' and 3' reference bases (trinucleotide context) flanking the somatic mutation position. Thus, we counted how many of the 64 possible context triplets are present in the spacer regions of a CRD and how many of them harbour a somatic mutation, to calculate an expected mutation rate for each of the 64 classes. We then counted the number of times these triplets are observed in the regulatory regions of a CRD and using the expected mutation rates for each of the 64 classes, we found the total number of expected mutations. Subsequently we counted the observed number of

mutations in the regulatory regions and using a one-tailed Poisson test calculated a nominal p-value for the excess of somatic mutations in a CRD. Lastly, we wanted to account for differential mutation rates and other unknown technical biases between open (regulatory regions) and closed chromatin (spacers) (Schuster-Bockler and Lehner, 2012). Thus, we devised a permutation scheme, in which for a triplet context in the regulatory region of a CRD we select a random position with the same context in a random CRD and its corresponding downstream spacer and ask if these positions harbour somatic mutations. We did this for all of the regulatory bases of a CRD, to generate a pseudo CRD, which conserves the mutation rates in open vs. closed chromatin, but breaks the clustering of the somatic mutations. Then, we iterate this process 10000 times for each of the CRDs and at each iteration check if the nominal p-value of the pseudo CRD is as significant as or more significant than the observed p-value, to come up with our adjusted p-value (**Supplementary methods, Supplementary figure 2 & 3**). Using this methodology, we control for the differential mutation rate across the genome, the local context of the types of somatic mutations, and the different rates of somatic mutations inferences between open and closed chromatin, therefore resulting in a robust estimate of the significance of excess somatic mutations in CRDs.

We aimed to find CRDs with significant excess of somatic mutations with the aforementioned methodology using publicly available somatic mutation calls. To this end, we acquired somatic mutation calls from WGS data for CLL (Puente et al., 2015) and through the International Cancer Genome Consortium data portal for skin cancer (project codes: MELA-AU and SKCA-BR) (Zhang et al., 2011). These cancers were chosen because they affect the cell types analysed in the SGX project. In CLL we found 149 CRDs that accumulate significantly more somatic mutations (FDR = 5%), 92 of which regulate 163 genes, and in skin cancer there were 465 significant CRDs (FDR = 5%), 142 of which regulate 187 genes (**Figure 1**, **Supplementary tables 1 & 2**). Due to limitations in power, not all modules are assigned to genes, which is more apparent in the smaller sample sized skin fibroblasts.

One possible technical issue that might affect our results is if the regulatory regions of the significant CRDs had systematically higher coverage in WGS than their spacers, then this would cause more somatic mutations to be called in regulatory regions, which would create false positives. In order to assess this effect, we compared the number of reads overlapping the somatic mutation calls in the regulatory regions of the significant CRD to their spacers. In both cancer types, there was no significant increase in coverage of somatic mutations in regulatory region compared to their spacers (**Supplementary figure 9**). In fact, the coverage in the regulatory regions is less than the spacers in both cancers (median 21 reads in regulatory vs. 31 in spacer in CLL and 60 vs. 61 in skin cancer). We also investigated whether inclusion of different spacers, e.g. including or excluding flanking spacers, is having an effect on our results, thus we reran the analysis without the flanking regions. We observed that the p-values are highly significantly positively correlated between the two analyses (rho = 0.826, $p < 2e-16$ in CLL and rho = 533, $p = 1.5e-13$ in skin cancer). Skin cancer showed a higher difference between the two analyses but in almost all cases inclusion of the flanking spacers made the p-value less significant (**Supplementary figure 10**). These results indicate that our permutation scheme accounts for any heterogeneity of coverage and variant call confidence and that our results hold with different ranges of spacer sequence.

We examined the per sample somatic mutation rates of the significant CRDs, and observed that similar to the protein coding drivers there are CRDs that are highly mutated in most samples, and others exhibiting high mutation rates in just few samples (**Supplementary figures 11 & 12**). This suggests that the patterns of selection in non-coding cis-regulatory regions are similar to protein coding sequences with some regulatory regions probably important for tumorigenesis in most or all cancers while others being more specific.

Among the genes identified in CLL are *BIRC3* (**Supplementary figure 4**), a putative driver previously identified due to excess protein altering mutations (Landau et al., 2015), *BCR*, which is one of the breakpoints of the Philadelphia chromosome first identified in chronic myeloid leukaemia (Salesse

and Verfaillie, 2002), *LPP*, where the CRD contains a variant that confers predisposition to CLL (Berndt et al., 2016). Overall the genes identified in CLL are enriched for B-cell receptor signalling pathway and its downstream pathways, one of the major pathways involved in CLL development (Stevenson et al., 2011), and pathways involved in the tumorigenesis of other types of cancers (**Supplementary table 3**). Seven of the significant CLL CRDs are in the hypermutated *IGHV* locus, however we show below that these are unlikely to be false positives. In skin cancer we find CRDs for genes such as *LRRC37A3* (**Supplementary figure 5**), *NOTCH2NL*, and *TMEM200A*, which have been previously identified as putative drivers in melanoma due to an excess of protein altering mutations (Cancer Genome Atlas, 2015b). Collectively these genes are enriched for Ras/Ref/MAPK signalling pathway (**Supplementary table 4**), an important pathway in skin cancer development (Fecher et al., 2008). We show that by using our methodology we can identify putative non-coding regulatory drivers for genes that are known to be involved in cancer development as well many novel ones.

