bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identification of genetic variants affecting vitamin D receptor binding and associations with autoimmune disease

Giuseppe Gallone[1,2,⋆], Wilfried Haerty[1,2], Giulio Disanto[2], Sreeram Ramagopalan[3,⋆,†], Chris P. Ponting[1,2,4,†] and Antonio J. Berlanga-Taylor[5,6,7⋆,†]

[1]MRC Functional Genomics Unit, University of Oxford, Oxford OX1 3PT

[2]Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, OX1 3PT

[3]Real-World Evidence, Evidera, London, W6 8DL

[4]MRC Human Genetics Unit, The Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU

[5]Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX3 7BN

[6]CGAT, MRC Functional Genomics Unit, University of Oxford, Department of Physiology, Anatomy and Genetics, Oxford, Oxford OX1 3PT

[7]*present address:* MRC-PHE Centre for Environment and Health, Department of Epidemiology & Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, St Mary's Campus, Norfolk Place, LONDON W2 1PG

[†]Authors contributed equally.

[⋆]Corresponding authors.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Background - Large numbers of statistically significant associations between sentinel SNPs and case-control status have been replicated by genome-wide association studies (GWAS). Nevertheless, currently few underlying molecular mechanisms of complex disease are known. Our approach has been to investigate whether variation in binding of a transcription factor, the vitamin D receptor (VDR) whose activating ligand vitamin D has often been proposed as a modifiable factor in multiple disorders, could explain any of these associations. VDR modifies gene expression by binding DNA as a heterodimer with the Retinoid X receptor (RXR).

Results - We identified 43,332 genetic variants significantly associated with altered VDR binding affinity (VDR-BVs) using a high-resolution (ChIP-exo) genome-wide analysis of 27 HapMap lymphoblastoid cell lines (LCLs). The VDR-BVs are over-represented in well conserved consensus RXR::VDR binding motifs. However, most fell outside of recognisable RXR::VDR motifs, implying that genetic variation often affects RXR::VDR binding affinity only indirectly. Finally, we compared 341 VDR-BVs replicating by position in multiple individuals against background sets of variants lying within VDR-binding regions that had been matched in allele frequency and were independent with respect to linkage disequilibrium. In this stringent test, these replicated VDR-BVs were significantly ($q < 0.1$) and substantially ($>$ 2-fold) enriched in GWAS intervals associated with autoimmune and other diseases, including inflammatory bowel disease, Crohn's disease and rheumatoid arthritis.

Conclusions Our findings are consistent with altered RXR::VDR binding contributing to immunity-related diseases. Replicated VDR-BVs associated with these disorders could represent causal disease risk alleles whose effect may be modifiable by vitamin D levels.

Keywords: Allele-specific binding; ChIP-exo; Vitamin D receptor; Autoimmune disease

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Genetic variants can alter transcription factor (TF) binding, chromatin status and gene expression [1, 2, 3] and can be explanatory of disease susceptibility variation. Rather than residing in protein-coding sequence, the majority (approximately 93%) of disease- or trait-associated variants lie in non-coding sequence [4]. Furthermore, non-coding variants in regulatory elements explain 8-fold more heritability of complex traits than protein-coding variants [5]. If we are to better understand the mechanistic bases to complex disease (and their interaction with environmental factors) then we will need to understand how sequence differences in TF binding sites (TFBSs) across multiple genotypes contribute to disease susceptibility. Variations in TF affinities will then need to be cross-referenced not just with the sentinel variant from genome-wide association studies (GWASs), which has a very low probability ($\sim$5%) of being causal and is on average 14 kb from the true causal variant [6], but also with all other variants with which it lies in strong linkage disequilibrium (LD). True causal variants will be enriched in *cis*-regulatory elements, especially in enhancers and, to a lesser extent, in gene promoters [6]. These elements are often active in only a few cell types [7] and at limited developmental time points [8], which provides an opportunity for relating genetic variation — via molecular observations (TF binding differences) — to cellular or organ-based aetiologies.

Understanding complex disease will also need the determination of how functional genetic variants interact with environmental factors. One such factor is Vitamin D, a class of fat-soluble secosteroids that enhance intestinal absorption of dietary minerals, which are synthesised in the skin upon exposure to sunlight. Many observational studies have yielded associations between low serum concentrations of 25-hydroxyvitamin D [25(OH)D] and the risks of developing diverse diseases [9, 10]. However, Mendelian randomisation studies, for example on susceptibility to multiple sclerosis [11] and all cause mortality [12] have indicated a causal role for this hormone. Identifying a molecular basis to these statistical associations would also indicate a direct causal role for vitamin D levels in these diseases. Vitamin D signalling occurs principally following the binding of calcitriol, the active form of vitamin D, to its cognate nuclear vitamin D receptor (VDR) which binds DNA in a heterodimer with the Retinoid X receptor (RXR [13]). We previously have shown that VDR binding in lymphoblastoid cell lines (LCLs) occurs at 2,776 locations and preferentially (>2-fold) within intervals genetically associated

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with diseases, such as Crohn's disease and Multiple Sclerosis [14]. On one hand, these enrichments could indicate a direct modifying effect of VDR-binding on disease risk. On the other hand, they could, more trivially, reflect a general association between regulatory regions and disease-associated genomic intervals. These enrichments of VDR binding within these intervals also are not informative of whether altered VDR binding to DNA functional elements explains genetic contributions to disease risk. If this is the case, we then also wish to determine whether it is a gain or loss of VDR binding that causally increases disease risk.

In this study, we inferred VDR-binding sites using ChIP-exo (chromatin immunoprecipitation combined with lambda exonuclease digestion followed by high-throughput sequencing [15, 16]) from each of 27 LCL samples using 3 complementary approaches. The first 'peak-calling' method models the variation in read density. The second and third approaches predict VDR-binding sites using genetic variation to explain differences in VDR-binding affinity, either by quantitative trait loci association testing ('QTL') across all samples or by allele-specific binding ('ASB') analysis of allelic imbalance based on read depth at sites with heterozygous single nucleotide variants. Sequence variants were identified that both alter VDR binding and have been statistically associated by GWAS with altered risk for particular diseases (or are in strong LD with such variants). These are excellent candidates for sequence variants that, through their alteration of VDR binding, directly alter disease susceptibility. We considered whether VDR-binding events have consequences on fitness, and thus when lost or gained either provide protection from or confer additional risk of disease. Compared against a background of all VDR-binding sites, we found that human variants associated with variable VDR-binding are enriched (by up to 2-fold) in genomic intervals previously associated with particular traits, including some autoimmune diseases.

Results

ChIP-exo yields finely resolved VDR binding peaks genome-wide

Calcitriol–stimulated lymphoblastoid cells for 30 HapMap samples were grown and prepared for ChIP essentially as described previously [14] with modifications for ChIP-exo [15, 16] (Methods; Supplementary Table 1). In total, 27 samples successfully passed quality checking (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Material). Unlike for data from the more traditional ChIP-seq approach, analysis of ChIP-exo data has yet to become standardised. We thus undertook an in-depth study using contrasting approaches to modelling the ChIP-exo data background, to handling duplicate reads, to performing cross-correlation analyses and to calling peaks; these approaches are discussed in detail in the Supplementary Material (see also Supplementary Figures 1 and 2, Supplementary Tables 2–9).

VDR peaks were both highly reproducible between samples, with 15,509 intervals containing peaks called in at least 3 samples (CP_{o3} peak set) (Figure 1A and B; Supplementary Figures 3 and 4), and highly concordant with our previously published VDR ChIP-seq data for calcitriol-stimulated LCLs (CP_{o3}: 76.3%; Figure 1C; Supplementary Table 11) [14] (Supplementary Material). For peaks identified in both experiments, binding intervals were considerably better resolved (typically > 5-fold) using ChIP-exo (Supplementary Figure 2A, pairwise Kruskal-Wallis Test, $p < 0.001$).

As expected, VDR binding peaks occurred preferentially at protein coding transcriptional start sites (TSS) (Figure 1D) and within regions of open chromatin (at LCL DNase 1 hypersensitivity sites, DHS), especially for those observed repeatedly (Supplementary Figures 5 and 6). VDR binding peaks were significantly concentrated within proposed functional elements, specifically regions of high regulatory factor binding (HOT regions [17] and clustered TF binding sites from ENCODE), active promoters, enhancers and insulators (Figure 1E; Supplementary Figure 7) and previously reported GWAS disease loci, principally for autoimmune disorders (Supplementary Figure 8). This latter observation could reflect a causal effect in which VDR-binding influences disease susceptibility and/or with gene loci being correlated non-causally with both disease susceptibility intervals and VDR binding sites.

VDR binding motifs occur preferentially in enhancers

De novo motif analyses of peaks within ChIP-exo VDR binding intervals revealed motifs strongly resembling those for the RXR::VDR DR3 heterodimer [18, 19] (Figure 2A; Supplementary Material, Supplementary Figure 9A and Supplementary Table 12). For example, 9,899 (63.8% of the CP_{o3} set) instances of this DR3 heterodimer motif occurred in binding sites (PScanChIP [20], score > 0.7). 874 instances of the monomeric VDR motif were also identified in these peak intervals (Supplementary Figures 9B; PScanChIP score = 1) which are known to have limited affinity for functional dimeric nuclear hormone receptor complexes [18]. The heterodimeric DR3 motif was strongly over-represented

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Supplementary Table 12) both locally (in the peak sequences, compared to neighbouring regions) and globally (in the peak regions, compared to LCL DHS sites [20]) and was significantly centrally enriched in the binding regions (Figure 2A).

VDR binding sites could be separated into functionally distinct classes based on the presence of RXR::VDR DR3-like motifs. *Class I* sites contain strong DR3-like motifs (top 20% of the PScanChIP score distribution, 3,094 regions, Supplementary Figure 10A) and tend to be located away from genes' transcriptional start sites (TSSs) and yet close to genes involved in immune response processes (Figure 2B,C). *Class I* VDR binding sites show greatest significance of proximity to genomic regions associated with cancer and autoimmune diseases (Figure 2C). We compare these to *Class II* sites which contain only weak or no DR3-like motifs (bottom 20% of PScanChip scores, 3,102 regions, Supplementary Figure 10A) (Figure 2B,D). Class *I* binding events occur preferentially in strong or weak enhancers defined by ENCODE, rather than in promoters (Supplementary Figure 10B and 10C).

