Genome supranucleosomal organization and genetic susceptibility to diseases

K. P. Jablonski, C. Fretter, L. Carron, T. Forné, M.-T. Hütt, and A. Lesne

Citation: AIP Conference Proceedings **1882**, 020027 (2017); doi: 10.1063/1.5001606
View online: http://dx.doi.org/10.1063/1.5001606
View Table of Contents: http://aip.scitation.org/toc/apc/1882/1
Published by the American Institute of Physics

Genome Supranucleosomal Organization and Genetic Susceptibility to Diseases

K. P. Jablonski[1], C. Fretter[1], L. Carron[2], T. Forné[3], M.-T. Hütt[1], and A. Lesne[2,3,a)]

[1] *Jacobs University, Bremen, Germany*
[2] *LPTMC, CNRS, UPMC Sorbonne Université, Paris, France*
[3] *IGMM, CNRS, Univ. Montpellier, Montpellier, France*

[a)] Corresponding author: lesne@lptmc.jussieu.fr

Abstract. The notion of disease-associated single-nucleotide polymorphisms (da-SNP), as determined in genome-wide association studies (GWAS), is relevant for many complex pathologies, including cancers. It appeared that da-SNPs are not only markers of causal genetic variation but may contribute to the disease development through an influence on gene expression levels. We argue that understanding this possible functional role of da-SNPs requires to consider their embedding in the tridimensional (3D) multi-scale organization of the human genome. We then focus on the potential impact of da-SNPs on chromatin loops and recently observed topologically associating domains (TADs). We show that for some diseases and cancer types, da-SNPs are over-represented in the borders of these topological domains, in a way that cannot be explained by an increased exon density. This analysis of the distribution of da-SNPs within the 3D genome organization suggests candidate loci for further experimental investigation of the mechanisms underlying genetic susceptibility to diseases, in particular cancer.

INTRODUCTION

In recent years, *genome-wide association studies* (GWAS) have been performed by international scientific consortiums on large cohorts of patients and healthy people in order to disentangle the genetic components from other factors, typically environmental factors, in the etiology of complex diseases, including cancer [1, 2]. These studies statistically relate *single-nucleotide polymorphisms* (SNPs) with disease development, but provide no immediate explanation of the role of the genetic variants in the biological mechanisms responsible for the disease [3, 4]. The need to go beyond statistical association and unravel functional aspects of cancer risk loci has been rapidly underlined [5]. While da-SNPs have originally been considered only as markers of the genetic variations causally involved in the disease, their possible direct implication in the pathologies, mainly through a dysregulation of gene expression levels, is now considered. On the other hand, recent experimental advances dramatically improved the knowledge of 3-dimensional genome organization and its functional role in transcriptional regulation [6]. Joining these two research domains opens a novel research direction addressing the role of genome architecture and its modifications in understanding the genetic risk to diseases [7]. We here briefly review the state-of-the-art about these questions, then provide statistical evidences showing how insights on the functional consequences of da-SNPs can be gained by considering their surrounding 3D genome organization as well as distal effects of genetic variations in regulatory elements.

GENOME-WIDE ASSOCIATION STUDIES AND CANCER SUSCEPTIBILITY LOCI

Disease-Associated Single-Nucleotide Polymorphisms (da-SNPs)

Single-nucleotide polymorphisms are genetic loci (single base pairs) where a variation between individuals has been observed in the human genome. They have been extensively investigated as markers of human lineages and

Physics of Cancer: Interdisciplinary Problems and Clinical Applications
AIP Conf. Proc. 1882, 020027-1–020027-11; doi: 10.1063/1.5001606
Published by AIP Publishing. 978-0-7354-1562-1/$30.00

migrations [8]. The identification of these elementary genetic variations has also been a starting point to unravel the molecular bases of the genetic risk to develop various diseases [2]. Genome-wide association studies (GWAS) have quantified the statistical association between a disease and a SNP, that is, individuals harboring a variant allele at this single-nucleotide location are significantly more affected by the disease than individuals harboring another allele. In general, the disease-associated allele is a minor allele, with a few exceptions [9]. Such *disease-associated SNPs* (da-SNPs) are characterized by a *p*-value measuring the statistical significance of the association, and an effect size measured by the odds ratio. As a rule, the size of the cohorts determines a threshold on the effect size under which the association cannot be statistically significant (the larger the cohorts, the smaller this effect-size threshold). Accordingly, GWAS are only able to detect common disease-associated variants [10].

Cancer-associated SNPs generally depend on the cancer type (affected organ), although pleiotropic associations have been found [11]. Note that GWAS investigate mutations present at birth in the genome of an individual and influencing the probability to develop a cancer later on, what is termed *cancer genetic susceptibility* or *genetic risk*. These mutations are different from somatic mutations that accumulate in cancer cells, provide a signature of their pathological state and drive cancer progression.

At first, da-SNPs were considered only as markers of the causal genetic variation, which lies in the variation (possibly more complex than a point mutation) of a gene [12]. This interpretation is based on *linkage disequilibrium* [13], that is, the fact that genetic variations are locally correlated within haplotype blocks. One can note that the size of these blocks strongly depends on the ethnic origin of the individuals, due to the varying duration of the recombination history in different populations (mean size 11kb in Yoruban and African-American samples and 22 kb in European and Asian samples) [14]. These correlations are actually exploited to impute additional associations from the observed ones (*imputed da-SNPs*). Moreover, we observed in our analysis of the GWAS catalog (gathering all da-SNPs from various studies, see Methods section), compared to the current knowledge of human polymorphisms [15, 16], that is less than 1% of all possible SNPs that are associated with a disease, meaning that the association reflects a nontrivial and non-typical biological situation.

