


OPEN

# Novel Mutation Hotspots within Non-Coding Regulatory Regions of the Chronic Lymphocytic Leukemia Genome

Adrián Mosquera Orgueira <sup>1,2,3\*</sup>, Beatriz Rodríguez Antelo<sup>1,2,3</sup>, José Ángel Díaz Arias<sup>1,2</sup>, Nicolás Díaz Varela<sup>1,2</sup>, Natalia Alonso Vence<sup>1,2</sup>, Marta Sonia González Pérez<sup>1,2</sup> & José Luis Bello López<sup>1,2,3</sup>

Mutations in non-coding DNA regions are increasingly recognized as cancer drivers. These mutations can modify gene expression in *cis* or by inducing high-order chromatin structure modifications with long-range effects. Previous analysis reported the detection of recurrent and functional non-coding DNA mutations in the chronic lymphocytic leukemia (CLL) genome, such as those in the 3' untranslated region of *NOTCH1* and in the *PAX5* super-enhancer. In this report, we used whole genome sequencing data produced by the *International Cancer Genome Consortium* in order to analyze regions with previously reported regulatory activity. This approach enabled the identification of numerous recurrently mutated regions that were frequently positioned in the proximity of genes involved in immune and oncogenic pathways. By correlating these mutations with expression of their nearest genes, we detected significant transcriptional changes in genes such as *PHF2* and *S1PR2*. More research is needed to clarify the function of these mutations in CLL, particularly those found in intergenic regions.

A major part of mutations in the cancer genome occur in non-coding DNA regions, and their function is still beginning to be understood<sup>1</sup>. Non-coding DNA comprises approximately 98% of the human genome, but recent research has proven that most of these regions are either part of regulatory motifs or actively transcribed to RNA<sup>2,3</sup>. These mutations can induce functional genomic changes by altering the binding of transcription factors or by inducing high-order chromatin structural modifications<sup>2,4</sup>. For example, mutations in 5' and 3' untranslated regions (UTRs) may disturb RNA structural conformation, modify microRNA binding sites or disrupt polyadenylation signals<sup>2</sup>. In a similar fashion, mutations affecting non-protein coding genes such as microRNA and long intergenic RNA genes (lincRNAs) are known cancer driver events<sup>2,5</sup>. Different studies have evidenced that the expression of genes such as *BRCA1*, *CDH10*, *CCND1*, *MALAT1*, *PAX5*, *RB1*, *SDHD*, *TERT*, *TOX3*, and *TAL1* is influenced by non-coding DNA mutations in regulatory regions of the cancer genome<sup>1,6,7</sup>. The *Pancancer Analysis of Whole Genomes* (PCAWG) project has revealed the existence of common and tumor-specific recurrently mutated functional elements near known cancer drivers<sup>7</sup>. Some of these driver mutations can induce long-range changes in genome organization and trigger abnormal expression of distant oncogenes and tumor suppressors<sup>8</sup>. Furthermore, the sequence distribution of these driver mutations is not random. Hornshøj *et al.* (2018) identified a significant enrichment in conserved CCCT-binding factor (CTCF) binding sites among recurrently mutated non-coding DNA regions with cancer specificity<sup>6</sup>. Similarly, Line *et al.* (2019) identified 21 recurrently altered CTCF-rich insulator regions in the cancer genome, and elegantly demonstrated that some of these mutations drive tumor proliferation<sup>9</sup>.

Chronic Lymphocytic Leukemia (CLL) is among the most frequent lymphoproliferative disorders, and it is characterized by its remarkable clinical heterogeneity. Recent efforts by Puente *et al.*<sup>10</sup> enabled the discovery of 24 recurrently mutated non-coding genomic regions in the CLL genome, some of which are associated with functional changes such as mutations in the 3'UTR of *NOTCH1* and in the *PAX5* super-enhancer. Nevertheless, both

<sup>1</sup>Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain. <sup>2</sup>Complejo Hospitalario Universitario de Santiago de Compostela (CHUS), Division of Hematology, SERGAS, Santiago de Compostela, Spain.