Thousands of variants explain differential VDR binding

We next used the QTL and ASB methods to explain differences in VDR binding affinity (Supplementary Material). The QTL analysis used Bayesian regression modelling of variants underlying binding peaks [21, 22] (Supplementary Material, Supplementary Tables 13 and 14, Supplementary Figures 12 to 15). The ASB analysis of allelic imbalance is based on read depth at heterozygous single-nucleotide variants [23] and uses sample-specific maternal and paternal reference maps, which we corrected for unannotated copy number variation and from which we excluded variants in ENCODE blacklisted repeat regions [24]. The two methods' variants associated with differences in VDR binding (VDR-Binding Variants, VDR-BVs) were annotated and prioritised based on the guidelines proposed by [25] and related algorithms (Funseq2, [26]).

We report a total of 43,332 VDR-BVs. Of these, 6.6% - 10.5% (2,867 or 4,447) lie within (or 'hit') RXR::VDR consensus motifs (Pscanchip score > 0.7 or > 0.5, respectively). A similarly low fraction of binding affinity-associated variants present in motifs is observed for other TFs [27]. This could be explained, in part, by variants that alter DNA-binding affinity either for other subunits of a larger multi-molecular complex ('collaborative binding' [28]) or for factors that inhibit RXR::VDR binding.

Consequently, we next considered whether sequence variation in motifs for potential binding

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cofactors of RXR::VDR could explain the remaining $> 90\%$ of binding variation. We found that position weight matrices for 26 additional TFs (Supplementary Table 15) were significantly enriched ($p < 0.1$) (i) relative to both open chromatin genome-wide and chromosomally adjacent regions (< 150bp), and/or (ii) at the peak summits within VDR-BV intervals (Pscanchip score > 0.5 or 0.7; Supplementary Methods). Virtually all (84.1-88.3%) of binding variation could be explained by VDR-BVs lying within a RXR::VDR consensus motif, or within one or more of these 26 additional motifs (Pscanchip score > 0.7 or > 0.5). These are upper-bound values because of the inaccuracy of motif prediction and because not all variants lying in motifs will alter binding affinity.

Some of the 26 TFs (Supplementary Table 16) could bind RXR::VDR. Direct interaction between VDR and CREB1 has been reported previously [29] as have functional interactions between vitamin D/VDR and oestrogen metabolism and MYC proteins [30, 31]. Co-occurence of GABPA and ESRRB motifs with VDR-binding sites had been reported previously [32]. ZNF423 is not known to bind VDR yet it might do so indirectly because it physically associates with its heterodimer partner, RXRα [33].

Functional annotation of VDR-BVs

To begin to understand whether VDR-BVs could contribute to particular human diseases, we first considered their enrichment in particular genomic annotations. For this, we were interested in whether these variably-bound sites are enriched relative to a background of all VDR binding regions genome-wide so that we account for the non-uniform chromosomal distribution of binding sites. We chose for our null expectation all variants that were tested for differential binding within replicated (CP_{o3}) VDR binding peaks (Supplementary Material). Relative to this stringent background, VDR-BVs are 20%-40% enriched within enhancers but not in promoters (Figure 3A). These patterns of enrichment are reinforced when we stringently select only the most significant VDR-BVs (VDR-sBVs, Supplementary Figure 16A; AlleleSeq $\mathrm{FDR} \leq 0.01$; 6,715 VDR-BVs) or the 357 recurrent VDR-BVs that are called at the same position in multiple samples (VDR-rBVs, Supplementary Figure 16D). We conclude that VDR-BVs are not uniformly distributed among all VDR binding genomic intervals, and are especially frequent in enhancer regions manifested in LCLs.

VDR-BVs occurred 60% more frequently in RXR::VDR motifs within replicated (CP_{o3}) VDR binding peaks and 120% more frequently in strong (class I) RXR::VDR motifs in the same regions

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Figure 3B). Again, these enrichments strengthened substantially for VDR-sBVs or VDR-rBVs (Supplementary Figure 16B and 16E, respectively).

We then considered whether VDR-BVs are significantly enriched within LCL expression QTLs (eQTLs) relative to 1000 Genomes variants under VDR ChIP-exo pileups tested for VDR binding variation potential (Supplementary Material). In addition, these background variants were matched by derived allele frequency and LD-dependence confounders (Methods). We also took care to only consider variants that do not share strong linkage disequilibrium (32,183 VDR-BVs, 5,808 VDR-sBVs and 341 VDR-rBVs.)

VDR-BVs were enriched with dsQTLs (1.4-fold) and eQTLs (1.1-1.2-fold; Figure 3C) particularly for the more stringent VDR-BV subsets (Supplementary Figures 16C and 16F). The enrichment in eQTLs was confirmed using an independent eQTL resource (GEUVADIS LCL eQTLs [34]; Supplementary Material).

These eQTL enrichments are intriguing because the eQTLs were inferred from LCLs not treated with calcitriol, and implied that these enhancers' activities may not be entirely vitamin D-dependent in these cells. Nevertheless, there was no clear concordance or discordance between the direction of change of VDR binding at VDR-BVs (in calcitriol-treated LCLs) with the direction of change of gene expression in calcitriol-minus human LCLs reported for the GEUVADIS eQTL variants (data not shown). This, and the lack of enrichment of the DR3 motif in the VDR binding sites for the basal (unstimulated) state in LCLs [14] indicate that a calcitriol-activated LCLs' study at a similar scale to those employing unstimulated LCLs [34] will be required to fully resolve this issue.

Complex trait-associated variants influence VDR binding

To investigate whether susceptibility to specific diseases could be influenced by VDR-BVs, we considered their locations relative to GWAS variants significantly associated with 472 diverse diseases or traits from 688 studies in the largest catalogue available (GRASP v2.0 [35]) with at least 5 associated genome-wide significant SNPs ($p < 5 \times 10^{-8}$) (Supplementary Material). Again, we employed stringent procedures, matching variants for allele frequency and LD-dependence and discarding VDR-BVs in the MHC region as well as again comparing against a background of all VDR-binding regions (Supplementary Material).

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

VDR-BVs were significantly and up to 2-fold enriched within LD intervals associated with 17 diseases or traits (Figure 4). It is notable that these are not drawn randomly from all 472 traits considered, but are mostly immune- or inflammatory-related disorders, such as sarcoidosis, Graves' disease, Crohn's disease, irritable bowel syndrome and type I diabetes (Figure 4A). For the 341 VDR-rBVs, intervals from six disorders contained more VDR-rBVs than expected from sets of DAF-matched LD-accounted regions that bind VDR (Figure 4B). We emphasise that this represents a highly stringent analysis that accounts for all known confounding effects, namely potential biases from called VDR-binding sites, population stratification and physical linkage. Three of these disorders were autoimmune disorders (inflammatory bowel disease, Crohn's disease and rheumatoid arthritis) whilst a fourth, endometriosis, is frequently comorbid with autoimmune disorders [36]. Deficiency of serum 25(OH)D levels is associated with cardiovascular disease risk factors in adults [10], while the association between vitamin D levels and coronary artery disease, as for many disorders, is debated [37].

Functional impact of genetic variation on loss or gain of VDR binding

This evidence is consistent with altered RXR::VDR binding contributing to immunity-related diseases. Nevertheless, because variants in TFBS often fail to alter binding affinity and/or have no measurable effect on gene expression levels [38, 39], we needed to demonstrate that VDR-BVs are functional, specifically that they influence gene expression and have been negatively selected over evolution [40]. As expected if they are enriched in functional variants, we found that 50%-100% more VDR-BVs than random samples lie near (< 10kb) genes that are differentially expressed, in LCLs, upon addition of calcitriol [14] ($p < 10^{-4}$, Supplementary Methods).

Also, as expected if as a class they are functional, replicated VDR-binding peaks exhibit variable cross-species conservation across RXR::VDR motif positions (Supplementary Figure 17A and B). This signature of uneven selection across the DR3 motif is evident for class I VDR binding sites (Supplementary Figure 17C and D; Kruskal Wallis test, $p = 1.9 \times 10^{-11}$ and *post-hoc* pairwise Kruskal Wallis comparisons), but not in class II sites (Kruskal Wallis test, $p = 0.14$) consistent with the DR3 heterodimer-binding motif regulating transcription.

Next we assessed the impact on VDR binding affinity of VDR-BVs with respect to their location within the consensus DR3-type motif (JASPAR database, Figure 5), the transcriptionally active binding

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

site motif for the RXR::VDR heterodimer [41, 42, 43]. Mapping of allele-specific ChIP-exo reads to multiply replicated VDR binding regions (CP_{o3} peaks) was highly predictive of whether variants substantially alter ('break') functional motifs. For 89% of observations, the change in the strength of the VDR binding motif correctly predicts the direction of VDR binding affinity change (Figure 5A), a proportion that rises to 100% in class I VDR binding sites (Supplementary Figure 18A). Motif breaks do not favour 5' RXR over 3' VDR hexamers (Figure 5A, Wilcoxon rank sum test, $p = 0.71$), but especially impacts conserved G and C nucleotides (positions 2, 5, 11 and 14), and a T nucleotide (position 13). Notably fewer events occur at the near-essential T nucleotides at positions 4 and 12. Interestingly, in class I sites there are no VDR-BVs disrupting the T nucleotide at position 12 (Supplementary Figures 18A) which could reflect either low mutation rates (mutational 'cold spots') or strong purifying selection against deleterious nucleotide substitutions at these binding sites [44].

To distinguish these possibilities we next compared the population frequencies of alleles associated with either historical Loss Of Binding (LOB) or Gain of Binding (GOB) affinity to VDR (Methods). LOB VDR-BVs cause significantly stronger effects when within either 5' or 3' hexamer than within the DR3 spacer sequence (Wilcoxon rank sum test, $p < 0.01$); GOB variants: $p = 0.11$. VDR-BVs affecting one of the most highly conserved motif positions (position 13) always result in motif-breaking LOB effect (Figure 5C and D; Supplementary Figure 18C and D).

We next found that RXR::VDR motif sequences have been under constraint during recent evolution, within the extant human population. We did so by comparing the frequencies of derived alleles (Derived Allele Frequency [DAF] test) for LOB or GOB VDR-BVs in either Northern and Western European (CEU) or Yoruban (YRI) populations. In absence of demographic confounders, stronger purifying selection is implied by an excess of low population frequency variants in one set of variants over another [45].