Comparison of significant CRDs vs. non-significant ones in both cancer types

There are several patterns that are different between the two cancer types. First, due to the sample size differences between the cell types of the SGX project where skin fibroblasts have about four times fewer samples, there is differential power in both discovering CRDs and assigning them to genes. Furthermore, for CLL we are directly assessing CRDs from the cell type that is giving rise to the cancer, whereas in skin cancer we are using the skin fibroblasts as the closest proxy available to the underlying cell type involved in tumorigenesis. Also, the underlying somatic mutation rate in skin cancer is significantly higher than CLL. All these reasons individually and combined result in differences in our findings which we tried to highlight by comparing the significant CRDs to the non-significant ones in both cancer types.

We observed that the average correlation between the chromatin peaks in significant CRDs is significantly higher than the correlation among the peaks in non-significant CRDs in both cancer types (**Figure 2a**), indicating that the CRDs identified as cancer drivers indeed tend to have strong

coordinated behaviour and therefore the accumulation of mutations is justified. In CLL the significant modules are significantly shorter, contain fewer chromatin peaks, regulate the expression of fewer genes that are further away from the CRD than in non-significant modules, highlighting again the discovery of strongly coordinated CRDs (**Figure 2b,c,d,e,f**). These patterns are not observed in skin cancer. The main reasons for this are on the one hand the smaller sample size in fibroblasts that leads to fewer gene assignments and the increased mutation rate in skin cancer, which leads to signal dilution.

Functional assessment of somatic mutations in the significant CRDs

We wanted to assess whether the somatic mutations found in the significant CRDs are perturbing likely functional nucleotide bases. In order to do so we have acquired the probabilities for purifying selection acting on the bases of the human genome as calculated by the LINSIGHT method (Huang et al., 2017). Subsequently we compared the purifying selection acting on the positions of the somatic mutations in the significant CRDs vs. the positions in the non-significant CRDs (**Supplementary methods**). We find that, in CLL, somatic mutations in significant CRDs are hitting bases that are significantly (Mann-Whitney $p < 2.2e-16$) more likely to be under purifying selection than the mutated bases in non-significant CRDs, but this is not the case for skin cancer (**Figure 3**). We further investigated functional impact of the somatic mutations in the 7 significant CRDs overlapping the hypermutated *IGHV* locus in CLL (Landau et al., 2015). We found that the somatic mutations in the regulatory regions of these 7 CRDs are also hitting bases that are significantly more likely to be functional than both the mutated bases in non-significant CRDs (Mann-Whitney $p < 2.2e-16$) and the non-significant CRDs in the *IGHV* locus (Mann-Whitney $p = 2.3e-6$, **Supplementary figure 6**). The CLL result is indicative of selection in tumorigenesis, even in the hypermutated region, and the result for skin cancer, which on average has ~29 times more somatic mutations than CLL (Mann-Whitney $p < 2.2e-16$, **Supplementary figure 7**), is most likely due to high number of somatic mutations diluting the relevant functional mutations.

In order to further investigate the selective pressure acting on the significant CRDs during tumorigenesis, we compared the types of somatic mutations observed between the regulatory regions and spacers in CRDs. To this end, we classified the single nucleotide somatic mutations into 96 classes representing six possible substitutions of the pyrimidine of the mutated Watson–Crick base pair with incorporating information on the immediate flanking bases (**Supplementary methods**) (Alexandrov et al., 2013). If there is selective pressure on the significant CRDs, then we would expect the mutational signatures to differ between their regulatory regions and spacers, which we expect not to be under selection. In both cancer types we observed that there were significant differences in the mutation signatures in regulatory vs. spacer regions in significant CRDs (**Figure 4**). Furthermore, in skin cancer there was a significant decrease (Fisher's exact test, OR= 0.78, $p < 2.2e-16$) in the C to T mutations, which are due to damage by U.V. light, in the regulatory regions vs. spacers of the significant CRDs. We also observed the same type of significant differences in the mutational signatures between the regulatory regions of the significant CRDs vs. the regulatory regions of the non-significant CRDs (**Supplementary figure 8**). These results signify that during tumorigenesis there is selection acting on the CRDs we identify, agnostic to the underlying mutation rate, verifying that the signal of excess somatic mutations is not due to hypermutation and corroborating their driver potential.

