Computational Annotation of Genetic Biomarkers using Topologically Associating Domains

Carlos Cano[1Φ], Michela Verbeni[1Φ], Carmen Navarro[1], Maria S. Benítez-Cantos[1], Antonio González-Aguilar[2], Gema Durán-Ogalla[3], Manuel Benavides[3], Susana Pedrinaci[4], Mercedes López de Hierro-Ruiz[4], Pilar Martínez-Tirado[5], Francisco Ruiz-Cabello[4], José Luis Martín-Ruiz[5], Armando Blanco[1], Paul Lizardi[1,6].

(1) Department of Computer Science and Artificial Intelligence, University of Granada, Spain.
(2) Instituto de Parasitología y Biomedicina Lopez-Neyra CSIC, Granada, Spain.
(3) Hospital Carlos Haya, Oncología Médica, Málaga, Spain.
(4) Hospital Virgen de las Nieves, Granada, Spain.
(5) Hospital Clínico San Cecilio, Granada, Spain.
(6) GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, Granada, Spain.
Φ These authors contributed equally to this work.
Corresponding authors: ccano@decsai.ugr.es, michelav@decsai.ugr.es, paul.lizardi@me.com.

Keywords: Topologically Associating Domains, Computational Annotation, Methylation biomarkers, Lynch Syndrome

Abstract. The study of the 3D organization of nuclear DNA is attracting increasing interest since it has been shown to have a direct impact in the regulatory machinery of the cell. In particular, topologically associating domains (TADs) are structural units of chromatin regions proven to be highly self-interacting. A TAD can span from hundreds of kilobases to few megabases, thus potentially including a set of different genes together with promoters and regulatory regions.
Improvements in sequencing and computational technologies are continuously delivering biomarker signatures for many different diseases. These signatures often involve hundreds or thousands of different genomic loci, thus transforming in a challenge interpreting and identifying the underlying regulatory mechanisms for the target condition.
We propose a tool for computational and statistical analysis of biomarker signatures and topologically associating domains, with the aim of determining DNA domains significantly enriched with loci of interest. We believe that this approach eases the interpretation and further study of large sets of biomarker loci. In this paper, we show the potential of this tool in a case study of methylation biomarkers for Lynch Syndrome. However, the proposed tool is of general purpose and can be run on any set of loci of interest to identify enriched DNA domains.

1 Introduction

Spatial organization of DNA is a key player for genome functionality and transcription [1]. Topologically associating domains (TADs) are primarily cell-type-independent genomic regions that define interactome boundaries and can aid in the designation of limits within which an association most likely impacts genome-encoded regulatory function, whether this function is mediated by protein-coding elements or by non-coding RNA. TADs are experimentally defined by observations of increased chromatin contact frequency, consistency across cell types and, interestingly, of enrichment of CTCF insulator element flanks [1, 2]. Therefore, TADs can be used as virtual genome-grammar

boundaries that demarcate locations where not only genes, but most notably non-coding causal variants will most likely impact regulatory function [3]. Given their increasing popularity, many computational tools have emerged to predict TADs from Hi-C data (for a review, see [4, 5]). Also, Way et al (2017) have described and validated a computational method that uses the genic content of TADs to prioritize candidate genes. Their method, called 'TAD_Pathways', performs a Gene Ontology analysis over genes that reside within TAD boundaries corresponding to GWAS signals for a given trait or disease [6].

DNA methylation is an epigenetic mark related to many biological aspects such as gene expression, diseases or immunity. Cytosines methylation occurs when a methyl group attaches to a cytosine (C) in the DNA sequence, affecting its chemical properties. Studying DNA methylation is drawing increasing relevance, since it is dynamic and reversible. Mature technologies exist to measure DNA methylation levels in particular regions of DNA, such as methylation arrays and Bisulfite-Sequencing (BS-Seq) technologies. Currently, the identification of individual differentially methylated regions (DMR) is attracting most of the effort to discover methylation biomarkers. However, methylation sequencing experiments often generate hundreds of potential biomarkers which need further annotation and interpretation. In this paper we present a tool that allows to identify TADs significantly enriched with biomarkers of interest.