We detected significant increases in DAF distribution for GOB over LOB VDR-BVs, both for CEU (Figure 6C, top-left panel) and YRI (Figure 6C, bottom-left panel) cohorts. Class I binding site VDR-BVs exhibited significant differences in GOB versus LOB DAF distributions (Supplementary Figure 19; $p < 0.05$ for both CEU and YRI). By contrast, variants in RXR::VDR hexamers not assigned as VDR-BVs showed no significant difference in GOB versus LOB DAF distributions (Figure 6C, top-right and bottom-right panels). Significant differences between GOB and LOB DAFs were attained

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

even when observing the RXR and VDR motif hexamers separately, and in both sub-populations ($p < 0.012$). Consequently, LOB variants are under substantially stronger purifying selection than are GOB variants. This is an important finding because it indicates that reduced binding of VDR at these genomic positions, and presumably reduced vitamin D levels, are commonly deleterious.

Discussion

We have demonstrated how genetic variation alters the binding affinity of the vitamin D receptor (VDR), a member of the nuclear receptor family of transcription factors. Genetic variants significantly associated with immune and inflammatory disease were found to disrupt VDR binding and to be located preferentially within enhancer regions, in line with a recent analysis of causal immune disease-related variants [6]. The study benefited from the ∼5-fold greater spatial resolution and signal-to-noise ratios provided by the ChIP-exo assay.

In its best studied model of allosteric modulation, the RXR::VDR heterodimer binds to DNA and enhances transcription via response elements that typically consist of two hexameric `((A/G)G(T/G)TCA)` half-sites, with RXR and VDR DNA binding domains occupying the upstream and downstream half-site, respectively [46]. This DR3 motif, which we previously identified using ChIP-seq data from two calcitriol-activated CEU LCLs [14], was also the predominant VDR binding interface in calcitriol-activated LCLs in the present ChIP-exo study. We identified strong enrichment of VDR binding within enhancers, with stronger enrichments associated with the more stringent subsets of VDR-BVs. Whilst VDR binding occurs preferentially within promoters, these interactions are only rarely associated with RXR::VDR DR3 motifs, and may be mediated by additional DNA-binding co-factors, including CTCF, and might reflect unproductive binding to open chromatin. In contrast, those VDR binding sites containing strong instances of the RXR::VDR DR3 motif were particularly enriched in enhancers, and also near to genes involved in immunity processes and related diseases. VDR thus is expected to exert substantial effect on the biology of the immune system as a sequence-specific enhancer regulator.

The VDR-bound RXR::VDR motif shows evidence of levels of sequence conservation across vertebrate evolution that mirror its information content. Furthermore, over modern human evolution loss of VDR-binding variants has occurred under stronger purifying selection than gain of binding

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variants. Most remarkably, we found variants weakening or strengthening the RXR::VDR PWM score to be highly predictive of the loss or gain of VDR binding affinity. Together these results imply that VDR-binding and thus presumably vitamin D levels, have in general been protective of disease.

Our study will have underestimated the true extent of regulatory variation at VDR binding sites. This is because the asymmetric binding analysis [23] by definition can only test variants which are heterozygous for a given LCL sample: therefore, true positive VDR-BVs will be unobserved if all LCL samples we considered are homozygous. Similarly, not all of the required three biallelic genotypes will have been present in these samples for the regression-on-genotype analysis [22, 21]. Consequently, a larger scale study would be expected to show improved power to detect VDR-BVs.

At least 90% of VDR-BVs lay outside of RXR::VDR DR3 motifs. The majority of variable binding could occur, as with other TFs, at weak or non-canonical binding motifs, or its affinity may alter because of genetic disruption to a cofactor binding site [47]. Another explanation is that VDR-BVs commonly exert their effect by altering the DNA conformation of regions flanking the core binding site [48]. Nevertheless, our *de novo* motif discovery identified several proteins, including CREB1, GABPA, ESRRB and ZNF423 as potential RXR::VDR cofactors, whose altered DNA-binding affinity could influence VDR binding.

Our most important results derived from a conservative analysis of a stringent set of replicated binding variants. From this, we report a significant excess of VDR-BVs coincident or in strong LD with genome-wide significant GWAS tag variants for 3 autoimmune and inflammatory diseases. These results associate a molecular phenomenon, the genetic disruption of nuclear receptor binding, with a narrow set of immune-related syndromes and diseases.

Based on the combined functional, evolutionary and GWAS-based evidence, we propose that VDR-BVs represent good targets for subsequent experimental validation using, for example, genome editing in cells. It is also expected that only a small minority of genes bound by VDR will be altered in expression. A large-scale expression QTL study in calcitriol-activated LCLs will thus be required to further narrow down the set of candidate disease genes whose regulatory elements are variably bound by VDR and which are differentially expressed across the human population.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Conclusions

We have provided the first quantitative association-based analysis to explain the genetic effect on binding affinity variation for a nuclear receptor, and to propose a mechanistic model for disease susceptibility. Our results support the hypothesis that DNA variants altering transcription factor binding at enhancers contribute to complex disease aetiology and suggest that altered VDR binding, and by inference variable vitamin D levels, explain, in part, altered autoimmune and other complex disease risk.

Methods

VDR ChIP-exo Sample Preparation

Calcitriol stimulated lymphoblastoid cells were grown and prepared for ChIP as described previously with some modifications [49] and adapted for ChIP-exo [15, 16]. Briefly, cells were incubated in phenol red free RPMI-1640, 10% charcoal stripped FBS, 2mM glutamine-L, penicillin with streptomycin solution (100 U/mL + 100 microG/mL) medium at 37C and 5% CO_2. Cells were harvested after stimulation for 36 hours with 0.1 μM calcitriol (Sigma) and crosslinked using a 1% formaldehyde buffer for 15 minutes at room temperature and quenched with 0.125 M glycine. Cells were lysed and chromatin sheared by sonication into fragments of $\sim$200-1000 bp. VDR-bound genomic DNA regions were isolated using a rabbit polyclonal antibody against VDR (Santa Cruz Biotechnology, sc-1008). Immunoprecipitated chromatin was then processed enzymatically on magnetic beads. Samples were polished, A-tailed, ligated to sequencing library adaptors and then digested with lambda exonuclease to remove nucleotides from 5' ends of double stranded DNA. Single-stranded DNA was eluted and converted to double-stranded DNA by primer annealing and extension. A second sequencing adaptor was ligated to exonuclease treated ends, PCR amplified, gel purified and sequenced. Libraries were prepared for Illumina as described in [16].

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data Processing and Peak Calling

Full data processing steps are presented in the Supplementary Material. Briefly, we mapped the ChIP-exo sequence reads (Peconics Inc., USA; single-end 40bp) against the hg19 build of the human reference genome [50] using BWA (v. 0.7.4, [51]) followed by Stampy (v 1.0.21 [52]). We filtered the reads to retain only uniquely mapping reads with $\mathrm{MAPQ} > 20$ and removed reads mapping to empirical blacklist regions identified by the ENCODE consortium [24].

For peak calling, we first computed strand cross-correlation profiles [53] of read start densities to obtain consensus estimates of the ChIP-exo digested fragment sizes, which we corrected for the presence of phantom peaks [54] using a mappability-corrected approach to cross-correlation inference (MaSC, [55]). We used the cross-correlation based estimates for the digested fragment size as a MACS2 (version 2.0.10, [56]) input parameter to obtain peak calls. Separately, to increase sensitivity, we called peaks using GPS/GEM (v. 2.4.1, [57]) with recommended ChIP-exo options `--smooth 3 --mrc 20`. We then pooled, for each sample, the two sets of peak calls (Supplementary Material). Finally, we used DiffBind [58] to manipulate the per-sample merged peaksets and to obtain consensus peakset (CP_{o_3}, $\mathrm{CP}_{o_{10}}$, $\mathrm{CP}_{o_{20}}$) based on an overlap threshold. To do this, we normalised the read numbers using DiffBind's embedded EdgeR routines, based on the Trimmed Means of M's (TMM) algorithm [59].

Interval Overlap Enrichment Analysis

We used the Genomic Association Tester [60] to perform the randomization-based interval enrichment analyses. All analyses were based on 10,000 randomisations over the chosen background, and backgrounds differed depending on the specific analysis performed. All analyses accounted for genome-wide patterns of GC content variability and, where relevant, for uniquely mappable regions (40bp reads). For the enrichment analysis of VDR-BVs in GWAS intervals and QTLs, we used an in-house bootstrapping pipeline to perform LD-filtering (to retain only LD-independent foreground and background variants) and DAF matching. Further details on annotations, background and analytical design are provided in the Supplementary Material.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Motif Discovery

Full motif analysis steps are documented in the Supplementary Material. Briefly, we carried out *de novo* motif analysis using multiple parallel approaches: MEME-ChIP [61] from the MEME suite [62], XXmotif [63] and PScanChIP [20] from the Weeder/MoDtools suite [64]. For the MEME-ChIP and XXmotif analyses, we used the peak intervals in $\mathrm{CP}_{o_{10}}$ and extracted the sequence underlying each interval ± 150bp from a repeat-masked version of the hg19 reference [50]. For the PScanChIP analysis we used as the input the full set of CP_{o_3} binding regions with Jaspar PWM descriptors, to which we added the best heterodimeric RXR::VDR PWM found by XXmotif and the best monomeric VDR PWM found by DREME (MEME-ChIP package). The global background chosen for the analysis consisted of built-in PScanChIP DNase I digital genomic footprinting data for the LCL CEPH individual NA12865. For the motif analysis at the 43,332 VDR-BVs we used PScanChIP [20] with Jaspar PWM descriptors and a mixed background of promoters and LCL DNase cut sites.

Differential Binding Analyses

We performed both allele-specific (VDR-ASB) and regression on genotype (VDR-QTL) differential binding analyses.

For the VDR-ASB tests, we utilised the sequence composition of ChIP-exo sequence reads overlapping heterozygous SNPs to determine the sequences originating from each allele separately [65, 66, 23] and to identify allele-specific binding events showing significant difference in the number of mapped reads between parental alleles. We employed a modified version of the AlleleSeq pipeline [23] to carry out the VDR-ASB analysis. Briefly, we tested for significant allelic imbalance among all read pile-ups intersecting a variant showing heterozygous genotype, given a null hypothesis of 50% paternal versus 50% maternal reads. We followed [23] in controlling for two major sources of bias typically encountered when running an allele-specific analysis of short read data: a bias due to mapping to the reference hg19 genome [67] and a bias due to unannotated copy number variants skewing the read counts for some of the SNPs being tested [68]. In addition, we corrected for a third source of potential erroneous ASB SNP calls by removing any significant allelic imbalance hypothesis falling under repeat regions contained in the ENCODE blacklist data [24].

