SCIENTIFIC REPORTS
nature research

OPEN

Extensive fragmentation and re-organization of transcription in Systemic Lupus Erythematosus

Vasilis F. Ntasis[1], Nikolaos I. Panousis[2,3,4,5], Maria G. Tektonidou[6,7], Emmanouil T. Dermitzakis[2,3,4,8], Dimitrios T. Boumpas[7,8,9,10], George K. Bertsias[11,12] & Christoforos Nikolaou[1,12,13]✉

Systemic Lupus Erythematosus (SLE) is the prototype of autoimmune diseases, characterized by extensive gene expression perturbations in peripheral blood immune cells. Circumstantial evidence suggests that these perturbations may be due to altered epigenetic profiles and chromatin accessibility but the relationship between transcriptional deregulation and genome organization remains largely unstudied. In this work we propose a genomic approach that leverages patterns of gene coexpression from genome-wide transcriptome profiles in order to identify statistically robust *Domains of Co-ordinated gene Expression* (DCEs). Application of this method on a large transcriptome profiling dataset of 148 SLE patients and 52 healthy individuals enabled the identification of significant disease-associated alterations in gene co-regulation patterns, which also correlate with SLE activity status. Low disease activity patient genomes are characterized by extensive fragmentation leading to overall fewer DCEs of smaller size. High disease activity genomes display extensive redistribution of co-expression domains with expanded and newly-appearing (emerged) DCEs. The dynamics of domain fragmentation and redistribution are associated with SLE clinical endophenotypes, with genes of the interferon pathway being highly enriched in DCEs that become disrupted and with functions associated to more generalized symptoms, being located in domains that emerge anew in high disease activity genomes. Our results suggest strong links between the SLE phenotype and the underlying genome structure and underline an important role for genome organization in shaping gene expression in SLE.

Systemic Lupus Erythematosus (SLE) is considered the prototype of systemic autoimmune diseases due to highly heterogeneous manifestations, variability in symptoms, affected organs and alternating periods of dormancy and increased activity (flares)[1]. Several studies of SLE transcription profiles[2,3] have reported consistent alterations in key biological pathways, with the Interferon (IFN) signaling pathway being the most prominent example[4,5]. A recent systematic transcriptomic and genetic analysis comparing SLE patients, with variable disease activity, against healthy individuals led to the definition of discrete susceptibility and severity gene signatures[6]. Beyond gene expression, changes have also been observed at the epigenetic and chromatin levels, with extensive DNA hyper-hydroxymethylation in SLE T-cells[7] and altered chromatin accessibility in naive B-cells from SLE patients under flare status[8].

[1]Department of Biology, University of Crete, 70013 Heraklion, Greece. [2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [3]Institute of Genetics and Genomics in Geneva (iG3), University of Geneva Medical School, Geneva, Switzerland. [4]Swiss Institute of Bioinformatics, Geneva, Switzerland. [5]Wellcome Sanger Institute, Hinxton, UK. [6]Department of Propaedeutic Internal Medicine, Medical School, National and Kapodistrian University of Athens, Athens, Greece. [7]Joint Academic Rheumatology Program, Medical School, National and Kapodistrian University of Athens, Athens, Greece. [8]Biomedical Research Foundation of the Academy of Athens, Athens, Greece. [9]4th Department of Medicine, Attikon University Hospital, National and Kapodistrian University of Athens Medical School, Athens, Greece. [10]Medical School, University of Cyprus, Nicosia, Cyprus. [11]Department of Rheumatology, Clinical Immunology, Medical School, University of Crete, 70013 Heraklion, Greece. [12]Institute of Molecular Biology and Biotechnology (IMBB), Foundation of Research and Technology (FORTH), Heraklion, Greece. [13]Institute of Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Athens, Greece. ✉email: cnikolaou@fleming.gr