While the results of these studies are available from the public GWAS catalog [17], the limits of GWAS and the difficulties in the interpretation of da-SNPs have been pointed out [18]. We observe that the number of SNPs associated to the same disease in the GWAS catalog, e.g. a cancer in a specific organ, is in the range of one hundred (depending on the thresholds on *p*-value and effect size in the definition of a da-SNP) (Fig. 1). This indicates that a single da-SNP is not sufficient to determine the appearance (or non-appearance) of a disease, but only participates, together with numerous other genetic and non-genetic factors, in the biological mechanisms leading to the pathology. Therefore, more mechanistic understanding of the disease association is required, in particular for the development of personalized medicine [19] and in oncology [5, 20]. Thanks to the 1000 Genome Project [15, 16] and similar works, the knowledge of the genome of individuals became available, allowing a better characterization of da-SNPs, and opening the way towards experimental functional investigations. Globally, a novel view thus emerged according to which SNPs can have functional consequences [5, 21–23] and directly increase the propensity to develop a pathology [24]. One can note that the modification of the expression of a gene due to the variant form of a SNP can be further amplified by the gene regulatory network, thus leading to a major functional disturbance.

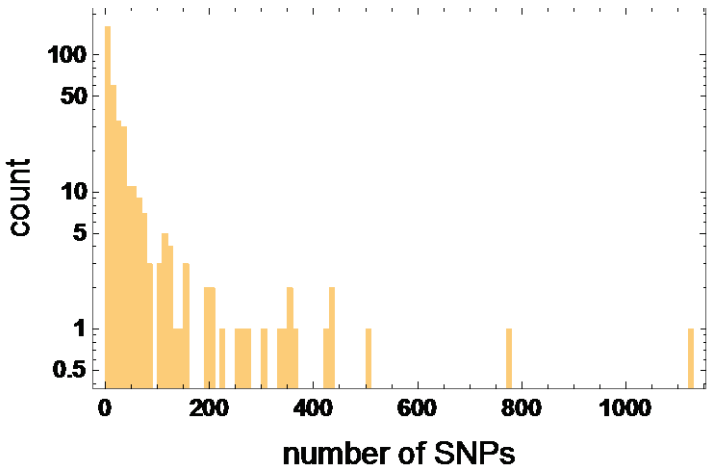

FIGURE 1. Histogram of the number of da-SNPs per disease. It is drawn over all diseases considered in the GWAS catalog

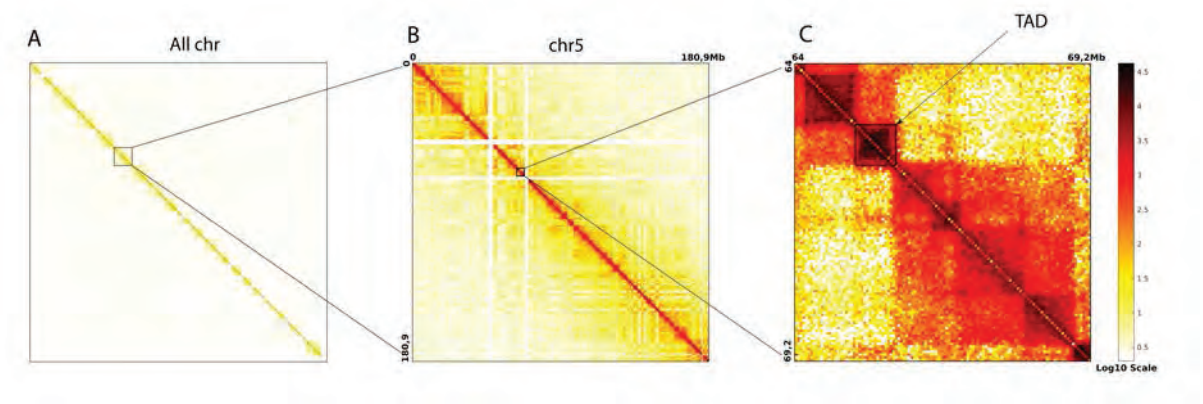

FIGURE 2. Topologically associating domains (TADs) in the human genome as derived from Hi-C data. The figure sketches how genome-wide chromosome conformational capture data (Hi-C contact maps, data from [34]) evidence the existence of disjoint self-contacting domains, known as topologically associating domains (TADs). The full contact map (i.e. of all chromosomes, resolution 160kb) for human embryonic stem cells is displayed on panel A, and a zoom on chromosome 5 on panel B (resolution 80 kb). The color indicates the contact frequency (the darker the more frequent). On panel C, the partition of the genome in TADs is illustrated for a part of chromosome 5, with an additional zoom (resolution 40 kb): TADs appear in the form of disjoint dark triangles along the diagonal of the contact map. One of them is underlined in black as an example. Fuzzy sub-structures of smaller size within TADs may feature chromatin loops

Different Relationships to Disease

A basic distinction lies in the dysfunction involved in the pathology, which can be due either to a defective biological factor (change in the sequence of a protein or non-coding functional RNA), or to a dysregulation of the amount of some relevant factors (typically their expression levels). This distinction reflects in the possible functional consequences of a SNP, as follows.

A first class, *coding SNPs*, corresponds to da-SNPs located in exons hence affecting the sequence of a protein [25]. In this case, the protein structure and presumably the protein function are affected, and the causal role of the SNP is thus straightforward. This class can be extended including *pseudo-coding SNPs,* located in genomic regions that are transcribed into various non-coding RNAs (tRNA, microRNA, siRNA) [26]. Another situation of SNPs having a direct impact on a protein structure are mutations disrupting the proper splicing events or affecting protein translation when located in 3' or 5' UTR [27].