Lastly, we wanted to examine whether there are specific transcription factor binding sites (TFBSs) perturbed by the somatic mutations in the regulatory regions of the significant CRDs. Thus, we used sequences from regions flanking the somatic mutations by 15 bp (total of 31 bp) in the regulatory regions of significant and non-significant CRDs. Then using the HOMER software package (Heinz et al., 2010) we searched for TFBSs enriched around somatic mutations in significant CRDs compared to the non-significant CRDs (**Supplementary methods**). We find that in both CLL and skin cancer, transcription factors known to be involved in tumorigenesis are being perturbed, e.g. in CLL p53 (Le Garff-Tavernier et al., 2011), NFKB (Furman et al., 2000), E2A (Kardava et al., 2011; Saborit-Villarroya et al., 2011), SCL (Crans and Sakamoto, 2001), and in skin cancer PAX6 (Li and Eccles, 2012) and CEBP

(Anand et al., 2014; Koschmieder et al., 2009) (**Figure 5**). Therefore, with this analyses we identify changes in transcription factor binding sites as a likely mechanistic cause for driver potential exhibited by these CRDs.

DISCUSSION

In this study, we describe a methodology to identify non-coding cis-regulatory drivers in cancer in an unbiased and genome-wide manner using the clusters of correlated regulatory elements to increase power. We show that this method can find likely non-coding drivers in CLL and skin cancer, cancers affecting cell types analysed in SGX project. This manuscript highlights the importance of population scale investigations of the non-coding genome; since studies like the SGX project are one of the most powerful ways of finding clusters of non-coding regions in the genome that directly affect gene expression. We confirm our previous finding that non-coding drivers are important players in tumorigenesis (Ongen et al., 2014) and that the number of genes involved is much higher than currently known for protein coding drivers. As the datasets like the SGX project are extended to more cell types, we will be able to find these types of non-coding drivers, which are currently under-studied due to the difficulty in interpreting the non-coding genome, in other cancers. We believe there should be a significant shift in focus to identify non-coding drivers in all cancer types, and this study describes a powerful methodology to do so. Our results with hundreds of cis regulatory regions having driver potential overall support a model under which cancer tumorigenesis and progression is a complex phenomenon, where hundreds of driver mutations with different effect sizes, coming both from the somatic and the germ-line genome both coding and non-coding, play an important role in the process. Under this model the well-known somatic driver mutations in key genes are the necessary components for tumorigenesis and define many of common properties of tumours and the larger number of driver mutations define the specific characteristics of each tumour. It is likely that the specific progression characteristics as well as the response of such tumours to specific therapies and our immune system depend to these larger sets of mutations in

much the same way they do in complex diseases and it provides a strong rationale to explore the non-coding cancer genome more deeply.

AUTHOR CONTRIBUTIONS

H.O and E.T.D designed the study. H.O. analysed the data and H.O and E.T.D interpreted the results. O.D. generated the SGX data, M.W.S and C.H. were involved in data analysis. H.O. wrote the manuscript and E.T.D edited it.

ACKNOWLEDGMENTS

This research was supported by grants from NIH-NIHM (GTEx), European Commission FP7, European Research Council, and Swiss National Science Foundation. The computations were performed at the Vital-IT Centre for high performance computing of the SIB Swiss Institute of Bioinformatics. We thank all members of the Sysgenetix for their crucial input in data generation and analyses.

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L.*, et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.
Anand, S., Ebner, J., Warren, C.B., Raam, M.S., Piliang, M., Billings, S.D., and Maytin, E.V. (2014). C/EBP transcription factors in human squamous cell carcinoma: selective changes in expression of isoforms correlate with the neoplastic state. PLoS One *9*, e112073.
Berndt, S.I., Camp, N.J., Skibola, C.F., Vijai, J., Wang, Z., Gu, J., Nieters, A., Kelly, R.S., Smedby, K.E., Monnereau, A.*, et al.* (2016). Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. Nature communications *7*, 10933.
Cancer Genome Atlas, N. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.
Cancer Genome Atlas, N. (2015a). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature *517*, 576-582.
Cancer Genome Atlas, N. (2015b). Genomic Classification of Cutaneous Melanoma. Cell *161*, 1681-1696.
Cancer Genome Atlas Research, N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068.
Cancer Genome Atlas Research, N. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.
Cancer Genome Atlas Research, N. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43-49.
Cancer Genome Atlas Research, N. (2014a). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202-209.