Given the complexity of the disease at both transcriptome and chromatin levels, an aspect that has not been adequately explored pertains to genome architecture. Over the last years, a number of genomic entities including chromatin loops[9], topologically associated domains[10], enhancer-promoter interacting domains[11], cis-regulatory domains[12,13] and domains of defined epigenetic characteristics[14,15] have been shown to define an ever more complex genomic landscape. In spite of their variable size, dynamics and the underlying principles governing their creation, a unifying property of these chromosomal entities is the co-ordination of gene expression[16,17]. At the same time, novel high-throughput methodologies have unravelled a strong link between nuclear compartments and transcriptional activity[18,19]. Positional effects in gene expression have been reported since relatively early and their evolutionary and regulation dynamics have been extensively studied[16,20–22]. The importance of gene clustering deregulation in disease has been demonstrated through epigenetics in the case of cancer[23] and genetic associations in the case of Down syndrome[24], but a comprehensive assessment of gene expression clustering has been lacking. Given the apparent extent and impact of genome organization, addressing gene expression changes from an architectural viewpoint could enhance our understanding of the genomic basis of complex pathological conditions, especially those that are accompanied by widespread gene expression alterations, such as SLE.

In this work, we have employed a genomic segmentation approach on an extensive SLE expression dataset[6], aiming to define regions of co-ordinated gene expression for the first time in the context of a complex disease. Our analysis leads to the definition of detailed patterns of transcriptional compartmentalization that vary significantly between SLE and healthy individuals. Interestingly, we find SLE patient genomes to exhibit more fragmented and thus, less structured co-expression patterns, a trend that correlates with the degree of disease activity. The defined *Domains of Co-ordinated Expression* (DCEs) exhibit intricate dynamics, that are associated with both molecular signatures and clinical features of the disease. This represents the first attempt to correlate the complex SLE phenotype with genome topology through detailed transcriptional analysis.

Methods

Analysis of gene expression. We obtained RNA sequencing data from a total of 142 SLE patients and 58 healthy individuals originally published in a SLE transcriptomics study[6]. Both groups contained individuals mainly of Caucasian ethnicity and an approximate ratio of 1:5 for male over female. The sequencing material was derived from whole blood samples. Extensive information regarding patient characteristics, mRNA extraction, sequencing protocol, quality control and mapping are thoroughly reported in[6].

We used FeatureCounts[25] to extract raw counts and quantify expression levels for a comprehensive set of human genes, as compiled under the GENCODE annotation v15 (https://www.gencodegenes.org/human/release_15.html, GRCh37). A fragment was counted in case of any overlap with an exon feature and the counts were grouped based on the "gene_name" attribute of the annotation entities. Only fragments with both ends successfully mapped were considered for summarization. Fragments that were chimeric, overlapping multiple meta-features (genes), not uniquely mapped, or having any read marked as duplicate were discarded.

The initial number of genes included in the raw count table was 51,716. A multi-step filtering approach was adopted. At first, the "type" of each gene was extracted from the annotation GTF file used in fragment summarization. Then, genes belonging to any of the following types were filtered out: "pseudogene", "processed transcript", "polymorphic pseudogene", "antisense", "sense intronic", "sense overlapping", IG_V pseudogene", "IG_C pseudogene", "TR_V pseudogene", "TR_J pseudogene", "IG_J pseudogene", "non_coding", "Mt-tRNA" and "Mt-rRNA". The total number of genes belonging to those categories were 20,190. Subsequently, 167 genes, which had multiple entries in the annotation file, with the same "gene_name", but different chromosome attribute, and could therefore generate errors in the fragment summarization process, were removed from our dataset as well. The number of genes that passed the filtering procedure was 31,318. Of those 27,061 with non-zero values were included in our analysis.

At the final stage, a two-step normalization was implemented on raw counts (filtered for the different irrelevant gene types), using relative log expression (RLE), followed by normalization for gene length.

Stratification of the patient cohort. We grouped patient samples according to a clinical SLE disease activity index (SLEDAI)[26]. A value of 0 for SLEDAI indicates inactive patients and it increases with higher disease severity. Patient samples were separated into three groups. A low disease activity group, with a maximum SLEDAI value of 2, an intermediate, with SLEDAI that ranged from 3 to 8, and a high disease activity group with SLEDAI greater than 8. The number of samples in each group were 55, 61 and 26 respectively. Both Differential and Topological analyses of gene expression have been performed on these three groups in comparison to the healthy group.

Differential gene expression analysis. Differentially expressed genes (DEGs) were called using zero inflated generalized linear models provided by the MDSeq tool[27]. For this analysis we applied an additional filtering layer. Genes with a mean cpm value lower than 0.05 were excluded. The remaining genes were 18,447. Furthermore, we incorporated gender and drug treatment as covariates in our models. DEGs were identified based on both statistical significance and effect size. They were defined as genes with corrected p value lower than or equal to 0.05 and absolute log2(Fold-Change) value greater than or equal to 0.5.