The second class is composed of *non-coding da-SNPs*, located in introns or intergenic regions. Their association to diseases is expected to lie in the modification, in individuals harboring the variant allele, of some levels of gene expression, and as such they are generally termed *regulatory SNPs* [21, 28]. It actually appeared that most da-SNPs (more than 90%) belong to this class [29–31]. A simple sub-class of regulatory SNPs is composed of da-SNPs located in the binding site of a regulatory protein, typically a transcription factor, so that the mutation (i.e. the variant allele) directly disturbs the transcription initiation of the gene regulated by this transcription factor, and accordingly its expression level [9, 32]. We will focus on another, more recently evidenced, sub-class composed of *architectural SNPs,* where the SNP variation first reflects in a local variation of the genome spatial conformation, which in turn affects some gene expression levels [7].

Previous approaches on cancer-associated SNPs typically used gene expression data to infer which SNPs may affect gene expression levels (eQTLs) and to delineate the associated pathways [9, 33]. We here adopt a complementary viewpoint and investigate da-SNPs, and in particular cancer-associated SNPs, in the context of the tri-dimensional (3D) genome organization.

3D GENOME ORGANIZATION AND GENETIC RISK

3D Genome Organization and Its Functional Role

Recently developed techniques of chromosome conformation capture combine chemical crosslinking and sequencing to identify genomic loci contacting each other *in vivo*. They have shown that the mammalian genome

displays three main architectural features at the large-scale level (supranucleosomal level, beyond the kb scale), nested in a hierarchical way (Fig. 2): *chromatin loops*, *topologically associating domains* (TADs) of larger size exhibiting more internal contacts than contacts between domains [34, 35], and a segregation in *active and inactive compartments* [36].

TADs existence and distribution, as well as the genome segregation into active and inactive compartments have been unraveled using genome-wide Hi-C. TADs organize the genome into a modular and presumably functional structure at the sub-megabase scale [37]. Chromatin loops and TADs have also been investigated experimentally using local quantitative and high-resolution chromosome conformation capture technique (3C-qPCR) [38, 39]. Such experiments using 3C-qPCR clarified the distinction between the TADs as originally evidenced in [34, 35], of size 200 kb to 1 Mb (median size about 800 kb) and structures of various sizes termed *chromosome contact domains* [40], among which the smallest ones rather correspond to chromatin loops embedding a single gene and its proximal regulatory sequences [41].

The above-mentioned architectural features are closely involved in the regulation of gene expression [6], and architectural changes are observed in pathologies. Significant changes in genomic contacts have been observed in bladder cancer and lymphomas [42], as well as in breast cancer [43, 44], revealing altered chromatin architecture. Another emblematic example is the observed disruption of chromatin enhancer-promoter loops at specific loci in β-thalassemia (β-globin gene, [7]) and more generally erythroid pathologies (MYB gene, [45]). Gene-specific multi-enhancer contacts playing a key role in this 3D genome spatial organization are similarly likely to be mutation-sensitive loci [46]. More generally, bioinformatic analyses show that a large number of da-SNPs (over all diseases) are located at enhancers inside TADs [31].

Changes in genome architecture, in particular TAD disruption, have also been evidenced in various cancers [47, 48]. The changes in long-range chromatin interactions associated with either the fusion of adjacent TADs or the creation of new TADs are potentially related to cancer progression, due to the ensuing dysregulation of oncogenes and tumor suppressors [49]. Specifically, TAD boundaries have been shown to be altered (microdeletion, mutations or epimutations in the binding sites of the architectural protein CTCF) in T-cell acute lymphoblastic leukemia (T-ALL) [50] and in gliomas [51]. More generally, variations of chromatin architectural patterns have been observed between individuals [52]. This prompts to investigate their possible relationships with genetic variations, possibly mediated by variation in the epigenetic marks [48], and their functional correlates.

Architectural SNPs

The understanding of the functional role of the 3D genome organization triggers an increasing interest in non-coding da-SNPs, whose effect on gene regulation is mediated by a change in chromatin 3D organization [7]. Such a scenario represents an instance of chromatin allostery, where the chromatin itself behave as an allosteric object [53]; here the allosteric transition would be triggered by the SNP and the effector would be some event necessary for transcription, e.g. TF binding or the formation of a promoter-enhancer loop [54].

These architectural SNPs can have an impact at multiple scales. A basic instance is a SNP in the binding site of the architectural protein CTCF (CCCTC-binding Factor), which have been shown to affect both loops and TAD boundaries [55], with stronger binding sites at TAD boundaries [56]. Another frequent instance is a SNP located in an enhancer, and whose variation affects the proper formation of enhancer-promoter loop [31, 57]. For instance, some risk loci for epithelial cancers, including colon, breast, and prostate cancers, have been found at enhancers forming a long-range chromatin loop with the MYC proto-oncogene in a tissue-specific way [58]. It is has been shown experimentally, using chromosome conformation capture techniques, that a single SNP can actually disrupt an enhancer-promoter loop and have deleterious effects through a dysregulation of the associated gene, potentially leading to disease, e.g. autoimmune diseases [59], asthma [60] or erythroid pathologies [45]. More generally, certain SNPs located in regulatory regions may affect not only the expression of nearby genes, e.g. through a modification of epigenetic marks, but also the expression of distant genes located in *cis* (that is, on the same chromosome, as opposed to *trans* acting factors) in the same TAD. This has been shown for SNPs associated with autoimmune diseases, with distal molecular coordination effects of range about 50 kb, within the same TAD [61].

We will here investigate the distribution of da-SNPs with respect to TADs and more specifically TAD borders. We here call *TAD border* the region limiting a TAD and across which very few physical contacts occur, thus corresponding to a marked insulation (slightly different from and more general than what is called *TAD boundary* in [34], namely the region separating two TADs along the genome). These borders can be observed using the 3C-qPCR technique (see e.g. the case of the *HoxD* locus in [41]). Certain TAD borders are involved in gene regulation events.