Particularly, we show the potential of this tool in a case study on methylation biomarkers for Lynch Syndrome (LS). LS represents between 4% to 5% of all colorectal cancer (CRC) cases and it is caused by different known mutations in DNA repair genes. Individuals that are carriers of LS mutations have a high risk of developing CRC, but also endometrial, ovarian, small intestine, and urinary tract cancers. LS has an autosomal-dominant transmission and causes an estimated lifetime risk for CRC as high as 80% to 85%, compared with 5% in the general population. Although screening with colonoscopy has been demonstrated to decrease cancer incidence in individuals that carry LS mutations, many patients continue developing CRC at a young age. Therefore, there is an urgent need for a better understanding of the pathogenesis of colorectal premalignant lesions, instrumental for the development of novel preventive strategies for Lynch syndrome. Chang et al (2018) have recently demonstrated that polyps arising in Lynch Syndrome before cancer onset are enriched with CD4+ T cells [7]. This is consistent with recently reported transcriptomic profiles detected in normal mucosa samples of patients with LS who harbored a CRC, which showed strong immune responses associated to microenvironment invasion by CD4+ T cells, expression of immune checkpoints, and HLA [8]. In view of this strong evidence for early involvement of CD4+ T cell responses during the emergence of premalignant lesions in the colon of individuals at risk, we decided to ask whether or not the appearance of DNA methylation alterations in CD4+ cells in peripheral blood could serve as a very early biomarker of CRC risk.

Although this paper illustrates a case study on methylation biomarkers for Lynch Syndrome, the proposed tool is of general purpose and can be run on any set of loci of interest to identify enriched TADs.

This paper is structured as follows. Section Methods describes the methylation data for the case study on Lynch Syndrome, the computational analysis for the identification of potential methylation biomarkers and the proposed pipeline for TAD enrichment. Section Results provides the results for the case study on Lynch Syndrome. Finally, we present the conclusions.

2 Methods

2.1 Methylation biomarkers for Lynch syndrome

Peripheral blood samples were obtained from a sample of 16 healthy individuals, who either carry or do not carry LS mutations, as determined by genetic analysis. Nine

individuals present genetic mutations linked to LS, but do not exhibit any clinical signs (cases), while the other seven are healthy individuals of the same family, without Lynch type mutations. The study is currently ongoing and it is planned to recruit at least 80 – 100 subjects with Lynch type mutations. CD4+ T cells were isolated from each sample and processed. Genome-wide bisulfite sequencing was carried out using the Agilent SureSelect Human Methyl-Seq platform; the data at our disposal describe methylation status of 1958283 CG sites. The primary objective of the study is finding differentially methylated loci, by comparing methylation level of each site between cases and controls, to be used as biomarkers for very early stage disease diagnosis. Methylation status of a CG site is usually quantified by the ratio $\beta = M/(M + U)$, where M and U denote methylated and unmethylated signal intensities, respectively. Treatment with bisulfite deaminates unmodified cytosines to uracils (which are then amplified to thymines), leaving 5-methylcytosines unaltered. Thus in a bisulfite sequencing experiment β values are expressed as the proportion of cytosines in the total number of reads of a given CG site, $\beta = N_i^{\mathrm{C}}/C$, where N_i^{C} indicates the number of cytosines and C the coverage. As we cannot assume a normal distribution for the data, we employed the non-parametric two-sample Wilcoxon rank sum test to calculate p-values for differential methylation, as measured by β values, between cases and controls. Since we are performing multiple hypotheses testing, we should control for false discovery rate, using for example Benjamini and Hochberg or Benjamini and Yekutieli methods. However for high-throughput methylation data involving thousands of comparisons, usually with small sample size, it is still a matter of debate which is the best approach to effectively reduce false positives. Our data derive from a very preliminary study and we were mainly interested in developing a procedure to underpin the role of TADs in genomics investigations, thus we directly used non-adjusted p-values from Wilcoxon test and selected a set of differentially methylated loci with $p < 0.005$, to analyze their relation with TAD structure.