Modular analysis of differential gene expression. We followed a gene set enrichment approach, in order to investigate the over-representation of specific functional modules in our dataset. In a gene set enrichment analysis, the objective is to detect functional modules, whose gene members tend to cluster towards the top (or bottom) of a ranked list. Here, we ranked genes according to absolute differential expression values ($\log_2$|Fold-Change|), and we tested the over-representation of functional groups of genes (modules). The tested

modules are closely related to blood tissue and immunity, as identified by two independent studies[28,29], that analyzed a plethora of blood gene expression datasets in a variety of conditions. Finally, we assessed statistical significance by applying a CERNO test[30]. We filtered modules according to statistical significance (corrected p value ≤ 0.05) and their DEG content, i.e. at least 15% of gene members had to be DEGs according to our previous analysis.

Weighted gene co-expression network analysis (WGCNA). In order to detect modules of genes with correlated expression independently of their genome topology, we implemented weighted gene co-expression network analysis (WGCNA)[31]. Briefly, WGCNA represents genes as nodes in a network. These are connected to each other by edges, to which an adjacency score is attributed. The adjacency score of a node pair is calculated by a power function of the absolute value of the correlation of the corresponding pair of genes. Here, the soft threshold parameter (the power in the adjacency function) was selected to be 10, according to scale free topology criterion, which suggests choosing the lowest possible value, such that approximate scale free topology is reached in the network. Modules of co-expressed genes were extracted from the network based on the hierarchical clustering of a topological overlap measure and the subsequent implementation of a dynamic cutter. Those initial modules were merged using hierarchical clustering of their eigengene vectors and by cutting the resulted tree at the height of 0.25. The final modules were functionally characterized by utilizing pathway enrichment and calculating the correlations of the module eigengene vectors with a variety of clinical traits.

Robust co-expression matrix calculations. Each chromosome was split in 10 kb bins, starting from the start of the first gene, till the end of the last gene. A mean gene count for every bin was then calculated from the normalized counts of the genes it contained. Bin counts were then grouped according to four sample categories (healthy, low, intermediate and high SLE activity). Next, we calculated the Spearman correlation matrix between all bins that resided in the same chromosome for all samples within each group. Chromosomal bins with zero expression were ignored for the rest of the analysis. This procedure produced a square correlation matrix for each chromosome. To statistically evaluate the correlation coefficients, a Monte-Carlo-like approach was implemented. The bin counts, of each individual separately, were shuffled randomly and afterwards the correlation matrix was re-constructed. That procedure was repeated 1000 times for each chromosome. In every iteration the calculated correlation coefficients were compared to the original correlation coefficients that were calculated using the intact bin counts. The p value for each coefficient was set as equal to the fraction of those 1000 permutations, in which the corresponding coefficient had the same or more extreme value compared to the actual one. The correlation coefficients with p value greater than 0.05 were discarded from the analysis (turned into 0 s).

Definition of domains of coordinated expression (DCEs). To call domains of co-ordinated expression, we modified a methodology that was introduced for the definition of topologically associated domains (TADs)[32], in our case, by using expression correlation data instead of chromosomal contact frequencies. Statistically robust expression correlation matrices (see "Methods" section) were used as input. Domains of co-ordinated expression (DCEs) were defined as genomic regions of consecutive chromosomal bins with correlation above average, delimited by statistically significant boundaries. More specifically, DCE detection is a four-step pipeline, which is repeated for each chromosome and for every study group (see Fig. 1a).