Cancer Genome Atlas Research, N. (2014b). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315-322.
Crans, H.N., and Sakamoto, K.M. (2001). Transcription factors and translocations in lymphoid and myeloid leukemia. Leukemia *15*, 313-331.
Delaneau, O., Zazhytska, M., Borel, C., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D., Ambrosini, G., Bielser, D.*, et al.* (2017). Intra- and inter-chromosomal chromatin interactions mediate genetic effects on regulatory networks. bioRxiv.
Fecher, L.A., Amaravadi, R.K., and Flaherty, K.T. (2008). The MAPK pathway in melanoma. Curr Opin Oncol *20*, 183-189.
Furman, R.R., Asgary, Z., Mascarenhas, J.O., Liou, H.C., and Schattner, E.J. (2000). Modulation of NF-kappa B activity and apoptosis in chronic lymphocytic leukemia B cells. J Immunol *164*, 2200-2206.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.
Hodgkinson, A., Chen, Y., and Eyre-Walker, A. (2012). The large-scale distribution of somatic mutations in cancer genomes. Human mutation *33*, 136-143.
Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K.*, et al.* (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959-961.
Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957-959.
Huang, Y.F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet *49*, 618-624.
International Cancer Genome, C., Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I.*, et al.* (2010). International network of cancer genome projects. Nature *464*, 993-998.
Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.
Kardava, L., Yang, Q., St Leger, A., Foon, K.A., Lentzsch, S., Vallejo, A.N., Milcarek, C., and Borghesi, L. (2011). The B lineage transcription factor E2A regulates apoptosis in chronic lymphocytic leukemia (CLL) cells. Int Immunol *23*, 375-384.
Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet *17*, 93-108.
Koschmieder, S., Halmos, B., Levantini, E., and Tenen, D.G. (2009). Dysregulation of the C/EBPalpha differentiation pathway in human cancer. J Clin Oncol *27*, 619-628.
Landau, D.A., Tausch, E., Taylor-Weiner, A.N., Stewart, C., Reiter, J.G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Bottcher, S.*, et al.* (2015). Mutations driving CLL and their evolution in progression and relapse. Nature *526*, 525-530.
Le Garff-Tavernier, M., Blons, H., Nguyen-Khac, F., Pannetier, M., Brissard, M., Gueguen, S., Jacob, F., Ysebaert, L., Susin, S.A., and Merle-Beral, H. (2011). Functional assessment of p53 in chronic lymphocytic leukemia. Blood Cancer J *1*, e5.
Li, C.G., and Eccles, M.R. (2012). PAX Genes in Cancer; Friends or Foes? Front Genet *3*, 6.
Liu, L., De, S., and Michor, F. (2013). DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nature communications *4*, 1502.
Ongen, H., Andersen, C.L., Bramsen, J.B., Oster, B., Rasmussen, M.H., Ferreira, P.G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A.*, et al.* (2014). Putative cis-regulatory drivers in colorectal cancer. Nature *512*, 87-90.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R.*, et al.* (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.
Puente, X.S., Bea, S., Valdes-Mas, R., Villamor, N., Gutierrez-Abril, J., Martin-Subero, J.I., Munar, M., Rubio-Perez, C., Jares, P., Aymerich, M.*, et al.* (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature *526*, 519-524.
Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M.*, et al.* (2017). Recurrent and functional regulatory mutations in breast cancer. Nature *547*, 55-60.
Saborit-Villarroya, I., Vaisitti, T., Rossi, D., D'Arena, G., Gaidano, G., Malavasi, F., and Deaglio, S. (2011). E2A is a transcriptional regulator of CD38 expression in chronic lymphocytic leukemia. Leukemia *25*, 479-488.
Salesse, S., and Verfaillie, C.M. (2002). BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. Oncogene *21*, 8547-8559.
Schuster-Bockler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504-507.
Stevenson, F.K., Krysov, S., Davies, A.J., Steele, A.J., and Packham, G. (2011). B-cell receptor signaling in chronic lymphocytic leukemia. Blood *118*, 4313-4320.
Vinagre, J., Almeida, A., Populo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L.*, et al.* (2013). Frequency of TERT promoter mutations in human cancers. Nature communications *4*, 2185.
Woo, Y.H., and Li, W.H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nature communications *3*, 1004.
Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B.*, et al.* (2011). International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database : the journal of biological databases and curation *2011*, bar026.