2.2 Pipeline for TAD enrichment

This section describes the pipeline for the statistical enrichment of TADs with a set of biomarkers of interest. For this process, two inputs are required: (1) the biomarker signatures of interest, i.e. a set of genetic loci; (2) the definition of TAD boundaries.

Our tool implements the following steps:

1. Load the TAD boundaries and perform a whole-genome binning guided by TAD coordinates. We define a bin as a region corresponding to either a TAD or the gap region between two TADs.
2. Perform basic statistics on the number of TADs, the distribution of bin sizes, etc.
3. Count the number of signature loci (hits) located in every bin.
4. Filter-out irrelevant bins based on the content of Ns.
5. Define the null hypothesis for the statistical test. For single-loci resolution methylation biomarkers, we chose to count the number of CG dinucleotides per bin.
6. Perform the statistical test.
   (a) Using an exact multinomial goodness of fit test, performed through a Monte Carlo simulation with the R package XNomial[1], verify if signature loci are distributed among TADs according to the null model probabilities given by the proportion of dinucleotides CG present in each bin, $p_i^{(0)} = N_i^{\mathrm{CG}}/N_{\mathrm{T}}^{\mathrm{CG}}$, where N_i^{CG} and $N_{\mathrm{T}}^{\mathrm{CG}}$ are the numbers of dinucleotides CG in the i-th bin and in all the genome, respectively.

[1] https://cran.r-project.org/package=XNomial

Table 1: Experimental Data.

No.	chr	start	end	Bin type	No. hits	pvalue	Genes	miRNAs
1	chr7	1480000	2799999	TAD	26	6.27E-10	INTS1, MAFK, TMEM184A, PSMG3, ELFN1, MAD1L1, FTSJ2, NUDT1, SNX8, EIF3B, CHST12, GRIFIN, LFNG, BRAT1,IQCE, TTYH3, AMZ1, GNA12	MIR4655, MIR6836, MIR4648
2	chr2	176080000	176599999	TAD	11	1.13E-08	EVX2, HOXD13, HOXD12, HOXD11, HOXD10, HOXD9, HOXD8, HOXD3, AC009336.19, HOXD4, HOXD1, MTX2	MIR10B, MIR7704
3	chr7	4640000	4799999	GAP	9	4.79E-08	FOXK1, AP5Z1, RADIL	
4	chr7	1	1199999	GAP	22	1.29E-07	FAM20C, WI2-2373I1.1, PDGFA, PRKAR1B, DNAAF5, SUN1, GET4, ADAP1, COX19, CYP2W1, C7orf50, GPR146, GPER1, ZFAND2A	MIR339
5	chr4	1	1679999	GAP	23	1.55E-07	ZNF595, ZNF718, ZNF732, ZNF141, ZNF721, PIGG, RP11-1263C18.3, PDE6B, ATP5I, MYL5, MFSD7, PCGF3, CPLX1, GAK, TMEM175, DGKQ, SLC26A1, IDUA, FGFRL1, RNF212, SPON2, CTBP1, MAEA, UVSSA, NKX1-1, FAM53A	MIR571
6	chr16	87280000	90338345	GAP	35	1.79E-07	C16orf95, RP11-178L8.4, FBXO31, MAP1LC3B, ZCCHC14, JPH3, KLHDC4, SLC7A5, CA5A, BANP, ZNF469, ZFPM1, ZC3H18, IL17C, CYBA, MVD, SNAI3, RNF166, CTU2, PIEZO1, CDT1, APRT, GALNS, TRAPPC2L, PABPN1L, CBFA2T3, ACSF3, CDH15, SLC22A31, ZNF778, ANKRD11, SPG7, RPL13, CPNE7, DPEP1, CHMP1A, SPATA33, CDK10, SPATA2L, VPS9D1, ZNF276, FANCA, SPIRE2, TCF25, MC1R, RP11-566K11.2, TUBB3, DEF8, RP11-566K11.6, DBNDD1, GAS8, PRDM7	MIR6775, MIR5189, MIR4722
7	chr3	48200000	48639999	TAD	10	2.34E-07	CAMP, ZNF589, NME6, SPINK8, FBXW12, PLXNB1, CCDC51, TMA7, ATRIP, TREX1, SHISA5, PFKFB4, UCN2, COL7A1, UQCRC1, TMEM89, SLC26A6, CELSR3	