1. First, we compute a signal that runs along the chromosomes and is indicative of the local average correlation of expression. We achieve this by sliding two juxtaposed windows of equal size along a chromosome with a single-bin displacement, until the whole chromosome has been covered. In every iteration, we use the correlation matrix that has already been constructed and statistically evaluated. We look up the correlation values concerning the relationship of the two regions and calculate their average. That value is assigned to the chromosomal bin located in the middle, more precisely, the downstream-most bin inside the upstream window.
2. Subsequently, the calculated signal is used to detect DCE boundaries. Hence, the second step of the pipeline is to compute a smoothed function of that signal, using a smoothing spline, and to detect all local minima of that function. DCEs are initially detected as regions between local minima with a value lower than 0.25, which is the average genome signal for the healthy, control group.
3. The third step is to statistically evaluate and refine the boundaries. We estimate the significance of the boundaries by utilizing a Mann–Whitney U test to compare "within" and "in-between" correlation coefficients. In case any of the initially calculated boundaries does not reach the required statistical significance threshold (p value > 0.05), we "chop" that boundary by one bin towards the centre of DCE, and repeat the test. DCEs with any remaining non-significant boundary are discarded.
4. In the last step we fuse neighbouring DCEs, according to the following criteria: (a) they are separated by at most two bins with bin signal < 0.25. (b) The total number of such "intervening" low signal bins is less or equal to 2. This means that no more than three neighbouring DCEs may be fused in one and that fusion events cannot span intervening sequences longer than two bin sizes. This step enhances the robustness of the pipeline and decreases the noise in our data. The window size used in bin-signal calculation and in boundary evaluation was set to be equal to 3, based on the maximization of average intra-DCE correlation of chromosome 1 of the healthy group.

Cell type estimation and entropy calculation. We used the results of CIBERSORT[33] for the estimation of the proportion of different immune cell types in whole blood. Shannon entropy was used as a metric,