The emblematic example is the case of *Hox* genes, where a displacement of a TAD border at the *HoxD* locus induces a switch of the Hoxd genes (from *Hoxd11* to *Hoxd8*) from a telomeric TAD to a more centromeric TAD thus changing their expression levels [62]. Therefore, regulation of gene expression can be altered by the presence of a SNP variant as soon as this mutation is able to disrupt the border. Engineered genetic alteration of TAD borders using CRISPR/Cas9 at the *WNT6/IHH/EPHA4/PAX3* locus induces limb malformation in mice; specifically, mutations in the binding sites of the architectural protein CTCF have been shown to cause the appearance of long-range interactions between some promoters of a TAD and enhancers located in the adjacent one, reflecting topologically in the fusion of adjacent TADs and functionally in the abnormal expression of these genes [63].

STATISTICAL INVESTIGATION

We performed a statistical investigation of the relationship between disease association (that is, genetic risk) and location with respect to TAD borders, considering first all da-SNPs then focusing the analysis to *cancer-associated SNPs* (that is, SNPs whose variant form is statistically associated with an increased risk of developing a cancer, not to be confused with somatic mutations accumulating in cancer cells during cancer development). We aim at a novel annotation of da-SNPs based on their potential impact on the functional 3D chromatin architecture, which could suggest candidate loci for a further experimental and mechanistic study.

Methods

We used the compilation of GWAS data freely available in the GWAS catalog (www.ebi.ac.uk/gwas/, version r2017-06-26) [17]. We translated the genomic coordinates given in the GWAS catalog (Genome assembly GRCh38.p10, in short hg38 reference) into the hg19 reference. We used both directly observed and imputed da-SNPs, where imputation is based on the belonging to the same haplotype block than an observed association. We constructed the list of da-SNPs and the subset of cancer-associated SNPs by filtering the EFO label (Experimental Factor Ontology) indicating the trait(s) or disease(s) to which each SNP of the catalog has been significantly associated. Only diseases (including cancers) displaying at least 10 associations are considered. The da-SNPs of the list are moreover annotated with their genomic location in exons, introns or intergenic regions.

TAD coordinates have been determined in [34] for human foetal lung fibroblasts (IMR90 cells) and human embryonic stem cells (hESC), from two replicates, with only minor variations between the two cell types. We used the data corresponding to hESC (human embryonic stem cells), for which more annotations are available). To account for the topological characterization of borders seen as regions across which the contact frequency displays a marked decrease [41], we defined TAD borders as zones of 20kb located inside the TAD at the limits of this TAD (recall that the median size of a human gene is 23 kb [64]). Results remain stable under variation of the border size.

Enrichment in da-SNPs of a given genomic region (e.g. TAD borders) has been assessed for each disease using a uniformly random null model, described by a binomial distribution of parameters the fraction of base pairs in the given genomic region and the total number of SNPs associated to the disease (this distribution corresponds to the large-size limit of the currently used hyper-geometric distribution, given that the total number of SNPs in the human genome is of order of tens of millions—the number is evolving each year as new results accumulate [15, 16]). Enrichment p-value is then computed as the probability (cumulative binomial distribution) to get by chance at least the observed number of da-SNPs in the considered genomic region. Multiple testing (over the different diseases in the GWAS catalog, considering independently cancer types and non-cancer diseases) was corrected for using the Benjamini-Hochberg procedure [65, 66]. A (corrected) p-value smaller than 0.05 assesses the statistical significance of the enrichment. The number of da-SNPs present in the TAD borders gives a measure of the effect size.

The active compartment has been determined at a resolution of 40 kb, by an analysis of the principal eigenvector of the contact correlation map that refines the method in [36] (data from [34], normalized according to [67]). The enrichment p-value has been computed using a uniform random null model and a cumulative binomial distribution.

TABLE 1. Distribution of da-SNPs. The table indicates the location of da-SNPs in the different genomic compartments, according to the data gathered in the GWAS catalog (over all diseases).

da-SNPs	In exons	In introns	Intergenic
	8%	47.3%	44.7%

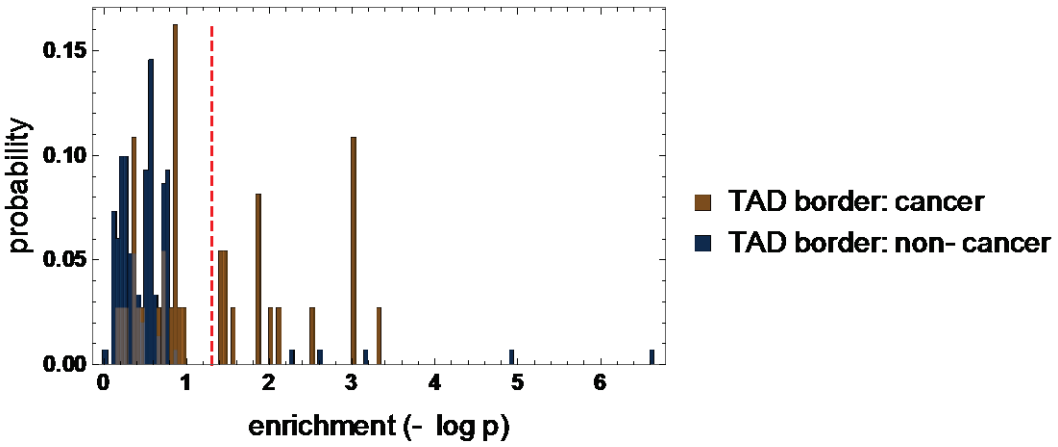

FIGURE 3. TAD border enrichment in disease-associated SNPs. The normalized histograms of (minus the logarithm of) *p*-values assessing TAD border enrichment for all non-cancer diseases (blue) and all cancer types (yellow) are compared (TAD border of size 20kb inwards). The grey color corresponds to the overlap of the two histograms. The vertical line indicates the significance threshold at $p = 0.05$. It means that for each disease and each cancer type contributing to the right part of the histograms, the associated genetic risk loci display a significant over-representation in TAD borders

