Review

Machine learning approaches infer vitamin D signaling: Critical impact of vitamin D receptor binding within topologically associated domains

Carsten Carlberg, Antonio Neme
School of Medicine, Institute of Biomedicine, University of Eastern Finland, FIN-70211 Kuopio, Finland

ARTICLE INFO

Keywords:
Vitamin D
VDR
Vitamin D target genes
TADs
Machine learning
SOM
Chromatin
Epigenome

ABSTRACT

The vitamin D-modulated transcriptome of highly responsive human cells, such as THP-1 monocytes, comprises more than 500 genes, half of which are primary targets. Recently, we proposed a chromatin model of vitamin D signaling demonstrating that nearly all vitamin D target genes are located within vitamin D-modulated topologically associated domains (TADs). This model is based on genome-wide binding patterns of the vitamin D receptor (VDR), the pioneer transcription factor PU.1, the chromatin organizer CTCF and histone markers of active promoter regions (H3K4me3) and active chromatin (H3K27ac). In addition, time-dependent data on accessible chromatin and mRNA expression are implemented. For the interrogation and in deep inspection of these high-dimensional datasets unsupervised and supervised machine learning algorithms were applied. Unsupervised methods, such as the vector quantization tool K-means and the dimensionality reduction algorithm self-organizing map, generated descriptions of how attributes, such as VDR binding and chromatin accessibility, affect each other as a function of time and/or co-localization within the same genomic region. Supervised algorithms, such as random forests, allowed the data to be classified into pre-existing categories like persistent (i.e. constant) and time-dependent (i.e. transient) VDR binding sites. The relative amounts of these VDR categories in TADs showed to be the main discriminator for sorting the latter into five classes carrying vitamin D target genes involved in distinct biological processes. In conclusion, via the application of machine learning methods we identified the spatio-temporal VDR binding pattern in TADs as the most critical attribute for specific regulation of vitamin D target genes and the segregation of vitamin D’s physiologic function.

1. Introduction

The transcription factor VDR is the key protein in intracellular vitamin D signaling, since it exclusively binds with high affinity the biologically active form of vitamin D₃, 1α,25-dihydroxyvitamin D₃ (1,25(OH)₂D₃) [1]. VDR is a member of the nuclear receptor superfamily characterized by a structurally conserved ligand-binding domain [2]. Within VDR’s ligand-binding domain 40 mostly non-polar amino acids form a pocket that fixes 1,25(OH)₂D₃ with high specificity [3]. Thus, all regulatory and modulatory effects of vitamin D within the cell nucleus, the so-called genomic actions of vitamin D, are mediated by VDR.

The VDR gene is rather ubiquitously expressed, since it is found in more than half of the 400 tissues and cell types forming the human body (www.proteinatlas.org/ENSG0000011424-VDR/tissue). Thus, the physiologic role of vitamin D comprises more than being the key regulator of calcium homeostasis and bone remodeling [4]. The most important extra-skeletal function of vitamin D is the modulation of the responsiveness of the innate and adaptive immune system [5]. Accordingly, vitamin D supports the immune system in its fight against infections and in parallel can prevent autoimmune disorders.

This review summarizes transcriptome- and epigenome-wide data, which had been obtained in THP-1 human monocytes. We outline how machine learning algorithms were instrumental in interpreting and categorizing these large-scale data. The results of these analyses were the basis of the chromatin model of vitamin D signaling [6] describing how 1,25(OH)₂D₃, via activating the VDR, induces epigenome-wide changes that modulate the transcriptome of the cells. The core of the model are chromatin loops, referred to as TADs, that are anchored by
1,25(OH)2D3-regulated binding sites of the transcription factor CCCTC-binding factor (CTCF) [7]. In THP-1 cells there are 425 TADs containing at least one VDR binding site and one of the 311 primary and 276 secondary vitamin D target genes [8,9]. We will conclude that the spatio-temporal VDR binding pattern within these TADs is the key attribute for specific regulation of vitamin D target genes and differentiating their physiologic function.