<sup>3</sup>University of Santiago de Compostela, Santiago de Compostela, Spain. \*email: [adrian.mosquera@live.com](mailto:adrian.mosquera@live.com)

the sparsity of annotations in non-coding DNA regions and the difficult functional classification of non-coding DNA mutations hinder a better understanding of the non-coding cancer genome, which probably harbors multiple deregulated elements yet to discover. In this analysis, we analyzed whole genome sequencing (WGS) data using a best-practice mutation detection pipeline. Then, we identified signals of positive selection of mutations in regulatory regions. Finally, our last attempt was to analyze if any of these recurrent mutations in non-coding DNA regions were associated with abnormal expression of the nearest gene. Our results point toward the existence of dozens of mutation-enriched regulatory regions near cancer and immune-related genes, some of which influence local gene expression.

## Methods

**Data origin.** Whole genome sequencing files produced by the *International Cancer Genome Consortium*<sup>11</sup> were obtained from the *European Genome-Phenome Archive* under accession code EGAD00001001466. Gene expression from microarray data of the same set of patients was obtained from EGAD00010000875.

**Data analysis.** 130 tumor-normal matched CLL whole genomes were processed using the bcbio-nextgen pipeline, which provides best practices for analyzing high throughput sequencing data<sup>12</sup>. Low complexity regions, areas with abnormally high coverage, sequences with single nucleotide stretches >50 bp and loci with alternative or unplaced contigs in the reference genome were not analyzed. Some polymorphic regions are prone to be classified as highly mutated due to artifacts or biases in the sequencing process, and suspicious elements were manually removed from downstream analysis. Single nucleotide and indel mutation detection was performed with *vardict*<sup>13</sup>, *varscan*<sup>14</sup>, *mutect2*<sup>15</sup> and *freebayes*<sup>16</sup> using default bcbio-nextgen parameters. Only variants with a minimum sequencing depth (DP) of 10 and a genotype quality (GQ) above 20 Phred in both tumor and normal samples were analyzed. A mutation was reported when detected by at least two different mutation callers. Mutations were annotated to the 1000G<sup>17</sup>, gnomAD<sup>18</sup> and ExAc<sup>19</sup> databases in order to filter likely germline variants. All mutations with a minimum allele frequency >0.001 in any population were discarded from the analysis.

**Region annotation.** Annotations corresponding to promoter regions, 5'UTR, 3'UTR and lincRNAs were retrieved from Genecode version 18<sup>20</sup>. DNase hypersensitivity (DHS) regions and Transcription Factor Binding Sites (TFBS) tracks from the ENCODE<sup>21</sup> project were obtained from Lochofsky *et al.*<sup>22</sup>. Similarly, we used enhancer regions from the GeneHancer database<sup>23</sup>, and analyzed those that were supported by two or more sources of evidence (“elite” enhancers). Regulatory regions within telomeric and centromeric positions were discarded.

Two different methods were used to identify areas with evidence of positive selection of mutations: *LARVA*<sup>22</sup> and *OncodriveFML*<sup>24</sup>. *LARVA* models the mutation counts of each target region as a  $\beta$ -binomial distribution in order to handle overdispersion. Furthermore, *LARVA* also includes replication timing information in order to estimate local mutation rate, and provides a  $\beta$ -binomial distribution adjusted for replication timing which is used to compute p-values. On the other hand, *OncodriveFML* is designed to analyze the pattern of somatic mutations across tumors in both coding and non-coding genomic regions. *OncodriveFML* uses functional predictions in order to identify signals of positive selection. *OncodriveFML* was run with CADD v1.3 scores and default parameters. TFBS tracks were not analyzed with *OncodriveFML* due to high computational demands. Regions were labeled as significantly mutated if the q-value was <0.05 with any of the two methods.