Location of Genetic Risk Loci with Respect to Topological Domain Borders

We first observed in the GWAS catalog (for all diseases investigated in GWAS) that only 8% of da-SNPs are located in exons, in agreement with the evaluation given in [7, 31]. This number confirms that most da-SNPs are non-coding and their potential effect lies in a dysregulation of gene expression. As such, these da-SNPs belong to the class of *expression quantitative traits loci* (eQTLs, [9, 52]). It is to note that introns may contain enhancers for another gene [68], so that intronic da-SNPs may participate to the (dys)regulation of another gene.

We then performed a statistical investigation of the location of da-SNPs with respect to genomic architectural features, considering first all diseases, then more specifically cancers. To detect a possible impact of some da-SNPs on TAD delimitation, we considered *TAD borders*, namely regions surrounding the limits of the TADs, and across which a step-wise decrease of the contact frequency occurs [41] (see section Methods). We found that for a small fraction of diseases, da-SNPs are over-represented in TAD borders (Fig. 3).

TABLE 2. Cancer types displaying a significant enrichment of TAD borders in da-SNPs. For each of these 12 cancer types, the total number of da-SNPs and the number of da-SNPs belonging to a TAD border are indicated. The list of cancer-associated SNPs located in TAD borders for these cancer type, is given in Table 4.

EFO term	Cancer type	p-value (TAD border enrichment)	Number of da-SNPs	Number of da-SNPs in TAD borders
EFO_0000095	Chronic lymphocytic leukemia	0.0043	124	13
EFO_0000178	Gastric cancer	0.0157	18	4
EFO_0000182	Hepatocellular carcinoma in hepatitis B infection	0.0145	10	1
EFO_0000305	Breast cancer	0.0055	244	22
EFO_0000571	Lung adenocarcinoma	0.0013	25	1
EFO_0000756	Melanoma	0.0055	55	13
EFO_0001071	Lung cancer	0.0034	50	7
EFO_0001075	Ovarian cancer	0.0002	35	8
EFO_0005088	Testicular germ cells cancer	0.0157	41	7
EFO_0005570	Oral cavity cancer	0.0145	9	3
EFO_0005842	Colorectal cancer	0.0112	181	12
EFO_1000650	Breast cancer (estrogen-receptor negative)	0.0004	27	3

TABLE 3. Preferential location of cancer-associated SNPs in the active compartment. For cancer types displaying a TAD-border enrichment (Table 2), the genetic risk loci located in TAD borders (see Table 4) are over-represented in the active compartment (which represents only 39.9% of the genome).

Cancer associated SNPs (for cancer types in Table 2)	Total number	Number in the active compartment	%	p-value of the enrichment
	73	47	64.4%	7.21×10^{-6}

We more specifically compared the set of cancers and the set of other diseases (Fig. 3). Both sets contain diseases for which TAD border enrichment is observed, and which can be identified in our analysis. This result supports a role of TAD borders and their modifications in these particular diseases, and suggests further investigation of the corresponding da-SNPs and surrounding loci. Noticeably, our results show that a significant enrichment of TAD borders in da-SNPs is present only for specific diseases and cancers, and thus represents a nontrivial feature. Noticeably, cancers dominate among the diseases displaying such TAD-border enrichment. The corresponding cancer types are listed in Table 2, together with the number of associated SNPs, while the list of the non-cancer diseases can be found at http://www.lptl.jussieu.fr/user/lesne/disease_list1.txt.

It has been observed in [34] that inter-TAD regions (termed TAD boundaries) are enriched in housekeeping genes. However, this fact cannot explain the association of embedded SNPs to a specific disease, since a flaw in a housekeeping gene would produce far more defects than a specific pathology. Our result about TAD borders enrichment in da-SNPs for only a small subset of diseases points at the existence of more specific mechanisms.

The cancer-associated SNPs located in TAD borders, for cancer types displaying TAD-border enrichment, appear to be preferentially located in the active compartment (Table 3). Moreover, the enrichment of TAD borders is mainly due to non-exonic SNPs (Fig. 4). Over-representation in TAD borders of exonic SNPs is observed for only two cancer types (gastric and ovarian cancers). Interestingly, exonic and non-exonic enrichments exclude one another, which points at two distinct mechanisms. Overall, these results support a functional interpretation of these SNPs as being involved in the deregulation of gene expression, presumably through a change in the TAD borders, and promote further experimental investigation to specify and validate this interpretation.