2. The VDR cistrome

The genome-wide binding pattern of a transcription factor is referred to as its cistrome. Traditionally, cistromes are assessed by applying the method chromatin immunoprecipitation sequencing (ChIP-seq), which uses an antibody highly specific for the investigated transcription factor [10]. Depending on the type of transcription factor, the respective cistromes typically comprise 1000–100,000 genome-wide binding sites, so-called “peaks” [11]. Moreover, for a given transcription factor there are tissue- and cell type-specific differences in the expression of its cistrome. In addition, the cistrome of inducible transcription factors, such as VDR, dynamically depends on the presence of their respective stimuli, such as 1,25(OH)2D3.

VDR ChIP-seq has been performed in human cellular systems, such as B lymphocytes (GM10855 and GM10861) [12], undifferentiated monococytes (THP-1) [9,13], colorectal cancer cells (LS180) [14], hepatic stellate cells (LX2) [15] and macrophages (lipopolysaccharide-polarized THP-1) [16]. Based on these data, the cistrome of ligand-stimulated VDR is formed by some 2000–10,000 sites per cell type [17]. Interestingly, a small subset of only 50 sites was found in all investigated human cell systems [16]. This means that the VDR cistrome is quite cell-specific providing an explanation why most VDR expressing tissues and cell types have a rather different set of vitamin D target genes [17,18]. However, common to all cell types is that 1,25(OH)2D3 treatment significantly increases the number of genomic VDR binding events by a factor of 2–10 [16].

Equal numbers of VDR binding sites are found both up-stream and down-stream of the transcription start sites (TSSs) of primary vitamin D target genes [13]. This Gaussian-type binding pattern is in accordance with findings of the Encyclopedia of DNA elements (ENCODE) project (www.encodeproject.org) [11]. The fact that the number of VDR binding sites largely exceeds the count for primary vitamin D target genes suggests that not all members of the VDR cistrome are equally important for genomic vitamin D signaling. A VDR ChIP-seq time course in THP-1 cells [9] demonstrated that out of a total number of 11,657 VDR binding sites only 510 persistent sites were present at all conditions, 2109 loci were transiently occupied and 9038 sites were detected only after 24 h stimulation with 1,25(OH)2D3 (Table 1). The machine learning approach self-organizing map (SOM, details follow below) was able to reduce the number of persistent VDR sites to 339 highly conserved loci [9]. In numbers this fits far better to the 311 primary vitamin D target genes in THP-1 cells [19], more than half of which are controlled by conserved persistent VDR sites [9]. The latter sites are equally distributed over the genome, i.e. they seem to be strategically positioned, in order to provide the whole human genome with vitamin D sensitivity. Similarly, primary vitamin D target genes are also equally distributed over the human genome [8]. Thus, a small subset of the VDR cistrome may be sufficient for regulating most primary vitamin D target genes.

3. Epigenome-wide effects of vitamin D

Chromatin is the complex of genomic DNA with nuclear proteins, such as nucleosome-forming histones [20]. The epigenome is the genome-wide information stored in form of covalent and structural modifications of chromatin. Epigenetic changes do not involve any alternation in the sequence of genomic DNA, although some of them are heritable. There are at least three levels of epigenome/chromatin organization, which are i) cytosine methylation, in particular at the sites of CpG islands, ii) more than 100 different types of post-translational modifications of histone proteins, such as lysine trimethylation or lysine acetylation, and iii) 3-dimensional organization of the nucleus, such as chromatin loops [21]. More than 90% of the genomic DNA of a differentiated cell is not accessible to transcription factors and other nuclear proteins, because chromatin largely acts as an intrinsic repressor of gene expression [22,23]. Accordingly, the “epigenetic landscape” of a cell type is restricted to some 100–200,000 loci, majorly representing TSS and enhancer regions being accessible to transcription factors and RNA polymerases [11].