## Gene expression analysis and association with recurrent non-coding DNA mutations.

Background correction, normalization and log<sub>2</sub>-transformation of microarray gene expression data was performed with the RMA algorithm<sup>25</sup>. In the case of genes targeted by multiple probes, the median expression was calculated. The Wilcoxon-Rank sum test was used to detect changes in gene expression between mutated and wild-type cases. Non-coding regulatory genomic regions cannot be directly ascribed to any gene, and they can affect the transcription of virtually any part of the genome. However, this study is underpowered to detect long-range interactions due to small sample size and the need of extreme p-values passing multiple-testing correction. Therefore, we centered our efforts on changes in expression of the nearest gene. We annotated the closest gene to each recurrently mutated non-coding genomic region as the nearest transcription start site to the middle position of the corresponding region. In the case of multiple overlapping regulatory regions, we selected the most significant one for downstream analysis. P-values were adjusted for multiple testing using the FDR method, with a significance threshold of 0.05.

## Results

**Mutation distribution.** 397,433 non-coding DNA mutations were detected in the genome of this CLL cohort. Most of these were either intergenic (45.46%) or intronic (42.12%). The remaining mutations were located in 5' flanks (5.83%), 3' flanks (5.30%), RNA genes (0.64%), 3'UTRs (0.52%) and 5'UTRs (0.13%). Most of the mutations were single nucleotide variants (92.96%), whereas 4.57% and 2.47% were short deletions and insertions, respectively.

**Regions significantly enriched in mutations.** *LARVA* detected significant mutation enrichments (q-value < 0.05) in 120 TFBS, 16 DHS regions, 10 enhancers, 4 promoters, 2 5'UTRs and 1 lincRNA (Table 1, Supplementary Tables 1–6). No relevant inflation in p-value distribution was observed. (Supplementary Fig. 1). These regions were located in 44 different genomic loci (Fig. 1). The most recurrently mutated promoters were those of *TCL1A* (q-value  $3.32 \times 10^{-4}$ ), *LCN6* (q-value  $4.17 \times 10^{-3}$ ), *ZFP36L1* (q-value  $3.25 \times 10^{-2}$ ) and *WDR97* (q-value 0.04); and the most significantly mutated enhancers were *GH01J229147* (intergenic region chr1:229283343–229284982, q-value  $5.79 \times 10^{-6}$ ) and *GH07J000467* (*PDGFA* gene, q-value  $8.53 \times 10^{-4}$ ). The DHS regions chr4:184474905–184475055 (*ING2/RWDD4* locus, q-value  $1.42 \times 10^{-5}$ ), chr21:46673965–46674115

Chromosome	Start	Stop	Mutation count	p-value (bbd)	FDR	Gene	SHM target	Type of Regulator
chr1	155666495	155666977	16	2.20E-16	3.14E-10	<i>DAP3</i>	No	TFBS
chr1	229283343	229284982	28	1.58E-10	5.79E-06	Intergenic	No	ENHANCER
chr3	186782686	186783907	26	1.34E-09	3.52e-04	<i>BCL6</i>	Yes	TFBS
chr4	184474905	184475055	13	3.92E-11	1.42E-05	<i>ING2/RWDD4</i>	No	DHS
chr7	507064	509696	17	4.65E-08	8.53e-4	<i>PDGFA</i>	No	ENHANCER
chr7	507220	508145	17	3.85E-09	8.15E-4	<i>PDGFA</i>	No	TFBS
chr9	115161245	115161395	11	2.98E-09	3.98e-4	<i>HSDL2</i>	No	DHS
chr11	65265233	65273940	10	2.04E-08	4.50e-4	<i>MALAT1</i>	Yes	lincRNA
chr14	96179060	96180273	25	2.34E-08	3.32E-4	<i>TCLIA</i>	Yes	PROMOTER
chr14	96179721	96180690	22	1.36E-09	3.52E-4	<i>TCLIA</i>	Yes	TFBS
chr14	96179799	96180653	21	6.70E-10	2.25E-4	<i>TCLIA</i>	Yes	TFBS
chr14	96179816	96180607	21	2.67E-10	1.38E-4	<i>TCLIA</i>	Yes	TFBS
chr14	96179960	96180110	12	3.03E-09	3.98E-4	<i>TCLIA</i>	Yes	DHS
chr21	46673965	46674115	12	7.33E-10	1.38E-4	<i>C21ORF89/LINC00334</i>	No	DHS