(b) If the p-value of the multinomial test is < 0.05, run a post-hoc test to determine bins responsible for the significant deviation from the expected number of hits. This is carried out for each bin using an exact binomial test.

7. Rank the bins significantly enriched with biomarkers based on the obtained p-values.
8. Annotate the enriched bins with coding and non-coding genes from the GENCODE project (`https://www.gencodegenes.org`).

All statistical and computational analysis were performed in R version 3.4.4.

3 Results

This section comments the results we obtained after applying the pipeline proposed in section 2.2 to the methylation biomarker signatures obtained for Lynch Syndrome from the statistical analysis detailed in section 2.1. As TAD boundaries, we chose the TAD predictions for the Lymphoblastoid cell line GM12878 (GEO number GSM1608505) provided by the method GITAR [9] [2].

Exact multinomial goodness-of-fit test gave a p-value $< 10^{-5}$, with a Monte Carlo simulation of 10^6 trials. The pipeline provides a table with TADs ordered according to ascending p-values together with their annotation (see Table 1).

A detailed look at the results provided in Table 1 yields many interesting annotations for the most significant TADs obtained with the proposed method. For example, TAD 1

[2]Data available at: `https://data.genomegitar.org`

in Table 1 contains many genes involved in inflammatory disease, immunity and tumors: MAD1L1, shown to be differentially methylated in CD4+ cells in inflammatory disease [10]; ELFN1-AS1, a novel primate antisense RNA gene expressed predominantly in tumors [11]; CHST12, for which hypomethylation of a CG site has been shown to have a sensitivity of 86% and specificity of 64% in detecting renal disease in patients with Lupus [12]; GRIFIN, which plays a role in self/non-self recognition receptors in innate and adaptive immunity [13]; LFNG shown to alter cytokine production and to be involved in allergic diseases [14]; GNA12, known to be differentially expressed in inflammatory bowel disease and ulcerative colitis [15].

TAD 2 contains the HOXD cluster, known to be differentially expressed upon activation of leukocyte sub-populations [16], including gene HOXD11, aberrantly methylated in breast cancer [17].

Our tool also provides annotations for ncRNA genes, which are also very useful for identifying potential functionality associated to the TADs where the differentially methylated loci are located. For example, the first TAD from the results Table contains the following microRNAs which are of interest for LS: miR-6836 acts as a serum marker for pancreatobiliary cancers [18]; miR-4648 is disregulated in colon cancer [19]; miR-339-5p inhibits breast cancer cell migration and invasion [20]; miR-3176 is associated with cervical cancer [21]; miR-662 is associated with invasive lung carcinoma [22]; miR-711 has an oncogenic role in breast cancer [23].

Furthermore we studied homogeneity in methylation patterns in three relatively small TADs among the 10 most significatives, TADs 2,7, and 9 (see Table 1), by estimating average β values for cases and controls. Actually, we found that in all three TADs considered cases present a prevalence of hypomethylated CG sites with respect to controls.

A detailed analysis of the results provided by this tool is to be completed with the clinicians involved in the study.