In contrast to the genome, which is supposed to stay constant over the lifetime of a person, the epigenome constantly responds in a dynamic fashion to many extra- and intracellular signals, such as vitamin D stimulation [6]. For example, in THP-1 cells vitamin D-induced genome-wide changes of chromatin accessibility have been measured via the method formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq) [19]. From the 62,000 accessible chromatin regions in this cellular system in total nearly 9000 were significantly (p < 0.05) affected by vitamin D stimulation: after 2 h already 3300 loci changed their accessibility, after 24 h even more than 4500 sites responded, while after 48 h only some 2400 regions were still affected by 1,25(OH)2D3 (Table 1). Similar observations were made on the level of histone modifications [24]: 550 of 22,998 genomic regions with H3K4me3 modifications (marks of active TSSs) responded significantly (p < 0.05) to a 24 h stimulation with 1,25(OH)2D3 and even 2473 of 45,578 regions with H3K27ac modifications (marks of active chromatin) were significantly affected (Table 1). This means that, at least in human monocytes, the epigenome responds at more than 500 promoter regions and some 2500 enhancer regions to a stimulation with 1,25(OH)2D3.

Most of the vitamin D-induced events within the epigenome seem to

---

**Table 1**

**Large-scale datasets.** Datasets of genome-wide assays performed under identical conditions in THP-1 cells are listed. The relation of these datasets is schematically depicted in Fig. 1.

<table>
<thead>
<tr>
<th>GEO dataset</th>
<th>Assay type</th>
<th>Stimulation time</th>
<th>Total number</th>
<th>Vitamin D sensitive</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE 69284</td>
<td>RNA-seq</td>
<td>2.5, 4 &amp; 24 h 1,25(OH)2D3</td>
<td>14,402 expressed genes</td>
<td>587 genes: 311 primary &amp; 276 secondary</td>
<td>[8,9,19]</td>
</tr>
<tr>
<td>GSE 69297</td>
<td>FAIRE-seq</td>
<td>2, 24 &amp; 48 h 1,25(OH)2D3</td>
<td>62,231 sites of open chromatin</td>
<td>2h: 3323 sites; 24 h: 4586 sites; 48 h: 2399 sites</td>
<td>[19]</td>
</tr>
<tr>
<td>GSE 69962</td>
<td>ChIP-seq</td>
<td>24 h 1,25(OH)2D3</td>
<td>40,078 sites</td>
<td>2130 sites defining 587 TADs</td>
<td>[7]</td>
</tr>
<tr>
<td>GSE 89178</td>
<td>H3K4me3 ChIP-seq</td>
<td>24 h 1,25(OH)2D3</td>
<td>122,319 sites</td>
<td>6,498 sites: 5747 up- and 751 down-regulated</td>
<td>[31]</td>
</tr>
<tr>
<td>GSE 89431</td>
<td>VDR ChIP-seq</td>
<td>2 &amp; 24 h 1,25(OH)2D3</td>
<td>21h 1903 sites</td>
<td>510 persistent, 2109 transient and 9038 &quot;24 only&quot; sites</td>
<td>[9]</td>
</tr>
<tr>
<td>GSE 107851</td>
<td>H3K27ac ChIP-seq</td>
<td>24 h 1,25(OH)2D3</td>
<td>22,998 regions</td>
<td>550 regions</td>
<td>[24]</td>
</tr>
<tr>
<td>GSE 107852</td>
<td>H3K4me3 ChIP-seq</td>
<td>24 h 1,25(OH)2D3</td>
<td>45,578 regions</td>
<td>2,473 regions</td>
<td>[24]</td>
</tr>
</tbody>
</table>

* All assays were performed in at least three biological repeats.

Statistically significant changes (p < 0.05).
be secondary effects, but at some 25% of the affected chromatin regions VDR binding was detected [24]. At the latter loci the response of the THP-1 cells (95% ChIP-seq overlap between both cellular systems [7]). Center VDR (red circle) binds accessible genomic DNA (gaps between light blue circles representing nucleosomes) and is supported by the pioneer factor PU.1 (green square). The genomic region that can be influenced by 1,25(OH)2D3 (via binding to VDR) is restricted by CTCF proteins (blue triangles) de