**Table 1.** Summary of the regions most significantly enriched in mutations according to *LARVA*.

(*C21ORF89/LINC00334* locus, q-value  $1.38 \times 10^{-4}$ ), chr14:96179960–96180110 (*TCLIA* locus, q-value  $3.98 \times 10^{-4}$ ) and chr9:115161245–115161395 (*HSDL2* locus, q-value  $3.98 \times 10^{-4}$ ) were the most recurrently mutated among their class (Supplementary Table 2). Furthermore, up to 120 significantly mutated TFBS regions were detected, affecting 19 different genes and 3 intergenic regions. The most recurrently mutated regions were located in chr1:155666495–155666977 (*DAP3* gene, q-value  $3.14 \times 10^{-10}$ ), chr14:96179816–96180607 (*TCLIA* gene,  $1.38 \times 10^{-4}$ ), chr3:186782686–186783907 (*BCL6* gene,  $3.52 \times 10^{-4}$ ), chr7:507220–508145 (*PDGFA* gene,  $8.15 \times 10^{-4}$ ) and chr18:12086057–12086469 (*ANKRD62* gene,  $8.30 \times 10^{-4}$ ) (Supplementary Table 4).

Other significant enhancer regions were located in the proximity of genes involved in apoptosis (*BCL2* and *BIRC3*), cell cycle control (*WBP2NL*), cytoskeleton and extracellular matrix formation (*ARPC3* and *ITIH5*), gene expression regulation and chromatin remodelling (*BCL7A*, *PAX5* and *PHF2*), genome integrity (*XRCC5* and *ZNF506*), gene expression regulation (*MALAT1* and *RBFOX3*), intracellular signalling (*DACT2*, *HIPK2*, *IMPA2*, *KCTD10*, *ROR2* and *SIPR2*), immune pathways (*BACH2*, *LTB* and *MADCAM1*) and metabolism (*AKR1B15*, *AMPD3*, *GSTM1/GSTM2*, *LRP5* and *ST6GAL1*) (Supplementary Tables 1–6). Recurrent mutations were also found near less well-characterized genes such as *TMEM54* and *CTBP2P5*, as well as within intergenic regions such chr14:26068671–26069217 and chr1:229283491–229285693.