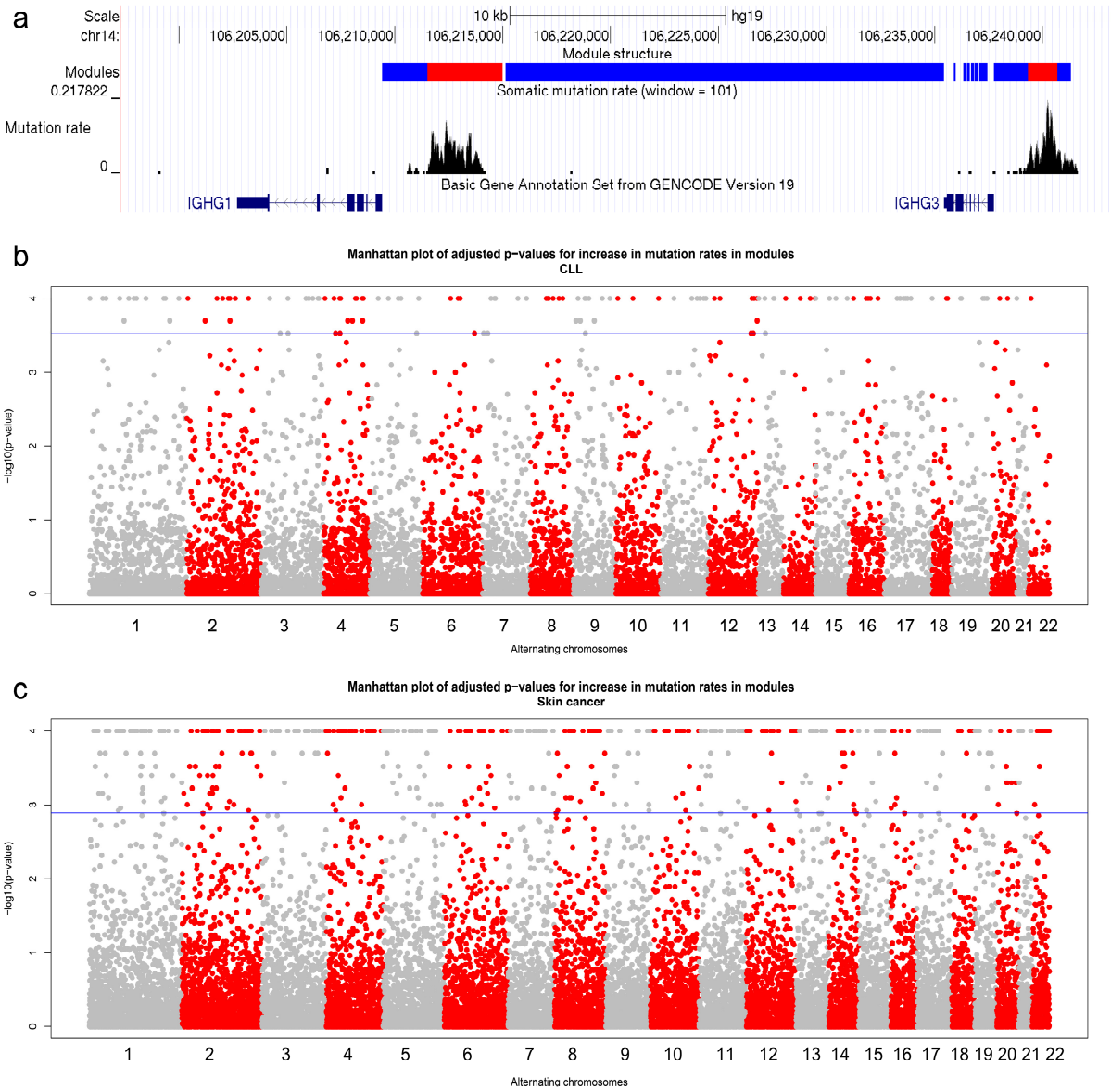

Figure 1 – (a) An example of a CRD that has significantly excess somatic mutations in CLL, which regulates the expression of the *IGHG3* gene. The blue red boxes are the positions of the CHiP-seq peaks of the CRD, i.e. the regulatory regions of this CRD, and the blue boxes are the spacers. The mutation rate averaged over a sliding window of 101 bp is shown as black bars. The gene structure in the region is represented as dark blue boxes. Plot is generated using the UCSC genome browser. Manhattan plot of the permuted p-values for increase in mutation rates in CRDs for (b) CLL and (c) skin cancer. The blue line indicates the 5% FDR threshold. Alternating chromosomes are coloured differently.