For the VDR-QTL tests, we analysed all 27 LCL samples simultaneously. We grouped the samples

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

based on the genotypes of the underlying variant, and carried out QTL association testing between the variant's genotype (under the assumption of an additive genetic model [69]) and the phenotype at that location (the VDR binding affinity based on the quantile-normalised ChIP-exo peak size at that location). We used an in-house pipeline based on SNPs and indels imputed with IMPUTE2 [70] and Bayesian regression modelling based on SNPTEST [21] and BIMBAM [22]. Bayesian methods for analysing SNP associations are now an established tool in GWAS analysis [71, 72, 73] and show advantages over the use of p-values in power and interpretation [69] though they require tighter initial modelling assumptions when compared to frequentist methods. For details on the underlying model and priors we refer the reader to the Supplementary document.

Variant Annotation

We annotated all variants associated with differences in VDR binding and obtained by pooling the VDR-QTL and the VDR-ASB variants (VDR-Binding Variants, VDR-BVs) according to the guidelines in [25] and using a customised version of the Funseq2 suite of algorithms [26]. We added additional VDR-centric information to Funseq2: VDR CP_{o_3} binding regions, motif intervals from the PScan-ChIP analysis for the VDR:RXR Jaspar DR3 motif and for the DREME VDR monomer, and VDR dimer/monomer PWM information. Gene annotation used Funseq2's default (Gencode, v. 16) and the HOT annotation in Funseq was filtered to only retain LCL HOT information. We ran Funseq using the `-m 2 -nc` options so that all annotation referred to the ancestral allele of the variant.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Availability of data and material

Sequencing data from this study have been submitted to the Gene Expression Omnibus archive (GEO; `http://www.ncbi.nlm.nih.gov/geo`; accession number GSE73254).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Medical Research Council (GG, AJBT, WH, CPP), Wellcome Trust and Genzyme (SVR), the Multiple Sclerosis Society UK (Grant 915/09) and the Council for Science and Technology (CONACyT, Mexico, Grant 211990) (AJBT).

Authors' contributions

AJBT and SR designed the experiment and initiated the study; AJBT performed the ChIP-exo experiments and generated the data; GG devised the analytical strategy and undertook all computational analyses with guidance from WH, SR and CPP; GD assisted in the design of the experiment and in data generation, and revised the manuscript; GG and CPP wrote the manuscript with contributions from AJBT and SR.

Acknowledgements

We thank Mark Gerstein, Jieming Chen, Joel Rozowsky, Alexej Abyzov, Ekta Khurana and Yao Fu for precious help and technical insight on the AlleleSeq, CNVnator and Funseq2 platforms. We also thank Matthew Stephens and Yongtao Guan for help and insight on the PIMASS and BIMBAM platforms. We thank Gosia Trynka and Kamil Slowikowksi for providing r^2-based LD block data; Yuchun Guo and David Gifford for precious help and technical insight on the GEM peak caller; Gerton Lunter for help on the STAMPY platform; Andreas Heger for help and technical insight on the GAT software and genomic randomization testing; Rory Stark and Simon Anders for insight on differential analysis of ChIP-seq data; Laura Clarke for help with the 1000 Genomes data and Daniel Gaffney for providing

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

helpful pointers regarding the GEUVADIS QTLs calls; We thank Li Shen and Ning-Yi Shao (Icahn School of Medicine at Mount Sinai) for insight into NGSplot utilisation; and Oscar Bedoya Reina and Christoffer Nellaker for helpful discussions.

Additional Files

Additional file 1 — `Supplementary_material.pdf`

Supplementary text and supplementary figures 1–20.

Additional file 2 — `Supplementary_tables.xlsx`

Excel tables documenting details about VDR-BV statistics, variant enrichment tests and GWAS enrichment tests.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

Figure 1: **ChIP-exo peaks are reproducible among samples and with previous ChIP-seq results, and are enriched in promoters and enhancers. A**: Numbers of overlapping ChIP-exo peaks across 27 LCL samples (log scale). Highlighted rows indicate numbers of peaks called in at least three (CP_{o3}), ten (CP_{o10}) or twenty (CP_{o20}) samples. **B**: Normalised peak read depths [59, 58, 74] (Supplementary Material) are concordant between samples. VDR ChIP-exo peak binding affinity heatmap, showing pair-wise binding affinity correlation among all 27 sample pairs. (*Inset*) histogram of correlation counts. **C**: 50-fold enriched overlap between CP_{o3} peaks and the consensus ChIP-seq peakset from [14] (enrichment test: GAT [60], Supplementary Material). **D**: Meta-gene profile of VDR binding reads. The panel shows a meta-profile of read pile-up for the pooled 27 LCL samples, averaged over gene bodies (Ensembl v. 75, protein coding genes). **E**: Enrichment of VDR binding sites in the CP_{o3} consensus with ENCODE chromatin states (chromHMM [75]) from LCL NA12878, TF-dense regions (HOT [17] and ENCODE clustered TF binding sites) and DNase I hypersensitive areas (ENCODE). Asterisk denotes significance at 1% False Discovery Rate (FDR) threshold (enrichment test: GAT [60], Supplementary Material).

Figure 2: **VDR binding intervals are immune-associated or are housekeeping-associated tending to be either present far from, or adjacent to, transcriptional start sites, respectively. A**: The most highly ranking motif cluster ensuing from a *de novo* motif analysis of the VDR CP_{o3} consensus intervals using MEME-ChIP [61], including a CentriMO [76] analysis of enrichment with respect to VDR peak centres. **B-D**: GREAT analyses [77] for subsets of VDR peaks from the CP_{o3} consensus set containing either a strong-to-perfect (class I) or weak-to-non-existent (class II) instance of the RXR::VDR heterodimer motif. **B**: Distributions of distances between known TSS sites (UCSC hg19, Feb 2009) and VDR binding locations for classes I or II, or all, VDR peaks. **C,D**: Enrichment for Gene Ontology Biological Process and Disease Ontology terms for class I (**C**) or class II (**D**) RXR::VDR containing sites, respectively. For **C** and **D**, bars map to fold change whereas Hinton plots map to $-\log_{10}$ of the enrichment significance p-value.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3: **VDR-BVs are enriched at enhancers, in RXR::VDR DR3-type motifs, and in LCL eQTLs.** Genomic association testing of VDR-BVs with functional annotation. **A**: Enrichment of VDR-BVs within ENCODE chromatin state segmentation tracks for LCL NA12878, TF-dense regions (HOT [17] and ENCODE clustered TF binding sites) and DNase hypersensitive areas (ENCODE). **B**: Enrichment of VDR-BVs in VDR consensus motif intervals present in VDR CP_{o3} binding regions. **C**: Enrichment of VDR-BVs at DNase-QTLs and eQTLs from the Pritchard resource and CEU/YRI LCL eQTLs from the GEUVADIS resource. Significant enrichments are indicated using blue histogram bars (f.c. = fold change of observed versus expected overlaps; $\mathrm{FDR} = q < 0.1$; numbers in a box close to each bar indicate 'observed [expected]' overlap counts; grey bars indicate lack of significance). For panel **A**, enrichments shown are above-and-beyond the previously observed (Figure 1E) enrichments of VDR CP_{o3} peaks in the same functional annotation classes (relative to a background of 20,330 1000 Genomes SNPs in CP_{o3} VDR binding regions). For panel **B**, enrichments are relative to a background of 114,155 1000 Genomes variants lying under VDR ChIP-exo read pileups (≥ 5 reads) which had been tested as potential VDR binding affinity modifiers. This background was further corrected for the analyses in panel **C**, were only LD-independent foreground VDR-BVs were tested and 10,000 DAF-matched random background sets were extracted with replacement from the main set of 114,155 background variants.

Figure 4: **VDR-BVs are enriched in GWAS LD blocks for autoimmune, inflammatory and other diseases. A-B**: Traits and diseases showing enrichment of VDR-BVs (**A**) or VDR-rBVs (**B**). Diseases/trait associations were acquired from the GRASP catalog v2.0 [35]. Statistical association was computed between VDR-BVs and strong-LD (Supplementary Material) intervals around genome-wide significant ($p < 5 \times 10^{-8}$) disease tag-SNPs from the GRASP catalog. All available GRASP diseases and traits represented by at least 5 SNPs were analysed and those showing significant enrichment of VDR binding beyond a 1% FDR threshold were retained; disease associations supported by only 1 VDR-BV were discarded (full data are presented in the Supplementary Material). The sizes of black squares indicate the statistical significance of enrichments.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5: **Genetic variation modulating VDR binding affinity in LCLs within the canonical RXR::VDR heterodimeric binding motif**. Genome-wide quantification of the effect of genetic variation on VDR binding, based on the analysis of VDR-BVs in CP_{o3} binding peaks which hit the canonical RXR::VDR DR3 heterodimeric consensus motif (motif shown in panel **B**, JASPAR database, [19]). **A**: Distribution across the RXR::VDR motif of VDR-BVs which cause significant motif disruption (significance assessed via FunSeq2 using TFMPvalue [78], with default threshold $p < 4 \times 10^{-8}$) and quantification of the concordance between directionality of VDR PWM disruption and directionality of resulting VDR binding affinity variation. 102/115 (89%) VDR-BVs in CP_{o3} VDR peaks that significantly break the RXR::VDR motif predict the direction of VDR binding affinity change. **C,D**: Genome-wide quantification of the phenotypic effect of all VDR-BVs intersecting the RXR::VDR motif (including those which do not generate a motif break at the above significance level). The vertical axis (*Magnitude of VDR binding affinity change*) indicates the fold change of read depth of the ancestral versus the derived allele (panel **C**) or derived versus the ancestral allele (panel **D**). **C**: Impact on VDR binding affinity of Loss of Binding (LOB, orange dots) VDR-BVs and Loss of Binding VDR-BVs which test for significant RXR::VDR motif break (red dots). **D**: Impact on VDR binding affinity of Gain of Binding (GOB) VDR-BVs (blue dots). Nucleotide position has a significant influence on LOB (Kruskal-Wallis rank sum test, $\chi^2(14) = 33.385$, $p < 0.01$) but not on GOB ($\chi^2(12) = 17.68$, $p = 0.12$) phenotype magnitude. **C,D**: Grey dots indicate impact on binding affinity of those 1000 Genomes variants carried by the LCL samples which do not test for significant asymmetric binding (AlleleSeq [23], FDR threshold $= 0.1$) and are not, therefore, VDR-BVs.