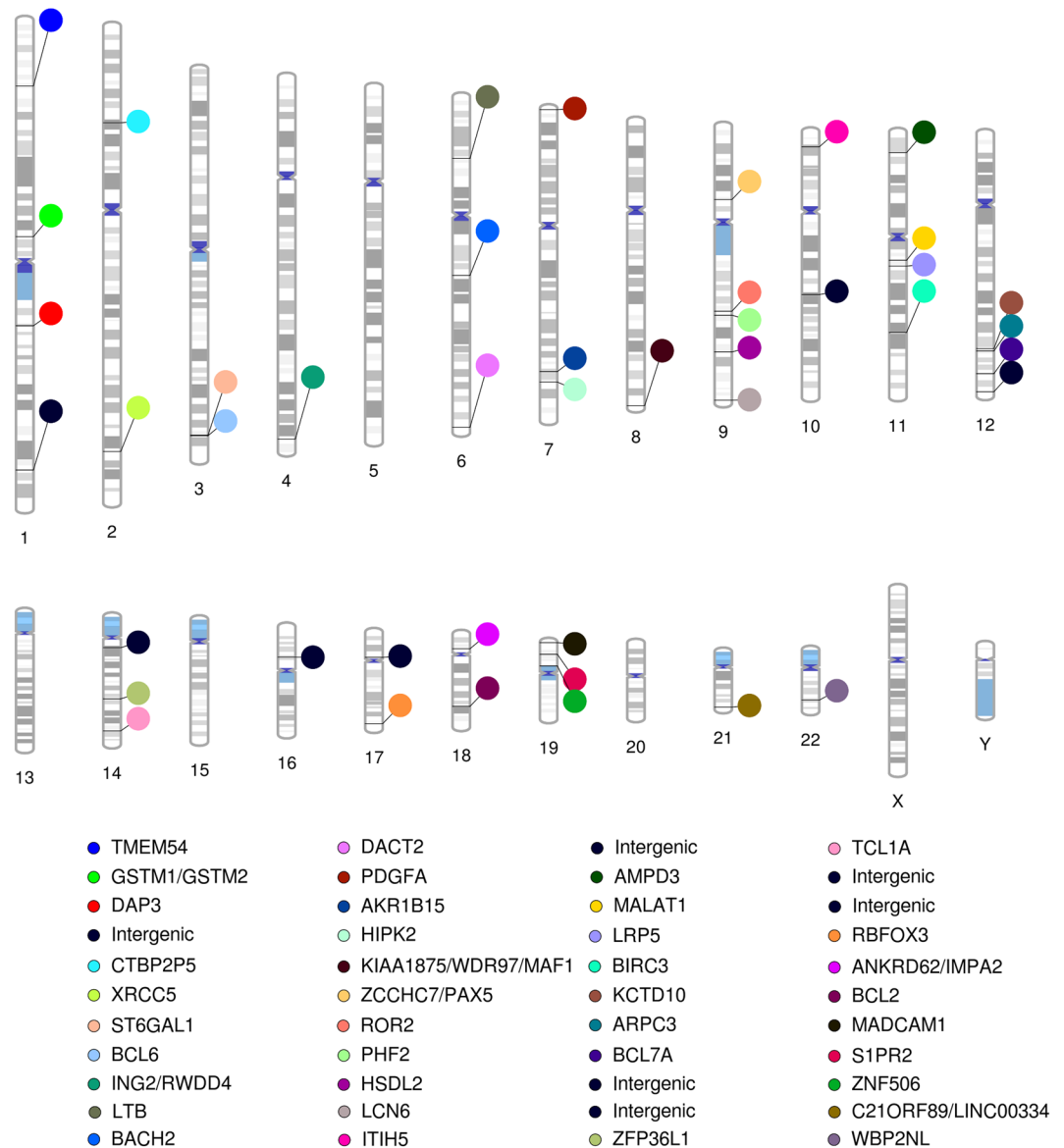
Finally, *OncodriveFML* identified 4 regions significantly enriched in likely functional mutations (Supplementary Tables 7 and 8). No relevant inflation in p-value distribution was observed (Supplementary Fig. 2). These regions were the enhancer *GH14J089855* (q-value  $2.54 \times 10^{-3}$ ) encoded within an intronic region of *EFCAB11*, two DHS regions in the proximity of *EGR* and *WBNPL2* (q-values 0.01 and 0.03, respectively), and one intergenic DHS region located in chr8:127155560–127155710 (q-value  $1.22 \times 10^{-3}$ ).

**Mutations associated with changes in gene expression.** We studied the association of regions enriched in mutations with changes in the expression of their respective nearest genes. Although this type of analysis is limited by low sample size, we detected significant associations in some cases. We tested if patients with at least one mutation in these regulatory regions were accompanied by changes in expression of the nearest gene. Significant associations were observed in 3 genes, namely *PHF2* (q-value 0.02, 95% CI [−0.295, −0.048]), *RPL39L* (q-value 0.04, 95% CI [0.018, 0.217]) and *SIPR2* (q-value 0.03, 95% CI [0.033, 0.38]) (Supplementary Table 9).