Fig. 1. Dataset integration for the chromatin model of vitamin D signaling. **Top** CTCF ChIA-PET data, which had been collected by the ENCODE project from K562 monocytes, were used to extrapolate the TAD borders in THP-1 cells (95% ChIP-seq overlap between both cellular systems [7]). **Center** VDR binding was detected [24]. At the latter loci the response of the THP-1 cells (95% ChIP-seq overlap between both cellular systems [7]). The genomic region that can be influenced by 1,25(OH)2D3 (via binding to VDR) is restricted by CTCF proteins (blue triangles) defining left and right TAD borders, i.e. only vitamin D target genes (red arrow) within the TAD will be stimulated to produce more mRNA copies. **Bottom** THP-1 datasets used for defining the chromatin model.

be secondary effects, but at some 25% of the affected chromatin regions VDR binding was detected [24]. At the latter loci the response of the epigenome is of primary nature. VDR can communicate with chromatin modifying proteins, such as the lysine demethylase KDM6B and the chromatin remodeler BRD7, via direct and indirect interaction, such as up- and down-regulating their genes [25] or being part of the same large protein complex in the nucleus [26,27].

The identification of another important interference of VDR with a nuclear protein was based on a screening for transcription factor binding motifs below VDR peaks using the HOMER algorithm [28]. In addition to traditional VDR binding motifs, top-ranking motifs were those of the pioneer transcription factor PU.1 [16]. A pioneer factor is a transcription factor with a large cistrome being related to some promiscuity in DNA binding and high diversity in protein-protein interactions [29]. In hematopoietic cells, PU.1 cooperates with VDR in the process of monocytes and granulocyte differentiation [30]. Accordingly, in THP-1 cells PU.1 ChIP-seq showed PU.1-VDR co-location at nearly 2/3 of all VDR binding sites [31] (Table 1 and Fig. 1). Thus, at many genomic loci VDR is supported by PU.1, in order to keep chromatin regions accessible. Moreover, at 5600 PU.1 sites (i.e. some 5% of all) 1,25(OH)2D3 significantly (p < 0.05) modulates the binding of PU.1, i.e. VDR also supports PU.1 binding [31].

4. Vitamin D-sensitive TADs

Like the genomes of other mammalians, also the human genome is organized into chromatin loops with an average size of some 700 kb, which are referred to as TADs [32,33]. The anchors of these loops function as insulator regions and are often conserved between tissues and even species [34,35]. Thus, TADs stabilize the architecture of the genome, so that enhancer-TSS contacts are restricted to the same chromatin loop. In this way, the epigenomic feature 3D-chromatin structure is linked to the control of gene expression [36]. The transcription factor CTCF binds to TAD anchors and is one of the key proteins organizing the 3D-chromatin structure [37]. In general, the CTCF cistrome comprises some 20,000 ubiquitous binding sites [38], but only a minority of them are found at TAD anchors. In K562 human monocytes, the method chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) was applied for CTCF [11] and created a database of chromatin loop-mediating CTCF loci.

In THP-1 cells, the CTCF cistrome is formed by 23,658 binding sites, 95% of which are identical to those in K562 cells [7]. Interestingly, 1321 of these CTCF sites are significantly (p < 0.05) affected (mostly up-regulated) by a 24 h stimulation with 1,25(OH)2D3 (Table 1). The comparison with K562 CTCF ChIA-PET data identified more than half of the vitamin D-sensitive CTCF sites in THP-1 cells as anchors of 587 TADs [7] (Fig. 1, top). Under the assumption that the human genome is segregated into some 3000 TADs [32], this suggests that nearly 20% of all TADs are modulated by stimulation with 1,25(OH)2D3. Thus, not only on the level of chromatin accessibility but also on the level of 3D-chromatin structure vitamin D has an effect on the human epigenome.