Figure 6: **Evolutionary conservation of VDR-BVs at RXR::VDR consensus motifs.** The diagrams show distributions of the Derived Allele Frequency (DAF) for variants in RXR::VDR consensus motifs within reproducible CP_{o3} binding peaks. DAF values are separated by ethnicity (panels **A** and **C**: CEU; panels **B** and **D**: YRI); variants are separated by their effect on VDR binding affinity (panels **A** and **B**: VDR-BVs; panels **C** and **D**: 1000 Genomes variants carried by the LCL samples which do not test for significant asymmetric binding and are not, therefore, VDR-BVs). Within each of the four panels, variants are split in two groups, based on their effect on VDR binding affinity direction (whether GOB or LOB). For all quadrants, only DAFs for variants hitting hexamer positions (i.e. hitting either the RXR recognition element at positions 1-6 or the VDR recognition element at positions 10-15) in the RXR::VDR motif are shown. An asterisk indicates significance (non parametric Wilcoxon rank sum tests, $\alpha = 0.05$). Panel **A**: Wilcoxon rank sum test, $p = 2.8 \times 10^{-4}$. Panel **B**: Wilcoxon rank sum test, $p = 4.6 \times 10^{-5}$. Panels **C** and **D**: $p = 0.55$ and $p = 0.14$ respectively. *ns* = not significant.

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., *et al.*: Extensive variation in chromatin states across humans. Science **342**(6159), 750–752 (2013)

[2] McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., Pritchard, J.K.: Identification of Genetic Variants That Affect Histone Modifications in Human Cells. Science **342**(6159), 747–749 (2013). doi:10.1126/science.1242429. http://www.sciencemag.org/content/342/6159/747.full.pdf

[3] Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., *et al.*: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science **342**(6159), 744–747 (2013)

[4] Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.*: Systematic localization of common disease-associated variation in regulatory DNA. Science **337**(6099), 1190–1195 (2012)

[5] Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjalmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., *et al.*: Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. **95**(5), 535–552 (2014)

[6] Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., *et al.*: Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature **518**(7539), 337–343 (2015)

[7] Heinz, S., Romanoski, C.E., Benner, C., Glass, C.K.: The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol **16**(3), 144–154 (2015)

[8] Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., *et al.*: Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cells. Science **347**(6225), 1010–1014 (2015). doi:10.1126/science.1259418. http://www.sciencemag.org/content/347/6225/1010.full.pdf

[9] Holick, M.F.: Vitamin D Deficiency. New England Journal of Medicine **357**(3), 266–281 (2007). doi:10.1056/NEJMra070553. PMID: 17634462. http://www.nejm.org/doi/pdf/10.1056/NEJMra070553

[10] Martins, D., Wolf, M., Pan, D., et al: Prevalence of cardiovascular risk factors and the serum levels of 25-hydroxyvitamin D in the United States: data from the Third National Health and Nutrition Examination Survey. Archives of Internal Medicine **167**(11), 1159–1165 (2007)

[11] Mokry, L.E., Ross, S., Ahmad, O.S., Forgetta, V., Smith, G.D., Leong, A., Greenwood, C.M.T., Thanassoulis, G., Richards, J.B.: Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. PLoS Med **12**(8), 1–20 (2015). doi:10.1371/journal.pmed.1001866

[12] Afzal, S., Brondum-Jacobsen, P., Bojesen, S.E., Nordestgaard, B.G.: Genetically low vitamin D concentrations and increased mortality: mendelian randomisation analysis in three large cohorts. BMJ **349**(g6330) (2014)

[13] Kliewer, S.A., Umesono, K., Mangelsdorf, D.J., Evans, R.M.: Retinoid X receptor interacts with nuclear receptors in retinoic acid, thyroid hormone and vitamin D3 signalling. Nature **355**(6359), 446–449 (1992)

[14] Ramagopalan, S.V., Heger, A., Berlanga, A.J., Maugeri, N.J., Lincoln, M.R., Burrell, A., Handunnetthi, L., Handel, A.E., Disanto, G., Orton, S.-M., *et al.*: A ChIP-seq defined genome-wide map of vitamin D receptor binding: Associations with disease and evolution. Genome Research **20**(10), 1352–1360 (2010). doi:10.1101/gr.107920.110. http://genome.cshlp.org/content/20/10/1352.full.pdf+html

[15] Rhee, H.S., Pugh, B.F.: Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell **147**(6), 1408–1419 (2011)

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[16] Rhee, H.S., Pugh, B.F.: ChIP-exo Method for Identifying Genomic Location of DNA-Binding Proteins with Near-Single-Nucleotide Accuracy. In: Current Protocols in Molecular Biology, p. . John Wiley & Sons, Inc., ??? (2012). doi:10.1002/0471142727.mb2124s100

[17] Yip, K., Cheng, C., Bhardwaj, N., Brown, J., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., Gerstein, M.: Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biology **13**(9), 48 (2012). doi:10.1186/gb-2012-13-9-r48

[18] Carlberg, C., Bendik, I., Wyss, A., Meier, E., Sturzenbecker, L.J., Grippo, J.F., Hunziker, W.: Two nuclear signalling pathways for vitamin D. Nature **361**(6413), 657–660 (1993)

[19] Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., *et al.*: JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. **42**(Database issue), 142–147 (2014)

[20] Zambelli, F., Pesole, G., Pavesi, G.: PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res. **41**(Web Server issue), 535–543 (2013)

[21] Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P.: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet **39**(7), 906–913 (2007)

[22] Servin, B., Stephens, M.: Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. PLoS Genet **3**(7), 114 (2007). doi:10.1371/journal.pgen.0030114

[23] Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., *et al.*: AlleleSeq: analysis of allele-specific expression and binding in a network framework. Molecular Systems Biology **7**(1), (2011). doi:10.1038/msb.2011.54

[24] Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., *et al.*: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[25] Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al.*: Integrative annotation of variants from 1092 humans: application to cancer genomics. Science **342**(6154), 1235587 (2013)

[26] Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X., Yip, K.Y., Khurana, E., Gerstein, M.: FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. **15**(10), 480 (2014)

[27] Deplancke, B., Alpern, D., Gardeux, V.: The genetics of transcription factor dna binding variation. Cell **166**(3), 538–554 (2016). doi:10.1016/j.cell.2016.07.012

[28] Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D., Glass, C.K.: Effect of natural genetic variation on enhancer selection and function. Nature **503**(7477), 487–492 (2013)

[29] Nakane, M., Ma, J., Ruan, X., Kroeger, P.E., Wu-Wong, R.: Mechanistic analysis of VDR-mediated renin suppression. Nephron Physiol **107**(2), 35–44 (2007)

[30] Salehi-Tabar, R., Nguyen-Yamamoto, L., Tavera-Mendoza, L.E., Quail, T., Dimitrov, V., An, B.S., Glass, L., Goltzman, D., White, J.H.: Vitamin D receptor as a master regulator of the c-MYC/MXD1 network. Proc. Natl. Acad. Sci. U.S.A. **109**(46), 18827–18832 (2012)

[31] Capellino, S., Straub, R.H., Cutolo, M.: Aromatase and regulation of the estrogen-to-androgen ratio in synovial tissue inflammation: common pathway in both sexes. Ann. N. Y. Acad. Sci. **1317**, 24–31 (2014)

[32] Tuoresmaki, P., Vaisanen, S., Neme, A., Heikkinen, S., Carlberg, C.: Patterns of genome-wide VDR locations. PLoS ONE **9**(4), 96105 (2014)

[33] Huang, S., Laoukili, J., Epping, M.T., Koster, J., Hölzel, M., Westerman, B.A., Nijkamp, W., Hata, A., Asgharzadeh, S., Seeger, R.C., Versteeg, R., Beijersbergen, R.L., Bernards, R.: Znf423 is critically required for retinoic acid-induced differentiation and is a marker of neuroblastoma outcome. Cancer Cell **15**(4), 328–340 (2009). doi:10.1016/j.ccr.2009.02.023

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[34] Lappalainen, T., Sammeth, M., Friedlander, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., *et al.*: Transcriptome and genome sequencing uncovers functional variation in humans. Nature **501**(7468), 506–511 (2013)

[35] Eicher, J.D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J.P., Leslie, R., Johnson, A.D.: GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res. **43**(Database issue), 799–804 (2015)

[36] Matarese, G., De Placido, G., Nikas, Y., Alviggi, C.: Pathogenesis of endometriosis: natural immunity dysfunction or autoimmune disease? Trends Mol Med **9**(5), 223–228 (2003)

[37] Kunadian, V., Ford, G.A., Bawamia, B., Qiu, W., Manson, J.E.: Vitamin D deficiency and coronary artery disease: A review of the evidence. American Heart Journal **167**(3), 283–291 (2014). doi:10.1016/j.ahj.2013.11.012

[38] Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., Gilad, Y.: The functional consequences of variation in transcription factor binding. PLoS Genet. **10**(3), 1004226 (2014)

[39] Spivakov, M.: Spurious transcription factor binding: non-functional or genetically redundant? Bioessays **36**(8), 798–806 (2014)

[40] Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., Turner, J.M., Bertelsen, M.F., Murchison, E.P., Flicek, P., Odom, D.T.: Enhancer Evolution across 20 Mammalian Species. Cell **160**(3), 554–566 (2015)

[41] Jimenez-Lara, A.M., Aranda, A.: The vitamin D receptor binds in a transcriptionally inactive form and without a defined polarity on a retinoic acid response element. The FASEB Journal **13**(9), 1073–1081 (1999). http://www.fasebj.org/content/13/9/1073.full.pdf+html

[42] Toell, A., Polly, P., Carlberg, C.: All natural DR3-type vitamin D response elements show a similar functionality in vitro. Biochem. J. **352 Pt 2**, 301–309 (2000)

[43] Zhang, J., Chalmers, M.J., Stayrook, K.R., Burris, L.L., Wang, Y., Busby, S.A., Pascal, B.D., Garcia-Ordonez, R.D., Bruning, J.B., Istrate, M.A., Kojetin, D.J., Dodge, J.A., Burris, T.P., Griffin,