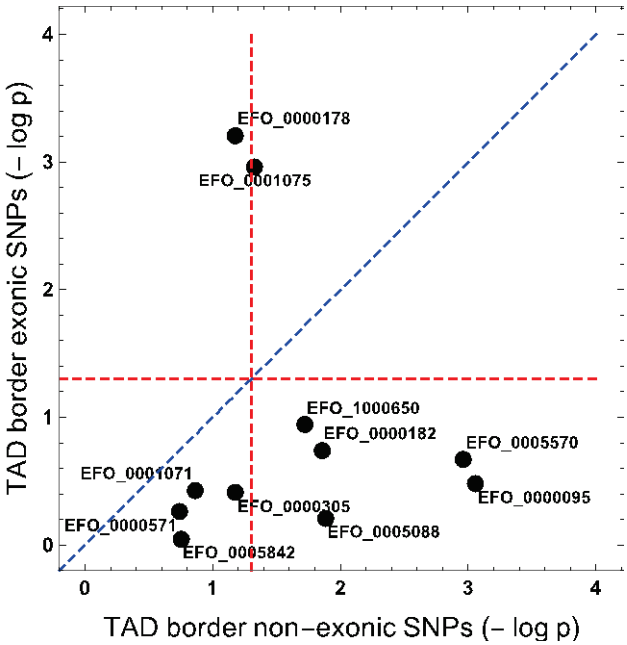

FIGURE 4. Enrichment of TAD borders in exonic and non-exonic cancer-associated SNPs. For each cancer type, described by its EFO label (Experimental Factor Ontology), the TAD border enrichment p-value is computed separately for the subsets of exonic and non-exonic da-SNPs. Enrichment in exonic SNPs is observed only for two cancer types (gastric and ovarian cancers). Non-exonic enrichment is observed for breast, oral cavity and testicular germ cell cancer, chronic lymphocytic leukemia, and hepatocellular carcinoma. The dotted red lines indicate the significance threshold at $p = 0.05$. The dotted blue line indicates equal p-values for exonic and non-exonic subsets

TABLE 4. Candidate architectural da-SNPs. The table provides a list of the cancer-associated SNPs located in TAD borders for cancer types displaying a significant TAD border enrichment in da-SNPs (see Table 1). Note that the number of SNPs associated to a cancer may be different from Table 2, in which the cancer-associated SNPs observed in independent studies were considered as different entries, like it is done in the GWAS catalog. The data presented in the above table can be downloaded at http://www.lptl.jussieu.fr/user/lesne/SNP_list2.txt

EFO term	Cancer type	da-SNPs in TAD borders	Chromosome	Genomic coordinates (hg19)
EFO_0000095	Chronic lymphocytic leukemia	rs41271473	chr1	228880296
		rs7558911	chr2	202023949
		rs9880772	chr3	27777779
		rs10936599	chr3	169492101
		rs31490	chr5	1344458
		rs4869818	chr6	154471225
		rs2236256	chr6	154478440
		rs11636802	chr15	56775597
		rs142215530	chr15	56777691
		rs72742684	chr15	56780767
		rs874460	chr19	47176752
		rs11083846	chr19	47207654
EFO_0000178	Gastric cancer	rs4072037	chr1	155162067
		rs1108143	chr2	235465858
EFO_0000182	Hepatocellular carcinoma in hepatitis B infection	rs2596542	chr6	31366595
EFO_0000305	Breast cancer	rs11903787	chr2	121088182
		rs653465	chr3	27343644
		rs6788895	chr3	150467808
		rs7716600	chr5	44875005
		rs2229882	chr5	56168712
		rs16886448	chr5	56170813
		rs3822625	chr5	56178111
		rs12655019	chr5	56195790
		rs10474352	chr5	90732225
		rs2180341	chr6	127600630
		rs11814448	chr10	22315843
		rs1926657	chr13	95874956
		rs2236007	chr14	37132769
		rs6504950	chr17	53056471
		rs8170	chr19	17389704
		rs8100241	chr19	17392894
		rs56069439	chr19	17393925
		rs4808801	chr19	18571141
		rs10411161	chr19	52372976
EFO_0000571	Lung adenocarcinoma	rs31489	chr5	1342714
EFO_0000756	Melanoma	rs3219090	chr1	226564691
		rs13097028	chr3	169464942
		rs401681	chr5	1322087
		rs1636744	chr7	16984280
		rs201131773	chr9	21805205
		rs7023329	chr9	21816528
		rs258322	chr16	89755903
EFO_0001071	Lung cancer	rs4254535	chr2	69198388
		rs10197940	chr2	152253918
		rs4975616	chr5	1315660
		rs402710	chr5	1320722
		rs401681	chr5	1322087
		rs4589502	chr15	67155069
		rs13181	chr19	45854919
EFO_0001075	Ovarian cancer	rs11782652	chr8	82653644
		rs183211	chr17	44788310
		rs8170	chr19	17389704
		rs2363956	chr19	17394124
EFO_0005088	Testicular germ cells cancer	rs3790672	chr1	165873392
		rs10510452	chr3	16625048
		rs17021463	chr4	95224812
		rs2720460	chr4	104054686
		rs4635969	chr5	1308552
		rs7040024	chr9	845516
EFO_0005570	Oral cavity cancer	rs10462706	chr5	1343794
		rs928674	chr9	133952024
		rs2398180	chr15	96863169
EFO_0005842	Colorectal cancer	rs10936599	chr3	169492101
		rs1370916	chr7	47238707
		rs2128382	chr8	130820039
		rs174537	chr11	61552680
		rs10849432	chr12	6385727
		rs11169552	chr12	51155663
		rs2286313	chr14	71514163
		rs1800469	chr19	41860296
EFO_1000650	Breast cancer (estrogen-receptor negative)	rs11903787	chr2	121088182
		rs8170	chr19	17389704
		rs56069439	chr19	17393925

It has recently been shown that the very existence of TADs (but not the compartmentalization into active and inactive domains) depends on the CTCF protein [69], which suggests analyzing the relationship between cancer-associated SNPs and CTCF binding sites. The link with cohesin binding sites is also a promising direction [31, 70, 71]. It is however to note that CTCF binding sites determined by sequence analysis have not the same strength (occupancy rate), and that only part of them are involved as insulators at the TAD boundaries [56] so that the analysis cannot be done in a straightforward way.