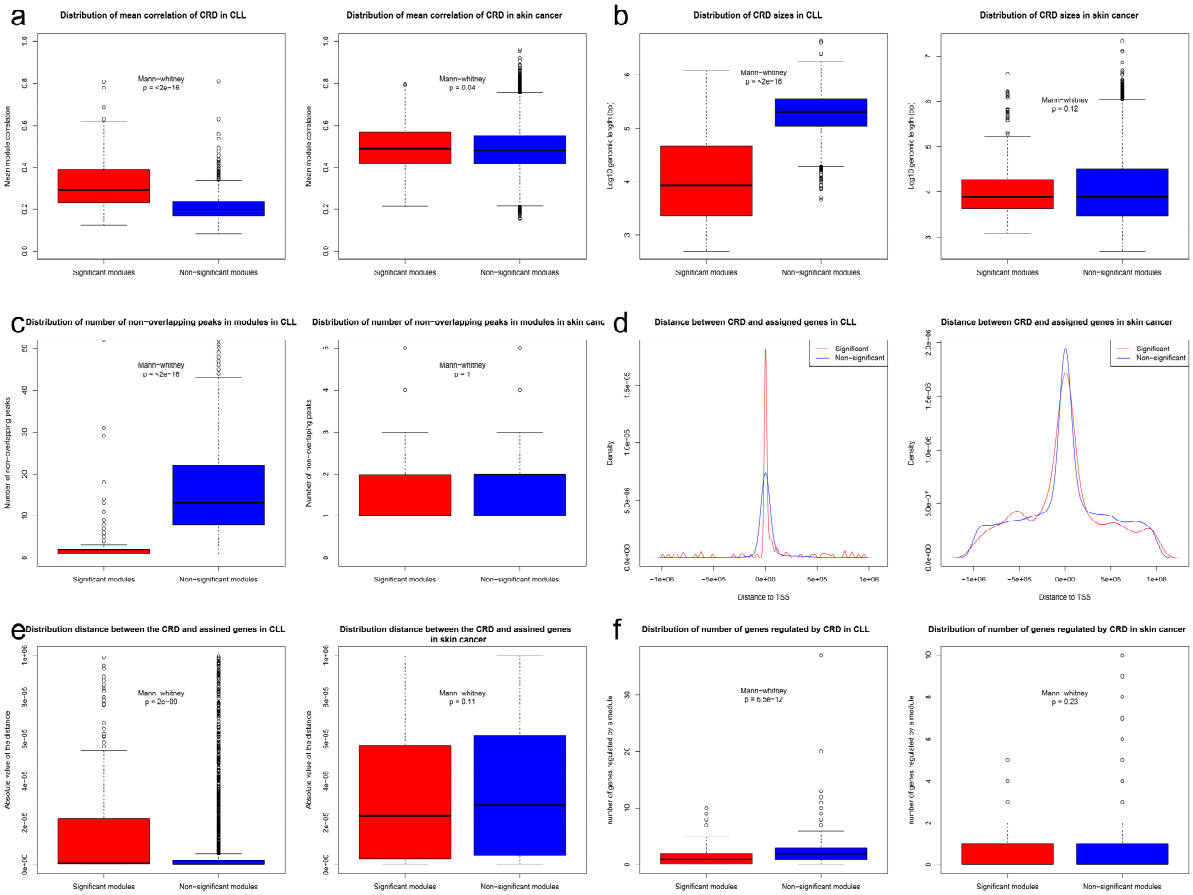

Figure 2 – Statistics of significant vs. non-significant CRDs in CLL and skin cancer. (a) The distributions of the average correlation among the CHiP-seq peaks of the CRDs. (b) The distributions of genomic lengths of significant and non-significant CRDs including their internal spacers. (c) The distributions of the number of non-overlapping CHiP-seq peaks in significant CRDs vs. the non-significant ones. The boxplots are truncated. (d) Density plot of the genomic distance between the CRDs and the genes they regulate for significant and non-significant CRDs. (e) The distributions of the absolute value of the genomic distance between the CRDs and the genes they regulate. (d) The number of genes the significant and non-significant modules regulate.