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

P.R.: DNA binding alters coactivator interaction surfaces of the intact VDR-RXR complex. Nat Struct Mol Biol **18**(5), 556–563 (2011). doi:10.1038/nsmb.2046

[44] Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S.E., Dermitzakis, E.T., Hirschhorn, J.N.: Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet **38**(2), 223–227 (2006)

[45] Fay, J.C., Wyckoff, G.J., Wu, C.I.: Positive and negative selection on the human genome. Genetics **158**(3), 1227–1234 (2001)

[46] Orlov, I., Rochel, N., Moras, D., Klaholz, B.P.: Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. EMBO J. **31**(2), 291–300 (2012)

[47] Reddy, T.E., Gertz, J., Pauli, F., Kucera, K.S., Varley, K.E., Newberry, K.M., Marinov, G.K., Mortazavi, A., Williams, B.A., Song, L., *et al.*: Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Research **22**(5), 860–869 (2012). doi:10.1101/gr.131201.111. http://genome.cshlp.org/content/22/5/860.full.pdf+html

[48] Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A.C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., Segal, E.: Unraveling determinants of transcription factor binding outside the core binding site. Genome Res. (2015)

[49] Ramagopalan, S.V., Maugeri, N.J., Handunnetthi, L., Lincoln, M.R., Orton, S.M., Dyment, D.A., Deluca, G.C., Herrera, B.M., Chao, M.J., Sadovnick, A.D., Ebers, G.C., Knight, J.C.: Expression of the multiple sclerosis-associated MHC class II Allele HLA-DRB1*1501 is regulated by vitamin D. PLoS Genet. **5**(2), 1000369 (2009)

[50] Kuhn, R.M., Haussler, D., Kent, W.J.: The UCSC genome browser and associated tools. Brief. Bioinformatics **14**(2), 144–161 (2013)

[51] Li, H., Durbin, R.: Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). doi:10.1093/bioinformatics/btp324. http://bioinformatics.oxfordjournals.org/content/25/14/1754.full.pdf+html

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[52] Lunter, G., Goodson, M.: Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. **21**(6), 936–939 (2011)

[53] Kharchenko, P.V., Tolstorukov, M.Y., Park, P.J.: Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. **26**(12), 1351–1359 (2008)

[54] Kundaje, A., Jung, L.Y., Kharchenko, P., Wold, B., Sidow, A., Batzoglou, S., Park, P.: Assessment of ChIP-seq data quality using cross-correlation analysis (2013). `http://code.google.com/p/phantompeakqualtools/`

[55] Ramachandran, P., Palidwor, G.A., Porter, C.J., Perkins, T.J.: MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. Bioinformatics **29**(4), 444–450 (2013)

[56] Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S.: Identifying ChIP-seq enrichment using MACS. Nat Protoc **7**(9), 1728–1740 (2012)

[57] Guo, Y., Mahony, S., Gifford, D.K.: High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Comput Biol **8**(8), 1002638 (2012). doi:10.1371/journal.pcbi.1002638

[58] Stark, R., Brown, G.: DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data. `http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf`

[59] Robinson, M., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology **11**(3), 25 (2010). doi:10.1186/gb-2010-11-3-r25

[60] Heger, A., Webber, C., Goodson, M., Ponting, C.P., Lunter, G.: GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics **29**(16), 2046–2048 (2013). doi:10.1093/bioinformatics/btt343. http://bioinformatics.oxfordjournals.org/content/29/16/2046.full.pdf+html

[61] Ma, W., Noble, W.S., Bailey, T.L.: Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc **9**(6), 1428–1450 (2014)

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[62] Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S.: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. **37**(Web Server issue), 202–208 (2009)

[63] Hartmann, H., Guthohrlein, E.W., Siebert, M., Luehr, S., Soding, J.: P-value-based regulatory motif discovery using positional weight matrices. Genome Res. **23**(1), 181–194 (2013)

[64] Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G., Pesole, G.: MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucleic Acids Research **34**(suppl 2), 566–570 (2006). doi:10.1093/nar/gkl285

[65] McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., *et al.*: Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. Science **328**(5975), 235–239 (2010). doi:10.1126/science.1184655. http://www.sciencemag.org/content/328/5975/235.full.pdf

[66] Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., *et al.*: Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature **464**(7289), 768–772 (2010)

[67] Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., Pritchard, J.K.: Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics **25**(24), 3207–3212 (2009). doi:10.1093/bioinformatics/btp579. http://bioinformatics.oxfordjournals.org/content/25/24/3207.full.pdf+html

[68] Pickrell, J.K., Gaffney, D.J., Gilad, Y., Pritchard, J.K.: False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics **27**(15), 2144–2146 (2011)

[69] Stephens, M., Balding, D.J.: Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. **10**(10), 681–690 (2009)

[70] Howie, B.N., Donnelly, P., Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. **5**(6), 1000529 (2009)

bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[71] Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., *et al.*: A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science **316**(5829), 1341–1345 (2007)

[72] Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., *et al.*: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447**(7145), 661–678 (2007)

[73] Wakefield, J.: Bayes factors for genome-wide association studies: comparison with P-values. Genet. Epidemiol. **33**(1), 79–86 (2009)

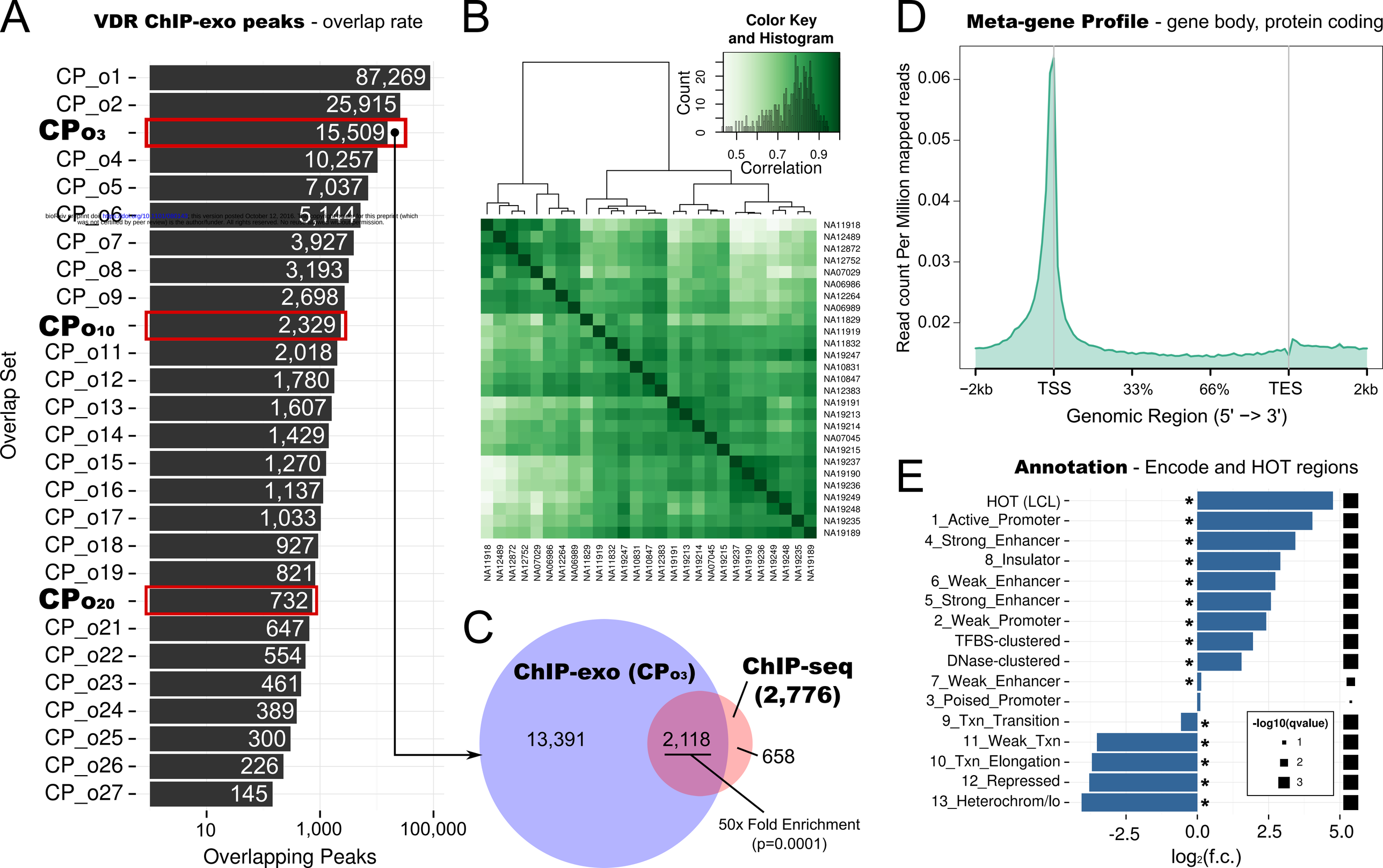
[74] Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D.: Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc **8**(9), 1765–1786 (2013)

[75] Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.*: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature **473**(7345), 43–49 (2011)