4 Conclusion

We described a computational and statistical tool intended to analyze TADs and pin down the ones with most significant presence of biomarkers of interest. In particular, to show the performance of the tool, we used methylation sequencing data from a pilot study aimed at identifying differentially methylated CG sites in subjects with genetic mutations related to LS. We selected a set of differentially methylated loci between cases and controls and find out partitions of the genome significantly enriched with these sites. We then analyzed annotations for protein coding genes and miRNAs of the most relevant TADs, devising interesting links to be further investigated.

Acknowledgments

This research has been funded by projects DPI2017-84439-R of MINECO, Madrid, Research Grant PI-0862-2012, Junta de Andalucia-Fundación Progreso y Salud and European Union FEDER.

References

[1] A. Pombo and N. Dillon, "Three-dimensional genome architecture: players and mechanisms," *Nature reviews Molecular cell biology*, vol. 16, no. 4, p. 245, 2015.

[2] M. D. Gallagher, M. Posavi, P. Huang, T. L. Unger, Y. Berlyand, A. L. Gruenewald, A. Chesi, E. Manduchi, A. D. Wells, S. F. Grant, G. A. Blobel, C. D. Brown, and A. S. Chen-Plotkin, "A dementia-associated risk variant near tmem106b alters chromatin architecture and gene expression," *The American Journal of Human Genetics*, vol. 101, no. 5, pp. 643 – 663, 2017.

[3] J. Dekker and E. Heard, "Structural and functional diversity of topologically associating domains," *FEBS letters*, vol. 589, no. 20PartA, pp. 2877–2884, 2015.

[4] M. Forcato, C. Nicoletti, K. Pal, C. M. Livi, F. Ferrari, and S. Bicciato, "Comparison of computational methods for Hi-C data analysis," *Nature Methods*, vol. 14, p. 679, July 2017.

[5] R. Dali and M. Blanchette, "A critical assessment of topologically associating domain prediction tools," *Nucleic acids research*, vol. 45, no. 6, pp. 2994–3005, 2017.

[6] G. P. Way, D. W. Youngstrom, K. D. Hankenson, C. S. Greene, and S. F. Grant, "Implicating candidate genes at gwas signals by leveraging topologically associating domains," *European Journal of Human Genetics*, vol. 25, no. 11, p. 1286, 2017.

[7] K. Chang, M. W. Taggart, L. Reyes-Uribe, E. Borras, E. Riquelme, R. M. Barnett, G. Leoni, F. A. San Lucas, M. T. Catanese, F. Mori, *et al.*, "Immune profiling of premalignant lesions in patients with lynch syndrome," *JAMA oncology*, 2018.

[8] H. Binder, L. Hopp, M. R. Schweiger, S. Hoffmann, F. Jühling, M. Kerick, B. Timmermann, S. Siebert, C. Grimm, L. Nersisyan, *et al.*, "Genomic and transcriptomic heterogeneity of colorectal tumors arising in lynch syndrome," *The Journal of pathology*, 2017.

[9] R. Calandrelli, Q. Wu, J. Guan, and S. Zhong, "Gitar: an open source tool for analysis and visualization of hi-c data," *bioRxiv*, 2018.

[10] T. Hughes, F. Ture-Ozdemir, F. Alibaz-Oner, P. Coit, H. Direskeneli, and A. H. Sawalha, "Epigenome-wide scan identifies a treatment-responsive pattern of altered dna methylation among cytoskeletal remodeling genes in monocytes and cd4+ t cells from patients with behçet's disease," *Arthritis & Rheumatology*, vol. 66, no. 6, pp. 1648–1658, 2014.

[11] A. Kozlov, "Expression of evolutionarily novel genes in tumors," *Infectious agents and cancer*, vol. 11, no. 1, p. 34, 2016.

[12] E. Weeding and A. H. Sawalha, "Deoxyribonucleic acid methylation in systemic lupus erythematosus: Implications for future clinical practice," *Frontiers in Immunology*, vol. 9, p. 875, 2018.