The expression of a primary vitamin D target gene is modulated, i.e. in most cases increased, when it co-locates with a prominent VDR binding site within the same TAD [9] (Fig. 1, center). An additional condition is that the TSS region of the target gene and an VDR-binding enhancer region are within accessible chromatin [39]. Interestingly, 425 of the vitamin D-sensitive chromatin loops contain at least one VDR binding site and one vitamin D target gene [7]. Vitamin D-dependent CTCF sites at TAD anchors insulate these prominent VDR binding sites from inappropriate activation of genes in a different TAD. This creates for most VDR sites and their target genes a chromosomal environment defining which target gene is controlled by which VDR loci in its vicinity. Accordingly, vitamin D-modulated TADs are the core of the chromatin model of vitamin D signaling (Fig. 1, center).

5. Principles of machine learning

Machine learning is a computational approach within the field of artificial intelligence [40]. The term “machine learning” is actually descriptive, since in this field algorithms learn from data, i.e. they are able to change internal parameters, in order to conduct a task in a more efficient way, for example, with a lower number of errors, once they are trained. Machine learning algorithms use inference, i.e. conclusions reached on the basis of evidence and reasoning, in order to recognize complex patterns in large-scale datasets. The latter are often obtained by the application of next-generation sequencing methods, such as RNA sequencing (RNA-seq), FAIRE-seq and ChIP-seq. In general, machine learning algorithms build a model from data inputs, then train it and finally turn it to a prediction-making device.

A key focus of machine learning is the identification of subtle or hidden structure of high-dimensional datasets, such as provided with examples in the next chapter. Importantly, for the learning of functional relationships from data the latter do not need to be functionally understood, i.e. they can be non-labeled. This means that machine learning methods can create predictive models without a detailed understanding of the underlying mechanisms [41].

In general terms, there are two types of learning observed in specialized algorithms, namely supervised and unsupervised approaches. Supervised methods are trained on well-defined, i.e. labeled datasets, such as a list of vitamin D target genes, and can then make predictions for non-labeled datasets. Discriminative classifier algorithms, such as random forests, incorporate prior knowledge on the investigated biological system [42]. Moreover, kernel algorithms, such as a K-means, use mathematical functions called kernels, which summarize full regions of the high-dimensional space, in order to capture the most relevant aspects of the distribution of the data in those spaces. In contrast, unsupervised methods do not rely in an external label or class, and their...
aim is to find relevant patterns in the data in an unbiased fashion. The SOM, an unsupervised machine learning method [43], can be used, when large-scale datasets, which are often non-labeled, need to be reduced in their dimensionality. An iterative process captures in a 2-dimensional lattice (often 3 × 3) the topology of the high-dimensional space, which is defined by a large number of tested attributes. SOMs have several advantages over more traditional dimensionality reduction methods, such as principal component (PC) analysis, since the former can capture high-order momentum of the data distribution, whereas the latter can only approximate first and second order momentum. Illustrative examples are presented in the next chapter.

Data mining and predictive machine learning algorithms integrate datasets from multiple sources [44]. The use of heterogenous data types, such as RNA-seq in combination with ChIP-seq, often allow more meaningful interpretations than the analysis of a single data type [45]. Many tasks for machine learning methods can also be solved using different statistical approaches indicating that there is a close relationship between statistics and machine learning [46]. However, machine learning mostly does not apply statistical standards, such as likelihood estimates, confidence estimations and power calculations [40].

6. Applications of machine learning

In genome biology, machine learning has been extensively applied to assign functional annotations, such as gene ontology terms, to genes and genomic regions. Input for such predictive machine learning algorithms are often next generation sequencing techniques, such as RNA-seq for gene expression profiling, FAIRE-seq for chromatin accessibility and ChIP-seq for patterns of histone modifications or transcription factor binding (Fig. 1, bottom). Outputs of these approaches are models that describe mechanistic relationships between data, for example, in gene expression. Such network models can be trained by additional gene expression and epigenomic data from the same or related cellular systems [47]. In the vitamin D field machine learning algorithms were applied to reduce the dimensionality of the heterogenous large-scale data obtained in 1,25(OH)2D3-stimulated THP-1 cells (Table 1) and to create the chromatin model of vitamin D signaling (Fig. 1, center).