CONCLUSION: CANDIDATE LOCI FOR EXPERIMENTAL STUDIES

The present approach proposes to extend GWAS results interpretation by including in the analyses some aspects of the 3D genome architecture. We here focused on the large-scale sub-megabase organization of the genome into topological domains, with the working hypothesis that modifications in their borders may mediate a functional role of some disease-associated single-nucleotide polymorphisms.

Our main result is a list of diseases, including some cancer types, and a list of cancer-associated SNPs whose variant forms could contribute to a dysregulation of gene expression at a TAD border. This result delineates candidate loci for experimental studies, for instance 3C-qPCR experiments measuring contact frequencies in a quantitative way and thus able to assess a displacement or a disruption of a TAD border [41]. Genome-wide Hi-C investigation can also been envisioned, as done in a study of breast cancer genetic risk focusing on the long-range interactions between putative regulatory elements, harboring cancer-associated SNPs, and distal target genes [72]. Chromatin immuno-precipitation (ChIP-qPCR) can then be exploited to find which protein binding, if any, is affected at these SNPs and elucidate the basic step of their functional role and diseases association.

ACKNOWLEDGMENTS

This work has been supported by the “Mission for Interdisciplinarity” of the French National Center for Scientific Research (CNRS), program InFIniTI 2017, project 3D-SNPs, Grant 232647 (to A.L.) and by the “Agence Nationale de la Recherche”, project CHRODYT, Grant ANR-16-CE15-0018-04 (to T.F.). The present paper is based on a talk given by AL during the international conference “Physics of Cancer: Interdisciplinary Problems and Clinical Applications”, Tomsk, May 23–26, 2017. AL thanks the organizers of this conference and ISPSM SB RAS (Tomsk) for hospitality. MTH thanks LPTMC (Paris) for hospitality and CNRS funding of his stay. MTH also acknowledges financial support from the German Ministery for Education and Research (sysINFLAME project within the e:med program, grant 01ZX1306D).

The study reported here was conducted according to accepted ethical guidelines involving research in humans and/or animals and was approved by an appropriate institution or national research organization. The study is compliant with the ethical standards as currently outlined in the Declaration of Helsinki. All individual participants discussed in this study, or for whom any identifying information or image has been presented, have freely given their informed written consent for such information and/or image to be included in the published article.

REFERENCES

1. H. C. Erichsen and S. J. Chanock, British J. Cancer **90**, 747–751 (2004).
2. J. Hardy and A. Singleton, New Engl. J. Med. **360**, 1759–1768 (2009).
3. D. Altshuler, M. J. Daly, and E. S. Lander, Science **322**, 881–888 (2008).
4. D. J. Hunter, New Engl. J. Med. **360**, 1701 (2009).
5. M. L. Freedman, A. N. Monteiro, S. A Gayther, ... and M. James, Nat. Genetics **43**, 513–518 (2011).
6. A. Pombo and N. Dillon, Nat. Rev. Mol. Cell Biol. **16**, 245–257 (2015).
7. P. H. L. Krijger and W. de Laat, Nat. Rev. Mol. Cell Biol. **17**, 771–782 (2016).
8. J. Z. Li, D. M. Absher, H. Tang, ... and R. M. Myers, Science **319**, 1100–1104 (2008).
9. F. W. Albert and L. Kruglyak, Nat. Rev. Genetics **16**, 197–212 (2015).
10. J. P. Ioannidis, G. Thomas, and M. J. Daly, Nat. Rev. Genetics **10**, 318–329 (2009).
11. G. Fehringer, P. Kraft, P. D. Pharoah, ... and R. S. Houlston, Cancer Res. **76**, 5103–5114.
12. L. J. Engle, C. L. Simpson and J. E. Landers, Oncogene **25**, 1594–1601 (2006).
13. D. E. Reich, M. Cargill, S. Bolk, ... and E. S. Lander, Nature **411**, 199–204 (2001).