## Discussion

Mutations in the non-coding part of the genome constitute the “dark-matter” of cancer genomics<sup>2</sup>. Growing evidence indicates that many of these mutations occur in conserved motifs and loci under epigenetic control, and some of these play fundamental roles in cancer biology and disease prognosis<sup>1–3,6–9</sup>. Using WGS data produced by the ICGC, we identified dozens of recurrently mutated regulatory regions in the CLL genome. Among these, 10 were previously reported by the original analysis performed by Puente *et al.*<sup>10</sup>, namely those near *BACH2*, *BLC2*, *BCL6*, *BCL7A*, *BIRC3*, *SIPR2*, *PCDH15*, *ZCCHC7/PAX5* and *ZFP36L1*. Numerous novel regions were also enriched in non-coding DNA mutations, including transcription factor binding sites, DNase hypersensitivity regions, 5'UTR regions, promoters, enhancers and non-coding RNAs. These events were frequently found in the vicinity of genes previously vinculated with oncogenic pathways. Indeed, the most significantly mutated regions were a SETB1 binding site within the first intron of *DAP3*, a GTP-binding protein that participates in the apoptosis pathway<sup>26</sup>, and a DNase hypersensitivity region downstream to *ING2*, a well-characterized tumor suppressor<sup>27</sup>. Other highly mutated regulatory regions affected cancer-related genes such as *DACT2*<sup>28</sup>, *ERG*<sup>29</sup>, *HIPK2*<sup>30</sup>, *ITIH5*<sup>31</sup>, *LRP5*<sup>32</sup>, *MAF1*<sup>33</sup>, *MALAT1*<sup>34</sup>, *PHF2*<sup>35</sup>, *PDGFA*<sup>36</sup>, *RBFOX3*<sup>37</sup>, *ROR2*<sup>38</sup>, *ST6GAL1*<sup>39</sup> and *XRCC5*<sup>40</sup>; and others were detected near genes involved in immunity, such as *LTB*<sup>41</sup> and *MADCAM1*<sup>42</sup>. Overall, only three of the novel genes (*LTB*, *MALAT1* and *ST6GAL1*) were previously defined as targets of somatic hypermutation in B cell lymphomas<sup>43</sup>. Finally, it is worthwhile to mention that recurrent and even highly significant enrichments were detected around barely characterized genes (e.g. *C21ORF89/LINC00334*) and intergenic regions.

## LARVA SIGNIFICANT REGULATORY REGIONS



**Figure 1.** Chromosomal ideogram representing the different gene affected by recurrent non-coding mutations according to *LARVA*.

The reported mutations can either be bystander or have functional implications related to their potential to modify gene expression or to induce high-order chromatin structural changes. Although limited by low sample size, we devised significant changes in the expression of *PHF2*, *S1PR2* and *RPL39L*. These three genes are involved in the regulation of important oncogenic processes. *PHF2* encodes a histone demethylase with tumor suppressor activity<sup>35</sup>. *S1PR2* participates in the TGF- $\beta$  pathway and acts as a tumor suppressor of B cell lymphomas<sup>44</sup>. Finally, *RPL39L*<sup>45</sup> is involved in cancer stem cell self-renewal and hypoxia response. These results are concordant with other reports of non-coding regulatory mutations driving gene expression changes in B-cell lymphomas<sup>46–48</sup>.

The combination of an optimized mutation detection pipeline with statistical tests specifically designed to handle non-coding DNA mutations has enabled the detection of novel putative regulatory driver regions in the CLL genome. These regions were mostly located in the vicinity of genes implicated in oncogenic and immune pathways, although several recurrently mutated intergenic regions were detected too. Furthermore, we could confirm the association of some of these events with altered expression of their respective genes. We expect that our results, along with those published by other groups, will promote an improved characterization of the non-coding mutational drivers of CLL.

### Data availability

There is not data to deposit.

Received: 7 May 2019; Accepted: 27 January 2020;