[13] G. R. Vasta, H. Ahmed, M. Nita-Lazar, A. Banerjee, M. Pasek, S. Shridhar, P. Guha, and J. A. Fernández-Robledo, "Galectins as self/non-self recognition receptors in innate and adaptive immunity: an unresolved paradox," *Frontiers in immunology*, vol. 3, p. 199, 2012.

[14] S. Mukherjee, A. J. Rasky, P. A. Lundy, N. A. Kittan, S. L. Kunkel, I. P. Maillard, P. E. Kowalski, P. C. Kousis, C. J. Guidos, and N. W. Lukacs, "Stat5-induced lunatic fringe during th2 development alters delta-like 4–mediated th2 cytokine production in respiratory syncytial virus–exacerbated airway allergic disease," *The Journal of Immunology*, vol. 192, no. 3, pp. 996–1003, 2014.

[15] C. A. Anderson, G. Boucher, C. W. Lees, A. Franke, M. D'Amato, K. D. Taylor, J. C. Lee, P. Goyette, M. Imielinski, A. Latiano, *et al.*, "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47," *Nature genetics*, vol. 43, no. 3, p. 246, 2011.

[16] R. Morgan and K. Whiting, "Differential expression of hox genes upon activation of leukocyte sub-populations," *International journal of hematology*, vol. 87, no. 3, pp. 246–249, 2008.

[17] K. Miyamoto, T. Fukutomi, S. Akashi-Tanaka, T. Hasegawa, T. Asahara, T. Sugimura, and T. Ushijima, "Identification of 20 genes aberrantly methylated in human breast cancers," *International journal of cancer*, vol. 116, no. 3, pp. 407–414, 2005.

[18] K. T. Lee, "The clinical utility of microrna as a prognostic biomarker of pancreatobiliary cancers," *Gut and liver*, vol. 10, no. 5, pp. 663–664, 2016.

[19] X. Wu, S. Li, X. Xu, S. Wu, R. Chen, Q. Jiang, Y. Li, and Y. Xu, "The potential value of mir-1 and mir-374b as biomarkers for colorectal cancer," *International journal of clinical and experimental pathology*, vol. 8, no. 3, p. 2840, 2015.

[20] Z.-s. Wu, Q. Wu, C.-q. Wang, X.-n. Wang, Y. Wang, J.-j. Zhao, S.-s. Mao, G.-h. Zhang, N. Zhang, and X.-c. Xu, "Mir-339-5p inhibits breast cancer cell migration and invasion in vitro and may be a potential biomarker for breast cancer prognosis," *BMC cancer*, vol. 10, no. 1, p. 542, 2010.

[21] A. Pedroza-Torres, J. Fernández-Retana, O. Peralta-Zaragoza, N. Jacobo-Herrera, D. C. de Leon, J. F. Cerna-Cortés, C. Lopez-Camarillo, and C. Pérez-Plasencia, "A microrna expression signature for clinical response in locally advanced cervical cancer," *Gynecologic oncology*, vol. 142, no. 3, pp. 557–565, 2016.

[22] M. Filipska, M. Skrzypski, K. Czetyrbok, T. Stokowy, G. Stasiłojć, A. Supernat, J. Jassem, A. Zaczek, and J. Bigda, "Mir-192 and mir-662 enhance chemoresistance and invasiveness of squamous cell lung carcinoma," vol. 118, 02 2018.

[23] J.-Y. Hu, W. Yi, M.-Y. Zhang, R. Xu, L.-S. Zeng, X.-R. Long, X.-M. Zhou, X.-F. S. Zheng, Y. Kang, and H.-Y. Wang, "Microrna-711 is a prognostic factor for poor overall survival and has an oncogenic role in breast cancer," *Oncology letters*, vol. 11, no. 3, pp. 2155–2163, 2016.