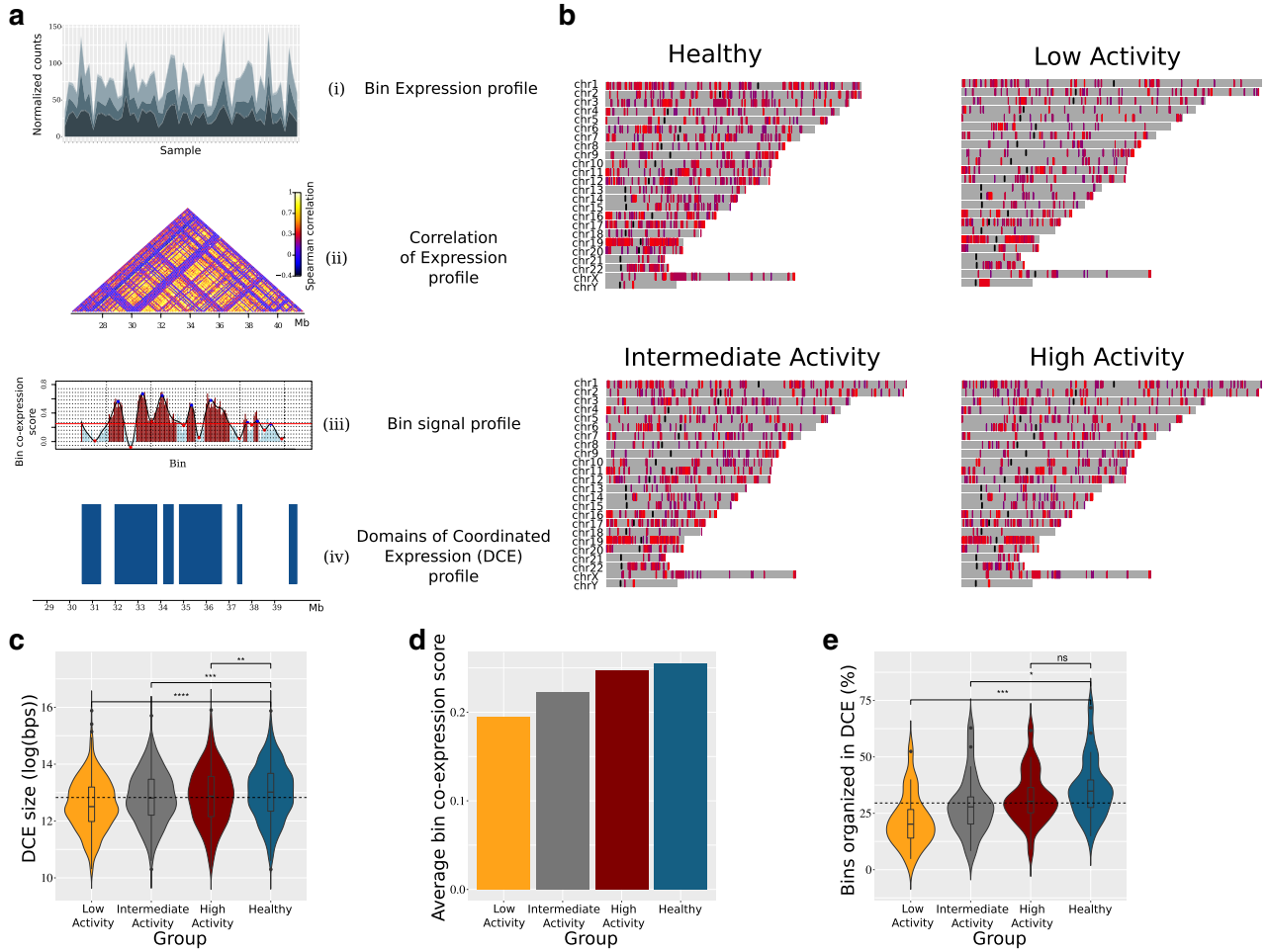

Figure 1. Differential patterns of domains of co-ordinated expression (DCEs) in healthy and patient groups. (**a**) The DCE detection pipeline is represented as a series of 'transformations' applied to the expression data. We start by calculating the expression profile of each chromosomal bin using the expression profile of the encompassed genes (i). We then calculate the correlation coefficients between the bins located on the same chromosome (ii). Next, the correlation profile of each chromosome is transformed into a one-dimensional binsignal profile (iii). We analyze that profile, detecting local minima and maxima in order to determine the borders of the domains. Finally, a statistical evaluation of those borders results in the final DCE coordinates (iv). (**b**) Domainograms depicting the distribution of DCEs for the healthy and the three patient groups studied. The color of DCEs represent the respective average binsignal of the chromosomal bins encompassed. (**c**) Violin plots illustrating the estimated distribution of DCE sizes in each group. Classic boxplots are included. The scale of the y axis is logarithmic (log(bps)). (**d**) Average bin signal (co-expression score) for each group. (**E**) Violin plots representing, for each chromosome, the percentage of chromosomal bins that contain genes, with non-zero expression value, and form DCEs. (**c**, **e**). The results of Mann–Whitney–Wilcoxon tests comparing each patient group to the healthy group are demonstrated by the significance level indicators.

in order to assess the variability/uncertainty in the proportions of the different cells types between healthy and SLE subjects.

$$H = -\sum_{i=1}^{n} p(x_i) \cdot \log_2(p(x_i))$$

where H is the Shannon (information) entropy, $p(xi)$ is the estimated proportion of x_i cell type in whole blood and n is the total number of estimated cell types. Entropy was calculated for every individual in the dataset. Subsequently, the difference between the distribution of entropies of healthy and SLE groups were statistically evaluated by a non parametric Wilcoxon–Mann–Whitney test.

DCE analysis. Two different metrics were applied to explore the differences of DCE sets of different groups. DCEs were handled as a set of chromosomal intervals. The first metric used was the Jaccard similarity coefficient. DCE pairs between two different groups (e.g. healthy and patient groups) with chromosomal coordinates that overlap were detected. For every pair the Jaccard index was calculated. The second metric was the BP distance score[34]. BPscore takes into account the relative chromosome size and thus may provide a more nuanced

assessment of coordinate similarity. For the calculation of BP-score we used a publicly available python script, (https://github.com/rz6/bp-metric), that is provided by the authors.

"Disruptor gene" definition. "Disruptor" genes are defined as those, which reside between two "patient DCEs" derived from a split of a "healthy DCE". More precisely, for each "split" characterized healthy DCE, attributed genes were listed. Subsequently, they were compared as a set with those genes, attributed to patient DCEs, that overlapped the initial healthy DCE. Genes absent from the patient DCEs and present in the healthy DCE were characterized as disruptors. Moreover, those genes were filtered, in order to capture only the genes located in the area between the patient DCEs (the locus of the "split" event). That process was repeated for every distinct patient group.

Results

Gene co-expression patterns are fragmented in SLE patients. Neighbouring gene expression correlation and modelling[16] have been recently introduced to define how gene expression propagates in space. We employed a topologically-inspired approach to quantify the correlation of gene expression genome-wide. After splitting each chromosome in fixed-size bins, we calculated the transcript count correlation and defined regions of significant co-expression based on a permutation test, followed by local minima localization (see "Methods" section). The *domains of co-ordinated expression* (DCEs) produced through this analysis are supported by permutation analysis involving 1000 random reshuffling events of transcript counts along each chromosome (see "Methods" section, Fig. 1a). In this respect, they correspond to statistically robust chromosomal domains, within which gene co-expression is significantly higher, when compared to the surrounding regions.