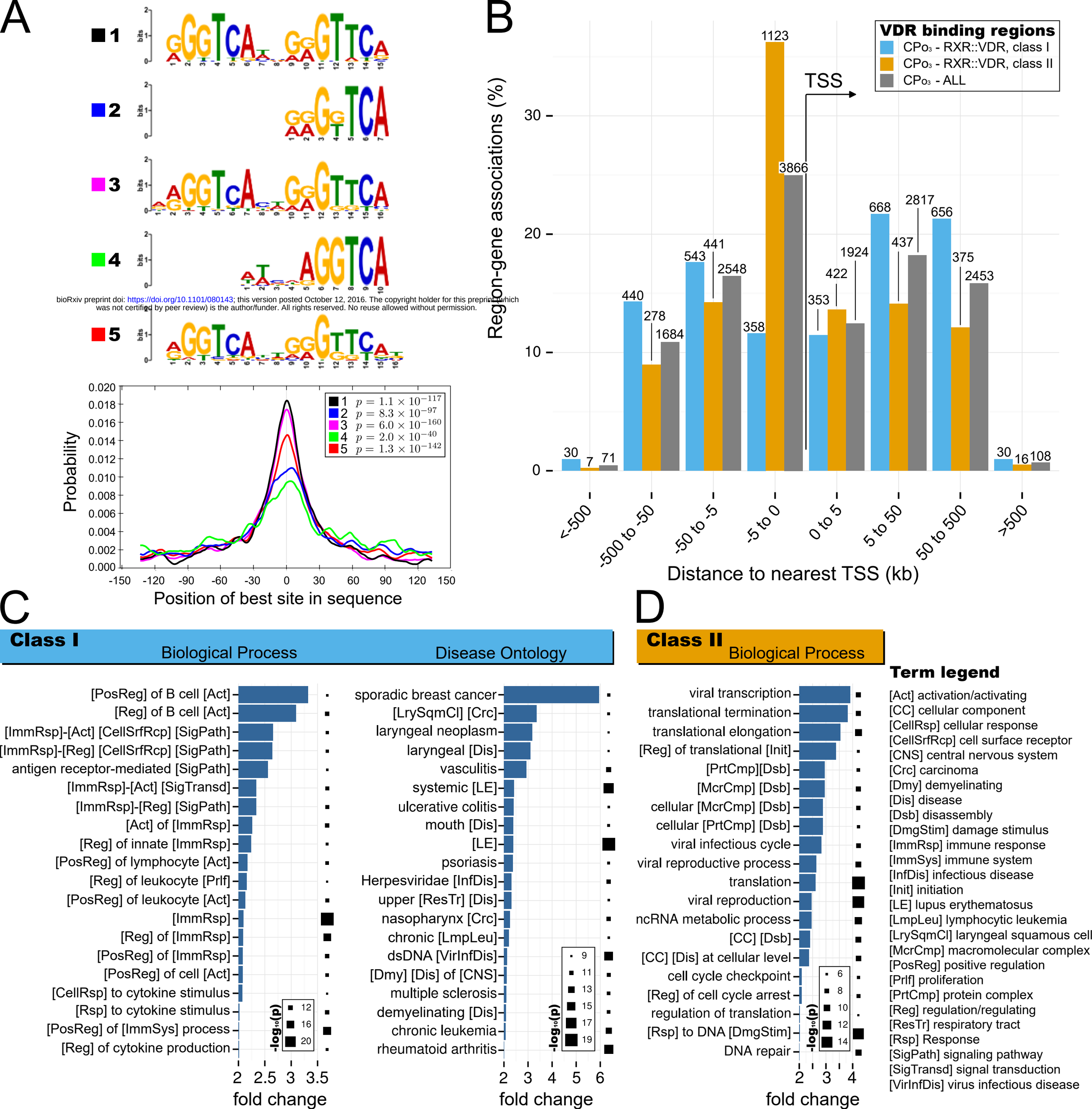
[76] Bailey, T.L., Machanick, P.: Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res. **40**(17), 128 (2012)

[77] McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G.: GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. **28**(5), 495–501 (2010)

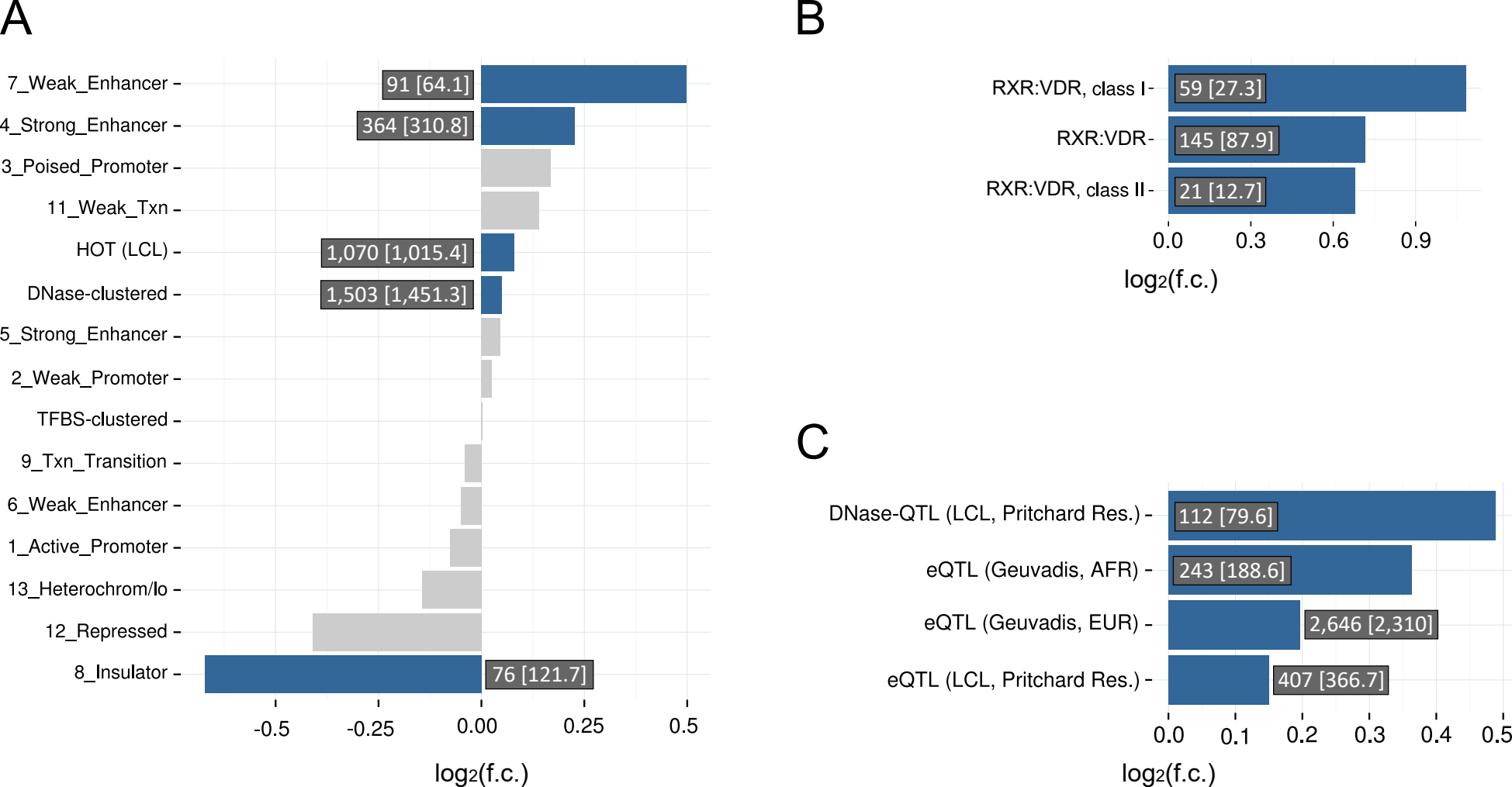
[78] Touzet, H., Varre, J.S.: Efficient and accurate P-value computation for Position Weight Matrices. Algorithms Mol Biol **2**, 15 (2007)

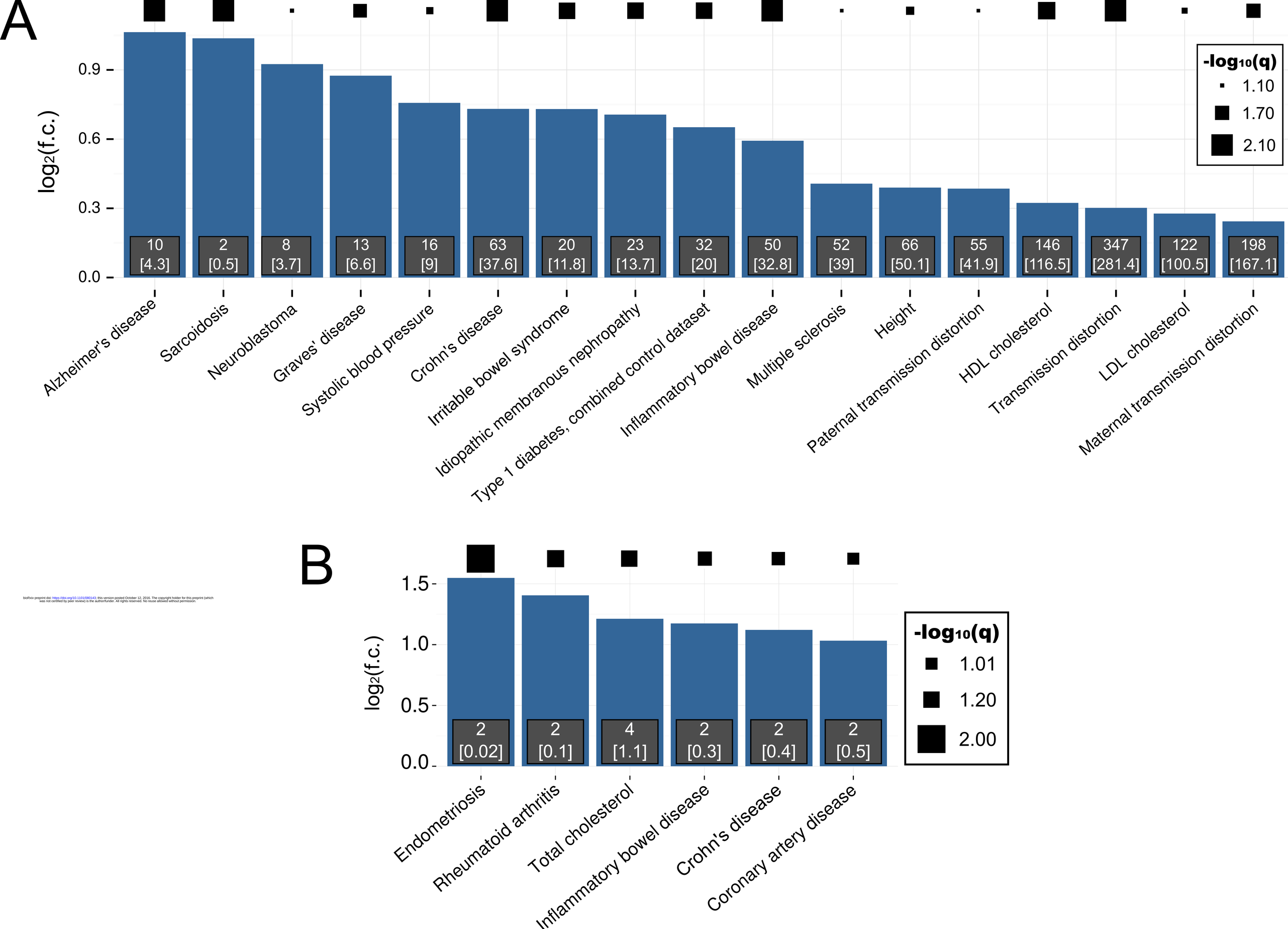


bioRxiv preprint doi: https://doi.org/10.1101/080145; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