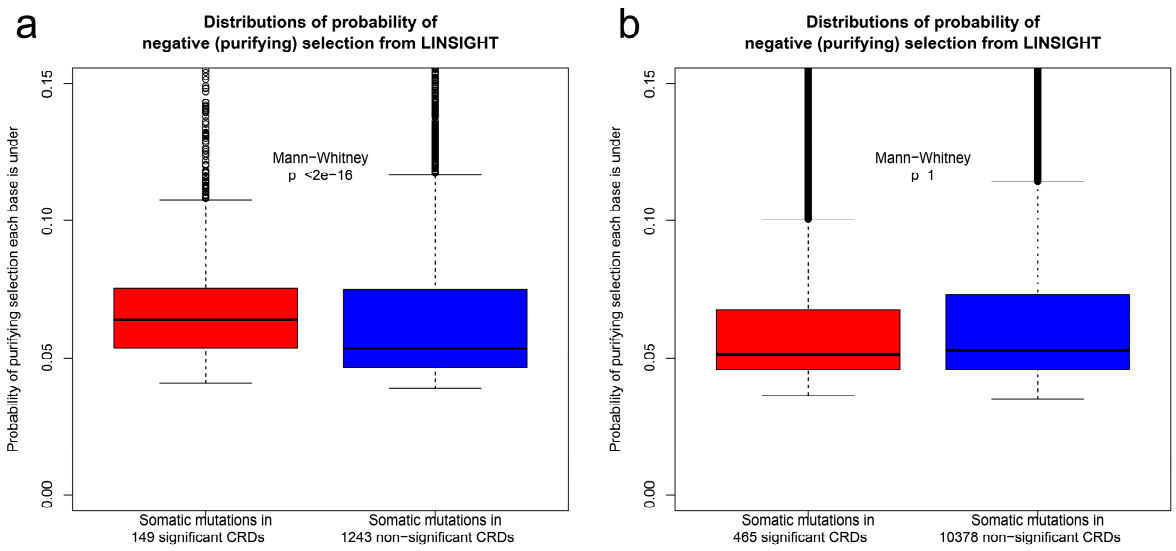

Figure 3 – The distributions of estimates of purifying selecting as calculated by LINSIGHT acting on somatic mutations in CRDs with significant excess of somatic mutations vs. ones that are not-significant in (a) CLL and (b) skin cancer. Boxplots are truncated at 0.15 probability of purifying selection. In CLL there is a significant increase in the purifying selecting acting on mutations in significant CRDs vs. non-significant ones, indicating that functional bases are being impacted in significant CRDs more so than in non-significant ones, suggesting selecting in tumorigenesis. This is not observed in skin cancer, likely due to very high number of somatic mutation seen skin cancer, diluting the signal.

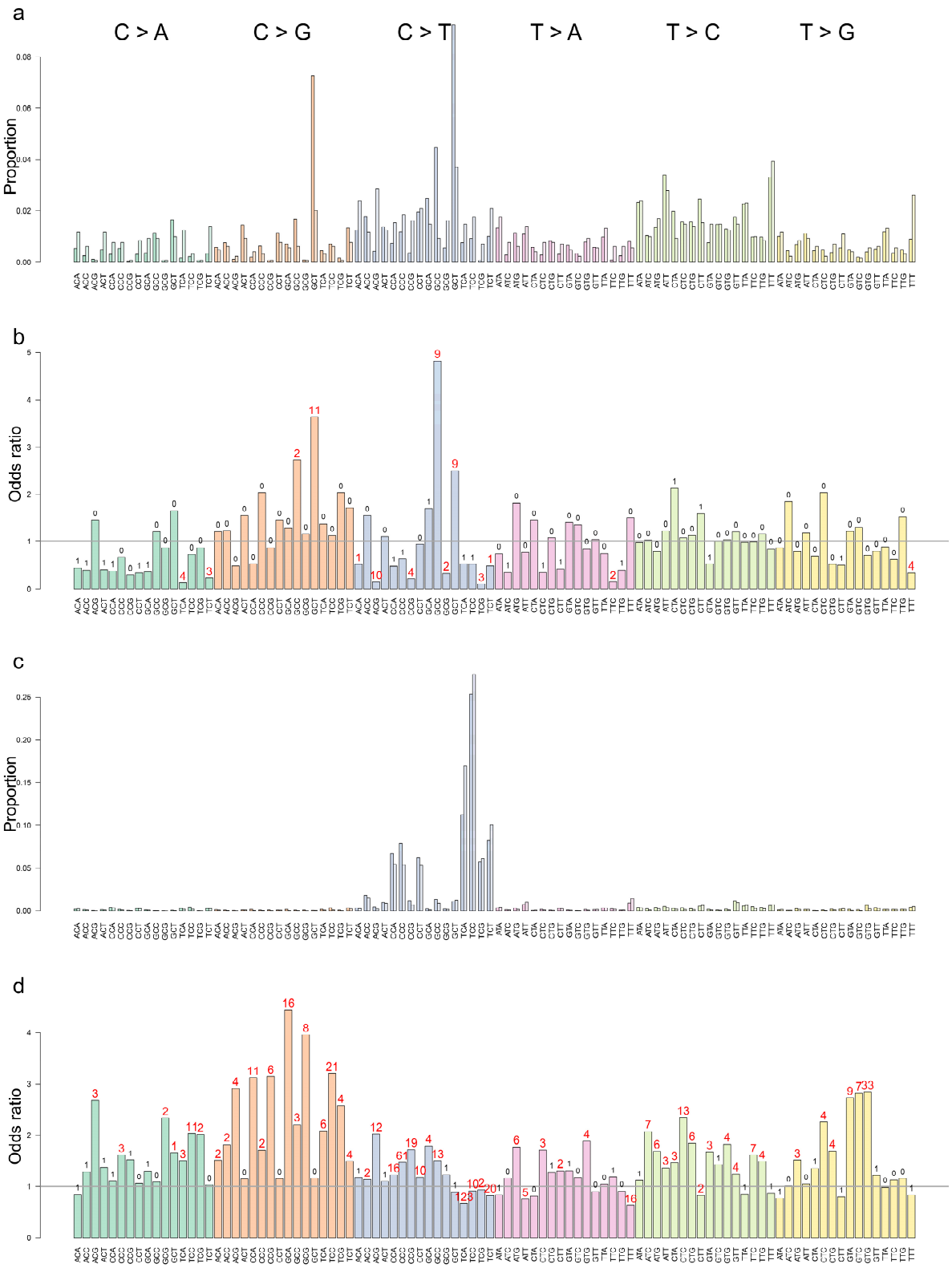

Figure 4 – The mutational signatures observed in the regulatory regions vs. their spacer. Bars are coloured by the nucleotide change observed and stratified by the local context. In (a) and (c) the proportion of each mutational signature in the regulatory regions are presented as darker shades whereas the spacers are shown as lighter shades of CLL and skin cancer, respectively. In (b) and (d) odds ratio, proportion in regulatory over proportion in spaces, of a given mutational signature is plotted for CLL and skin cancer, respectively. The numbers above the bars are the Benjamini-Hochberg corrected –log10(p-value) of the observed enrichment or depletion where the significant classes are shown in red.