As a first example, we discuss the dimensionality reduction of RNA-seq data by the application of SOM algorithm (Fig. 2). For a RNA-seq experiment in three biological repeats THP-1 cells had been treated for 2.5, 4 and 24 h with 1,25(OH)2D3 and vehicle as a reference (Table 1). Initially, the sequencing reads were aligned using Bowtie and Cuffdiff statistics for differential expression and indicated the rather large number of 1284 vitamin D target genes [19]. The re-analysis of the raw data by STAR alignment [48] and DESeq2 statistics [49] (assuming negative binomial distribution of the data) even detected 3650 out of 14,402 expressed genes to be significantly (p < 0.05) modulated by 1,25(OH)2D3 [9] (Table S1). In order to reduce the dimensionality of the basal state of these 3650 genes and their change in expression at three time points (i.e., representing four dimensions), the unsupervised machine learning approach SOM was applied with a 3 × 3 lattice. This sorted the rather heterogenous group of all vitamin D target genes into nine more homogenous classes containing 11–1390 genes with similar expression changes (Fig. 2, Table S1). In parallel, the SOM also indicated via the distant classes 1 and 9 the most different expression profiles of 11 strongly down-regulated genes and 19 remarkably up-regulated genes, respectively. In addition, there are 49 moderately down-regulated genes in class 4 as well as 109 and 401 highly and moderately up-regulated in classes 8 and 6, respectively (Table S1). In order to reduce the dimensionality of the basal state of these 3650 genes and their change in expression at three time points (i.e., representing four dimensions), the unsupervised machine learning approach SOM was applied with a 3 × 3 lattice. This sorted the rather heterogenous group of all vitamin D target genes into nine more homogenous classes containing 11–1390 genes with similar expression changes (Fig. 2, Table S1). In parallel, the SOM also indicated via the distant classes 1 and 9 the most different expression profiles of 11 strongly down-regulated genes and 19 remarkably up-regulated genes, respectively. In addition, there are 49 moderately down-regulated genes in class 4 as well as 109 and 401 highly and moderately up-regulated in classes 8 and 6, respectively (Table S1). These in total 587 genes were considered as reasonable targets of vitamin D, while the remaining 3063 genes in classes 2, 3, 5 and 7 were sorted out, i.e., the application of SOM resulted in an 3.7-fold
enrichment of prominent vitamin D target genes.

Following the same principles, SOM (mostly with a $3 \times 3$ lattice) had been used to filter the 339 most conserved persistent VDR sites out of a pool of 510 loci [9] (already mentioned above) or categorizing 2130 vitamin D-sensitive CTCF sites by attributes, such as overlap with TSS regions, VDR or accessible chromatin, the presence of binding motifs and the use as chromatin loop anchor [7]. The advantage of SOM is that it handles a multi-dimensional dataset as an entity and does not rely on threshold setting for individual attributes. In this way, heterogeneous data can simultaneously be integrated and categorized.

A second example for the application of machine learning in the vitamin D field is a categorization of the 587 vitamin D-sensitive TADs by their epigenomic properties [9] (Fig. 3). Vitamin D target genes have a heterogeneous epigenomic environment within their respective TADs, which is based on several attributes, such as the number, strength and inducibility of neighboring binding sites for VDR, PU.1 and CTCF as well as 1,25(OH)$_2$D$_3$-sensitive histone markers and chromatin accessibility (Fig. 3A). First, the unsupervised machine learning algorithm K-means was used for the segregation of all TADs into different classes, also referred to as tessels. K-means involves an iterative process, referred to as vector quantization, in which data is partitioned into $k$ different centroids. Each centroid defines the center of the corresponding class and its location is based on the distribution of the high-dimensional data. This so-called Voronoi tessellation creates a partitioning of the high-dimensional space (Fig. 3C). The average Euclidean distance between the center of a class and the data points within the class is referred to as quantization error. The lowest quantization error was obtained via a distinction into five TAD classes and outperformed K-means results with three, four, six and seven TAD classes. At the same time, for each Voronoi tessellation pathway and gene ontology enrichment was conducted over all vitamin D target genes contained in the TADs. For $k = 5$, the overlap of pathways/ontologies between the five categories was the lowest. The supervised machine learning approach random forest then identified the TAD-size weighted numbers of persistent and transient VDR sites as the most relevant attributes, i.e., the attributes that have the greatest contribution in differentiating the five Voronoi categories (Fig. 3E). By principal component analysis all 587 TADs were mapped from the high-dimensional space created by all tested attributes to a 2-dimensional graph (Fig. 3F).