Published online: 12 February 2020

## References

- Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165, <https://doi.org/10.1038/ng.3101> (2014).
- Diederichs, S. *et al.* The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol. Med.* **8**, 442–457, <https://doi.org/10.1525/emmm.201506055> (2016).
- Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* **11**, 559–571, <https://doi.org/10.1038/nrg2814> (2010).
- Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Sci.* **346**, 1373–1377, <https://doi.org/10.1126/science.1259037> (2014).
- Palamarchuk, A. *et al.* 13q14 deletions in CLL involve cooperating tumor suppressors. *Blood* **115**, 3916–3922, <https://doi.org/10.1182/blood-2009-10-249367> (2010).
- Hornshøj, H. *et al.* Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. *NPJ Genom. Med.* **3**, 1, <https://doi.org/10.1038/s41525-017-0040-5> (2018).
- Rheinbay, E. *et al.* Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv* 237313, <https://doi.org/10.1101/237313>.
- Wadi, L. *et al.* Candidate cancer driver mutations in superenhancers and long-range chromatin interaction networks. *bioRxiv* 236802, <https://doi.org/10.1101/236802>
- Liu, E. M. *et al.* Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. *Cell Syst.* **8**, 446–455, <https://doi.org/10.1016/j.cels.2019.04.001> (2019).
- Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nat.* **526**, 519–524, <https://doi.org/10.1038/nature14666> (2015).
- International Cancer, G. C. *et al.* International network of cancer genome projects. *Nat.* **464**, 993–998, <https://doi.org/10.1038/nature08987> (2010).
- Valls-Guimera, R. Bcbio-nextgen: Automated, distributed, next-gen sequencing pipeline. *EMBnet J.* **17**, 30, <https://doi.org/10.14806/ej.17.B.286> (2012).
- Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108, <https://doi.org/10.1093/nar/gkw227> (2016).
- Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576, <https://doi.org/10.1101/gr.129684.111> (2012).
- do Valle, I. F. *et al.* Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinforma.* **17**, 341, <https://doi.org/10.1186/s12859-016-1190-7> (2016).
- Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*, 1207.3907 [q-bio.GN] (2012)
- 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526** 68–74, <https://doi.org/10.1038/nature15393> (2015).
- Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210, <https://doi.org/10.1101/531210>.
- Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nat.* **536**, 285–291, <https://doi.org/10.1038/nature19057> (2016).
- Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774, <https://doi.org/10.1101/gr.135350.111> (2012).
- Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, 794–801, <https://doi.org/10.1093/nar/gkx1081> (2018).
- Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res.* **43**, 8123–8134, <https://doi.org/10.1093/nar/gkv803> (2015).
- Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, <https://doi.org/10.1093/database/bax028> (2017).
- Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128, <https://doi.org/10.1186/s13059-016-0994-0> (2016).
- Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2013).
- Wazir, U. *et al.* The role of death-associated protein 3 in apoptosis, anoikis and human cancer. *Cancer Cell Int.* **15**, 39, <https://doi.org/10.1186/s12935-015-0187-z> (2015).
- Guérillon, C., Larrieu, D. & Pedoux, R. ING1 and ING2: multifaceted tumor suppressor genes. *Cell Mol. Life Sci.* **70**, 3753–3772, <https://doi.org/10.1007/s00018-013-1270-z> (2013).
- Li, J. *et al.* Methylation of DACT2 promotes breast cancer development by activating Wnt signaling. *Sci. Rep.* **7**, 3325, <https://doi.org/10.1038/s41598-017-03647-3> (2017).
- Adamo, P. & Ladomery, M. R. The oncogene ERG: a key factor in prostate cancer. *Oncogene* **35**, 403–414, <https://doi.org/10.1038/nc.2015.109> (2016).
- D'Orazi, G., Rinaldo, C. & Soddu, S. Updates on HIPK2: a resourceful oncosuppressor for clearing cancer. *J. Exp. Clin. Cancer Res.* **31**, 63, <https://doi.org/10.1186/1756-9966-31-63> (2012).
- Rose, M. *et al.* ITIH5 induces a shift in TGF- $\beta$  superfamily signaling involving Endoglin and reduces risk for breast cancer metastasis and tumor death. *Mol. Carcinog.* **57**, 167–181, <https://doi.org/10.1002/mc.22742> (2018).
- Ren, D. N. *et al.* LRP5/6 directly bind to Frizzled and prevent Frizzled-regulated tumour metastasis. *Nat. Commun.* **6**, 6906, <https://doi.org/10.1038/ncomms7906> (2015).
- Li, Y. *et al.* MAF1 suppresses AKT-mTOR signaling and liver cancer through activation of PTEN transcription. *Hepatology* **63**, 1928–1942, <https://doi.org/10.1002/hep.28507> (2016).
- Liu, J., Peng, W. X., Mo, Y. Y. & Luo, D. MALAT1-mediated tumorigenesis. *Front. Biosci.* **22**, 66–80 (2017).
- Lee, K. H. *et al.* PHF2 histone demethylase acts as a tumor suppressor in association with p53 in cancer. *Oncogene* **34**, 2897–2909, <https://doi.org/10.1038/nc.2014.219> (2015).
- Palomero, J. *et al.* SOX11 promotes tumor angiogenesis through transcriptional regulation of PDGFA in mantle cell lymphoma. *Blood* **124**, 2235–2247, <https://doi.org/10.1182/blood-2014-04-569566> (2014).
- Liu, T. *et al.* RBFox3 Promotes Tumor Growth and Progression via hTERT Signaling and Predicts a Poor Prognosis in Hepatocellular Carcinoma. *Theranostics* **7**, 3138–3154, <https://doi.org/10.7150/thno.19506> (2017).
- Debebe, Z. & Rathmell, W. K. Ror2 as a therapeutic target in cancer. *Pharmacol. Ther.* **150**, 143–148, <https://doi.org/10.1016/j.pharmthera.2015.01.010> (2015).