14. S. B. Gabriel, S. F. Schaffner, H. Nguyen, ... and S. N. Liu-Cordero, Science **296**, 2225–2229 (2002).
15. *1000 Genome Project Consortium*, Nature **491**, 56–65 (2012).
16. *1000 Genome Project Consortium*, Nature **526**, 68–74 (2015).
17. D. Welter, J. MacArthur, J. Morales, ... and H. Parkinson, Nucl. Acids Res. **42**, D1001–D1006 (2014).
18. J. F. Brookfield, BMC Biology **8**, 41 (2010).
19. A. C. J. Janssens and C. M. van Duijn, Hum. Mol. Genetics **17**, R166–R173 (2008).
20. C. Q. Chang, A. Yesupriya, J. L. Rowell, ... and S. D. Schully, Eur. J. Hum. Genet. **22**, 402–408 (2014).
21. J. C. Knight, Clinical Science **104**, 493–501 (2003).
22. S. Mooney, Briefings Bioinformatics **6**, 4456 (2005).
23. C. Chelala, A. Khan, and N. R. Lemoine, Bioinformatics **25**, 655–661 (2008).
24. P. H. Lee and H. Shatkay, Bioinformatics **25**, 1048–1055 (2009).
25. M. Cargill, D. Altshuler, J. Ireland, ... and E. S. Lander, Nat. Genetics **22**, 231–238 (1999).
26. V. Kumar, H. J. Westra, J. Karjalainen, ... and C. Wijmenga, PLoS Genet. **9**, e1003201 (2013).
27. A. A. Gheyas, C. Boschiero, L. Eory, ... and D. W. Burt, DNA Res., dsv005 (2015).
28. Y. G. Tak and P. J. Farnham, Epigenetics Chromatin **8**, 57 (2015).
29. M. T. Maurano, R. Humbert, R. Thurman, ... and J. Stamatoyannopoulos, Science **337**, 1190–1195 (2012).
30. M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder, Genome Res. **22**, 1748–1759 (2012).
31. X. Ji, D. B. Dadon, B. E. Powell, ... and R. A. Young, Cell Stem Cell **18**, 262–275 (2016).
32. L. Prokunina and M. E. Alarcón-Riquelme, Expert Rev. Molecular Med. **6**, 1–15 (2004).
33. Q. Li, J. H. Seo, B. Stranger, ... and M. L. Freedman, Cell **152**, 633–641 (2013).
34. J. R. Dixon, S. Selvaraj, F. Yue, ... and B. Ren, Nature **485**, 376–380 (2012).
35. E. P. Nora, B. R. Lajoie, E. G. Schulz, ... and E. Heard, Nature **485**, 381–385 (2012).
36. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, ... and J. Dekker, Science **326**, 289–293 (2009).
37. E. P. Nora, J. Dekker, and E. Heard, Bioessays **35**, 818–828 (2013).
38. F. Court, J. Miro, C. Braem, ... and T. Forné, Genome Biology **12**, R42 (2011).
39. V. Ea, T. Sexton, T. Gostan, ... and T. Forné, BMC Genomics **6**, 607 (2015).
40. S. S. Rao, M. H. Huntley, N. C. Durand, ... and E. Lieberman Aiden, Cell **159**, 1665–1680 (2014).
41. V. R. Ea, M. O. Baudement, A. Lesne, and T. Forné, Genes **6**, 734–750 (2015).
42. K. S. Sandhu, G. Li, H. M. Poh, ... and Y. Ruan, Cell Rep. **2**, 1207–1219 (2012).
43. M. J. Zeitz, F. Ay, J. D. Heidmann, ... and A. R. Hoffman, PLoS One **8**, e73974 (2013).
44. A. R. Barutcu, B. R. Lajoie, R. P. McCord, ... and J. Dekker, Genome Biology **16**, 1–14 (2015).
45. R. Stadhouders, S. Aktuna, S. Thongjuea, ... and E. Soler, J. Clin. Invest. **124**, 1699–1710 (2014).
46. R. A. Beagrie, A. Scialdone, M. Schueler , ... and J. Fraser, Nature **543**, 519–524 (2017).
47. T. Misteli, Cold Spring Harbor Perspectives Biol. **2**, a000794 (2010).
48. O. B. Naimark, A. S. Nikitiuk, M. O. Baudement, T. Forné, and A. Lesne, AIP Conf. Proc. **1760**, 020051 (2016).
49. A. L. Valton and J. Dekker, Curr. Op. Gen. Dev. **36**, 34–40 (2016).
50. D. Hnisz, A. S. Weintraub, D. S. Day, ... and R. A. Young, Science **351**, 1454–1458 (2016).
51. W. A. Flavahan, Y. Drier, B. B. Liau, ... and B. E. Bernstein, Nature **259**, 110–114 (2016).
52. S. M. Waszak, O. Delaneau, A. R. Gschwind, ... and E. T. Dermitzakis, Cell **162**, 1039–1050 (2015).
53. A. Lesne, N. Foray, G. Cathala, T. Forné, and J. M. Victor, J. Phys. Cond. Mat. **27**, 064114 (2015).
54. N. Matharu and N. Ahituv, PLoS Genetics **11**, e1005640 (2015).
55. C. T. Ong and V. G. Corces, Nature Rev. Genetics **15**, 234–246 (2014).
56. M. Liu, M. T. Maurano, H. Wang, ... and G. Stamatoyannopoulos, Nat. Biotech. **33**, 198–203 (2015).
57. C. T. Ong and V. G. Corces, Nature Rev. Genetics **12**, 283–293 (2011).
58. N. Ahmadiyeh, M. M. Pomerantz, C. Grisanzio, ... and M. L. Freedman, Proc. Nat. Acad. Sci. USA **107**, 9742–9746 (2010).
59. D. J. Verlaan, S. Berlivet, G. Hunninghake, ... and A. Naumova, Am. J. Hum. Genet. **85**, 377–393 (2009).
60. S. Berlivet, S. Moussette, M. Ouimet, ... and A. K. Naumova, Hum. Genet. **131**, 1161–1171 (2012).
61. F. Grubert, J. B. Zaugg, M. Kasowski, ... and M. Snyder, Cell **162**, 1051–1065 (2015).
62. N. Lonfat and D. Duboule, FEBS Lett. **589**, 2869–2876 (2015).

63. D. G. Lupiáñez, K. Kraft, V. Heinrich, ... and S. Mundlos, Cell **161**, 1012–1025 (2015).
64. S. Scherer, *Guide to the Human Genome* (Cold Spring Harbor Laboratory Press, New York, 2010).
65. Y. Benjamini and Y. Hochberg, J. Roy. Stat. Soc. B **57**, 289–300 (1995).
66. A. Reiner, D. Yekutieli, and Y. Benjamini, Bioinformatics **19**, 368–375 (2003).
67. A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci, BMC Genomics **13**, 436 (2012).
68. M. R. Mansour, B. J. Abraham, L. Anders, ... and A. T. Look, Science **346**, 1373–1377 (2014).
69. E. P. Nora, A. Goloborodko, A. L. Valton, ... and B. G. Bruneau, Cell **169**, 930–944 (2017).
70. M. Merkenschlager and E. P. Nora, Annual Rev. Genomics Human Genetics **17**, 17–43 (2016).
71. W. Schwarzer, N. Abdennur, A. Goloborodko, ... and F. Spitz, bioRxiv, 094185 (2016).
72. N. H. Dryden, L. R. Broome, F. Dudbridge, ... and O. Fletcher, Genome Res. **24**, 1854–1868 (2014).