The five TAD classes contain between 34 and 156 vitamin D target genes (see [9], Table S5). Gene ontology analysis indicated for each TAD class an enrichment of different biological processes, in which the respective vitamin D target genes are involved (Fig. 3F): TAD class 1 genes are enriched for immune function, class 2 genes for differentiation and adhesion, class 3 for membrane-based signal transduction pathways, class 4 genes for antigen-receptor signaling, RAS kinase signaling and lipid storage and class 5 genes for cellular localization and transport. Thus, the TAD classes allow the segregation of the
pleiotropic functional profile of vitamin D, in which the relative number of persistent and transient VDR sites in the local environment of the vitamin D target genes is the most distinctive attribute.

7. Summary: VDR cistrome in TAD context

The large VDR cistrome of more than 10,000 peaks contains a subgroup of a few hundred persistent VDR loci that seem to act as “hotspots” in vitamin D signaling. VDR binding to these sites changes significantly over time after ligand stimulation, but they are always occupied. In this way, persistent VDR loci are not only the primary contacts of the human genome with the nuclear hormone 1,25(OH)2D3, but also coordinate the functional consequences of vitamin D stimulation over time. Thus, persistent VDR sites reflect best the spatio-temporal response of the (epi)genome to the extracellular changes in vitamin D levels.

In addition, transient VDR sites, i.e. genomic loci that are not always occupied, modulate the response of the epigenome to vitamin D and support persistent VDR sites in the regulation of primary vitamin D target genes. In a given TAD the combination of persistent and transient VDR sites create a local genomic environment that affects histone modifications at enhancer and TSS regions as well as chromatin accessibility. Moreover, the local binding of PU.1 and other (pioneer) factors, such as GABPA [50], and CTCF acting as TAD anchor is affected. Machine learning algorithms that analyzed these heterogeneous data in a high-dimensional space indicated that the relative number of persistent and transient VDR sites is the most critical attribute for the functional distinction of vitamin D-modulated TADs (Fig. 3). This implies that all other vitamin D-triggered epigenomic changes in activity and accessibility, transcription factor binding and 3-dimensional chromatin organization within the TAD are secondary effects of the activity of persistent and transient VDR sites.

Each vitamin D-sensitive TAD represents an individual gene regulatory scenario, in which vitamin D-triggered epigenetic changes in VDR and pioneer factor binding as well as in chromatin accessibility result in a significant modulation of the expression of one or more primary vitamin D target genes located within the TAD. Machine learning approaches allowed the segregation all 857 TADs into only five classes being characterized by an average ratio of the number of persistent and transient VDR sites and an individual functional profile of the vitamin D target genes within these TAD classes. Thus, vitamin D target genes are clustered by the similitudes of the epigenetic landscape of their corresponding TADs. For example, TADs with vitamin D target genes that are important for immune function (class 1, Fig. 3F) are primarily regulated by transient VDR sites, i.e. sites that respond in an “on/off” modus. This regulatory mechanism seems to be more suited for a tight control of immune functions.

8. Conclusion

High-throughput and next generation sequencing technologies generate large amounts of data. In some cases, such as depicted here for vitamin D and its influence on the epigenome and transcriptome, the combination of the number of cases and attributes make the use of more traditional hypothesis testing challenging. As an alternative, the use of machine learning is better suited, since the assumptions over the data distribution are much more flexible. In this review we discussed first examples of a systematic application of machine learning algorithms in the vitamin D field. Although the here presented analyses were exclusively based on experiments performed in in vitro systems, machine learning approaches can also be applied on data obtained from in vivo observations. A first example was provided by our study on in vivo epigenome changes in white blood cells of an individual obtaining every month a vitamin D3 bolus [51].