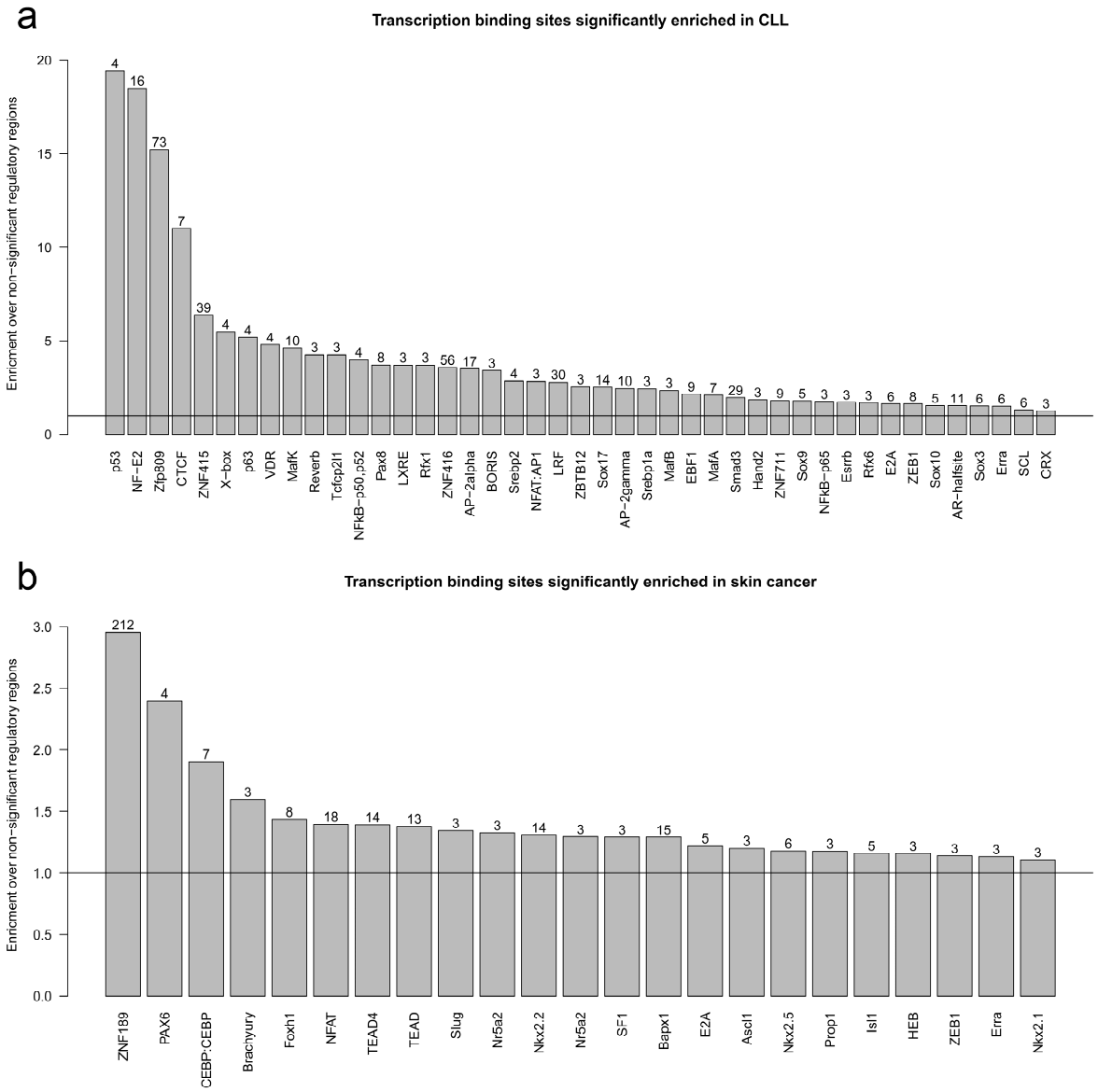

Figure 5– The transcription factors that are significantly enriched around somatic mutations in significant CRDs vs. the ones in non-significant CRDs for (a) CLL and (b) skin cancer. The horizontal black line indicate the base line, and the numbers above the bars are the Benjamini-Hochberg corrected –log10(p-value) of the observed enrichment.