39. Antony, P. *et al.* Epigenetic inactivation of ST6GAL1 in human bladder cancer. *BMC Cancer* **14**, 901, <https://doi.org/10.1186/1471-2407-14-901> (2014).
40. Zhang, Z. *et al.* XRCC5 cooperates with p300 to promote cyclooxygenase-2 expression and tumor growth in colon cancers. *PLoS One* **12**, e0186900, <https://doi.org/10.1371/journal.pone.0186900> (2017).
41. Nagy, B. *et al.* Lymphotoxin beta expression is high in chronic lymphocytic leukemia but low in small lymphocytic lymphoma: a quantitative real-time reverse transcriptase polymerase chain reaction analysis. *Haematologica* **88**, 654–658 (2003).
42. Sakai, Y. & Kobayashi, M. Lymphocyte ‘homing’ and chronic inflammation. *Pathol. Int.* **65**, 344–354, <https://doi.org/10.1111/pin.12294> (2015).
43. Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308–1319 (2012).
44. Stelling, A. *et al.* The tumor suppressive TGF- $\beta$ /SMAD1/S1PR2 signaling axis is recurrently inactivated in diffuse large B-cell lymphoma. *Blood* **131**, 2235–2246, <https://doi.org/10.1182/blood-2017-10-810630> (2018).
45. Dave, B. *et al.* Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. *Proc Natl Acad Sci USA* **111**, 8838–8843, <https://doi.org/10.1073/pnas.1320769111>.
46. Batmanov, K., Wang, W., Björås, M., Delabie, J. & Wang, J. Integrative whole-genome sequence analysis reveals roles of regulatory mutations in BCL6 and BCL2 in follicular lymphoma. *Sci. Rep.* **7**, 7040, <https://doi.org/10.1038/s41598-017-07226-4> (2017).
47. Arthur, S. E. *et al.* Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001, <https://doi.org/10.1038/s41467-018-06354-3> (2018).
48. Mathelier, A. *et al.* Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.* **23**(16), 84, <https://doi.org/10.1186/s13059-015-0648-7> (2015).

## Acknowledgements

We would like to thank the *International Cancer Genome Consortium* for facilitating the data, and to the *Supercomputing Center of Galicia* (CESGA) for providing informatics support for the analysis. The content of this paper is part of the doctoral thesis of Adrián Mosquera Orgueira to obtain a PhD at the Department of Medicine, University of Santiago de Compostela.

## Author contributions

A.M.O. designed the research and performed the analysis. A.M.O., B.A.R. and J.L.B.L. analyzed the results and wrote the paper. J.A.D.A., N.D.V., N.A.V. and M.S.G.P. critically evaluated the paper.

## Competing interests

The article processing fee of this paper has been partially funded by Roche Pharmaceuticals. Notwithstanding, this company did not have any influence on the study design, data analysis, results interpretation or article writing.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-59243-5>.

**Correspondence** and requests for materials should be addressed to A.M.O.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020