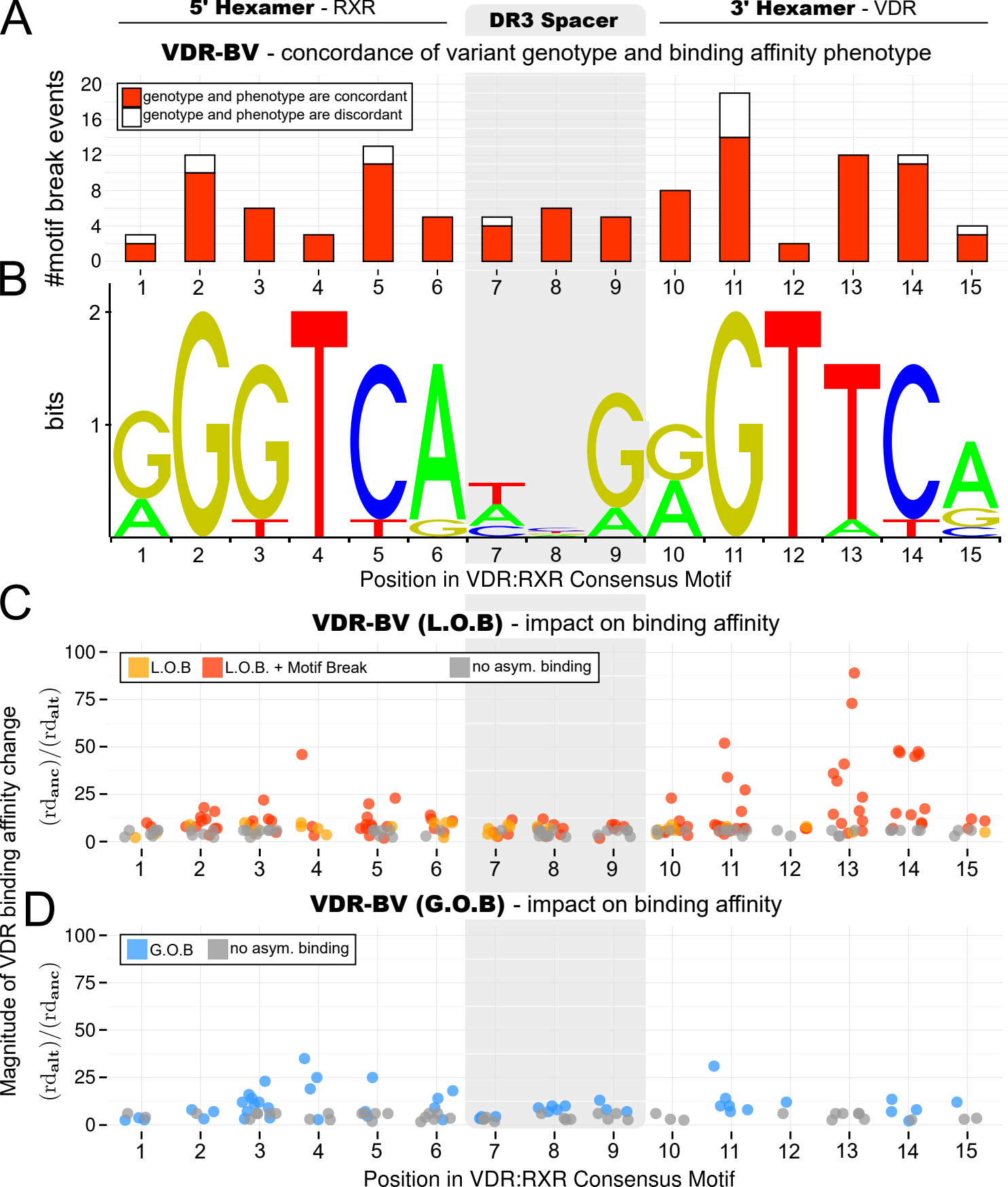


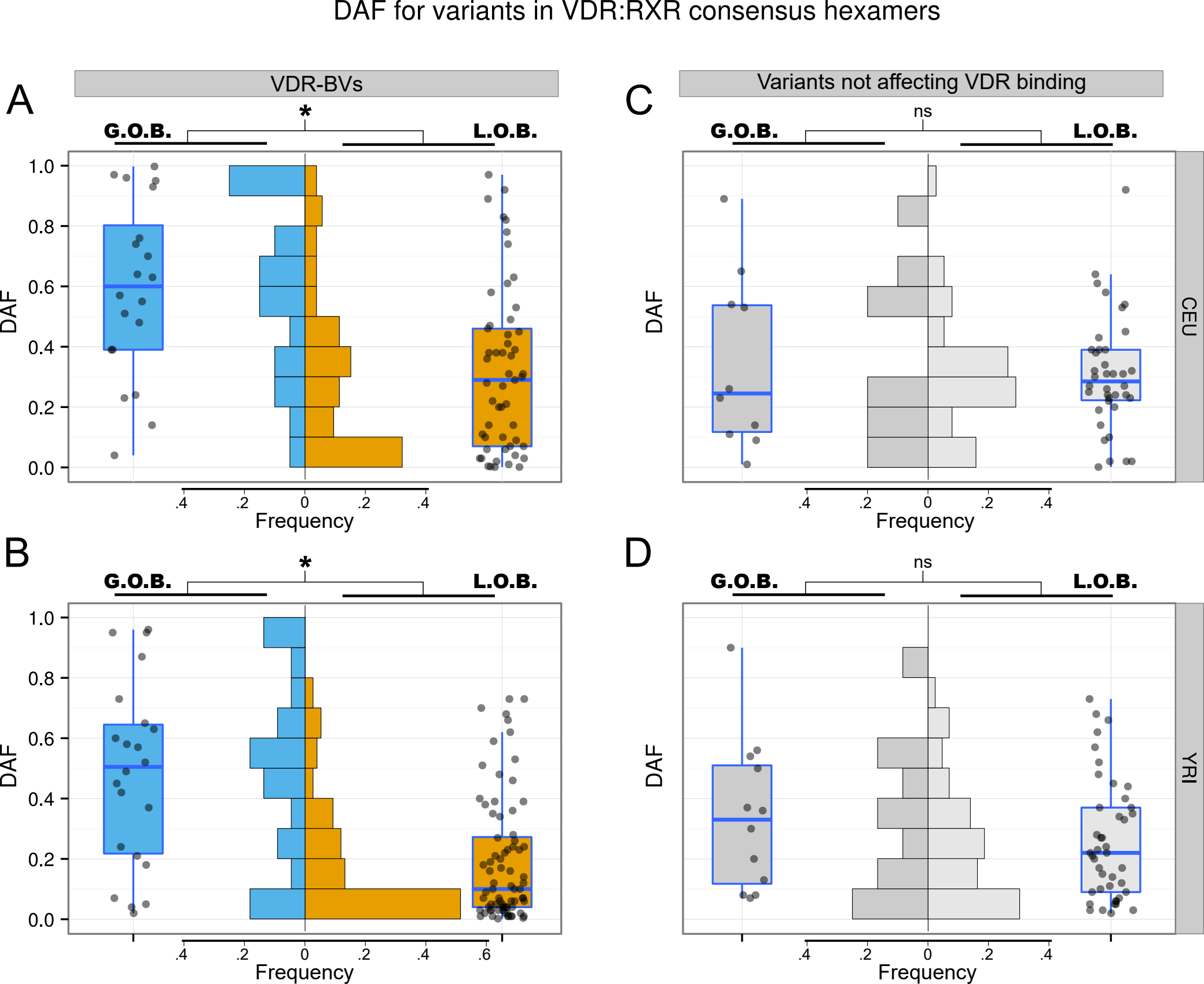
bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.


A
7_Weak_Enhancer
4_Strong_Enhancer
3_Poised_Promoter
11_Weak_Txn
HOT (LCL)
DNase-clustered
5_Strong_Enhancer
2_Weak_Promoter
TFBS-clustered
9_Txn_Transition
6_Weak_Enhancer
1_Active_Promoter
13_Heterochrom/lo
12_Repressed
8_Insulator
91 [64.1]
364 [310.8]
1,070 [1,015.4]
1,503 [1,451.3]
76 [121.7]
-0.5
-0.25
0.00
0.25
0.5
log2(f.c.)
B
RXR:VDR, class I
RXR:VDR
RXR:VDR, class II
59 [27.3]
145 [87.9]
21 [12.7]
0.0
0.3
0.6
0.9
log2(f.c.)
C
DNase-QTL (LCL, Pritchard Res.)
eQTL (Geuvadis, AFR)
eQTL (Geuvadis, EUR)
eQTL (LCL, Pritchard Res.)
112 [79.6]
243 [188.6]
2,646 [2,310]
407 [366.7]
0.0
0.1
0.2
0.3
0.4
0.5
log2(f.c.)



bioRxiv preprint doi: https://doi.org/10.1101/080143; this version posted October 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.


A
5' Hexamer - RXR
DR3 Spacer
3' Hexamer - VDR
VDR-BV - concordance of variant genotype and binding affinity phenotype
genotype and phenotype are concordant
genotype and phenotype are discordant
#motif break events
B
bits
Position in VDR:RXR Consensus Motif
C
VDR-BV (L.O.B) - impact on binding affinity
L.O.B
L.O.B. + Motif Break
no asym. binding
Magnitude of VDR binding affinity change
$(\mathrm{rd_{anc}})/(\mathrm{rd_{alt}})$
D
VDR-BV (G.O.B) - impact on binding affinity
G.O.B
no asym. binding
$(\mathrm{rd_{alt}})/(\mathrm{rd_{anc}})$
Position in VDR:RXR Consensus Motif


DAF for variants in VDR:RXR consensus hexamers
A
VDR-BVs
*
G.O.B.
L.O.B.
DAF
Frequency
B
*
G.O.B.
L.O.B.
DAF
Frequency
C
Variants not affecting VDR binding
ns
G.O.B.
L.O.B.
DAF
Frequency
CEU
D
ns
G.O.B.
L.O.B.
DAF
Frequency
YRI