Analysis of DCE patterns between SLE and healthy individuals shows significant differences, with SLE gene co-expression being organized into smaller and more fragmented regions. This finding is not confined to specific chromosomes, although gene-dense chromosomes with a more compact transcript pattern show increased overall signal (Fig. 1b). Notably, DCE patterns correlate with the activity of the disease (SLEDAI). We found that DCE sizes are smaller in low activity patients, where the percentage of the genome organized in DCE does not exceed 9% as compared to 13% and 17%, for intermediate and high activity respectively, and 19% for healthy individuals. Decreased gene co-expression in SLE patients is evidenced by the: (a) significantly lower numbers of total DCE for low and intermediate disease activity (Fig. 1b), (b) smaller DCE sizes (Fig. 1c) (c) decreased co-expression signal (Fig. 1d) and, (d) smaller overall percentage of the genome covered by DCEs (Fig. 1e).

It is important to note, that the observed differences cannot be explained by batch effects in either sequencing output or the genomic distribution of reads. Sequencing throughput was very similar for all disease activity groups (Supplementary Figure 1), as was the overall distribution of mapped reads in the annotated transcriptome (Supplementary Figure 2). Differences in DCE patterns cannot be attributed to cell type heterogeneity either, as shown by an entropy analysis of cell type variability (Supplementary Figure 3). Thus, the more fragmented expression patterns in low activity SLE genomes are most likely due to generalized perturbations in gene regulation, which could provide a mechanistical explanation for the recurrent flares that tend to develop in patients who are inactive. This may indicate that, while a desirable outcome, clinical remission may not necessarily be lacking a molecular fingerprint and the combination of the recently suggested susceptibility signature[6] with our fragmented DCE pattern may provide an interesting framework for the assessment of its stability.

DCEs are dynamically redistributed in SLE. To gain additional insight into the dynamics of DCEs, we classified DCE patterns into four main groups according to their changes between patient and healthy genomes. We used an implementation of the Jaccard Index to group the domains into: (a) DCEs that were left *intact*, (b) DCEs that were absent (*depleted*) in patients while present in healthy individuals, (c) DCEs that were only present (*emerged*) in patients and, (d) DCEs whose coordinates were altered between patient and healthy genomes. The last group was further categorized into DCEs that were *split* (one fragmented into two or more smaller sub-DCEs) or *merged* (two or more joined into one larger) and *expanded* or *contracted*.

Low and high disease activity patients showed the most extensive changes in the pattern of DCEs as compared to the healthy state (Fig. 2a). A detailed analysis shows that, in agreement with the changes observed at genome-scale level (Fig. 1), there is extensive fragmentation and redistribution of domains in SLE versus healthy genomes. Contraction and depletion of DCEs are more pronounced in low activity patients, with *contracted* and *depleted* DCEs corresponding to nearly 73% of DCEs in low activity, as compared to 56% and 48% for intermediate and high disease activity genomes, respectively (Fig. 2a). Conversely, *expanded* and *emerged* DCEs comprise over 30% in high activity versus less than 10% in low activity patients (Fig. 2a). These observations are suggestive of different modes of dynamic changes in co-expression domains, with low SLE activity genomes characterized by DCE fragmentation and high activity ones featuring a redistribution of co-expression with increased percentages of *expanded* and *emerged* DCEs. This redistribution was also supported by a simple value measure of DCE pattern similarity, calculated with the implementation of BPscore[34], which showed that in spite of being comparable in genome coverage, the DCEs between high activity patients and healthy controls were radically different in terms of genomic localization (Supplementary Figure 4).

Gene expression changes are reflected upon DCE dynamics. Changes in the patterns of co-expression may be linked to differential gene expression and underlying chromatin dynamics. To address this, we employed Modular and Weighted Gene Co-Expression Network Analysis (WGCNA)[31] of differential gene expression on the three disease activity groups against healthy individuals. The results were strongly suggestive of quantifiable phenotypic variability between patients with different clinical activity states, in agreement with the previously defined susceptibility and severity gene signatures (Supplementary Figure 5). In addition, we were

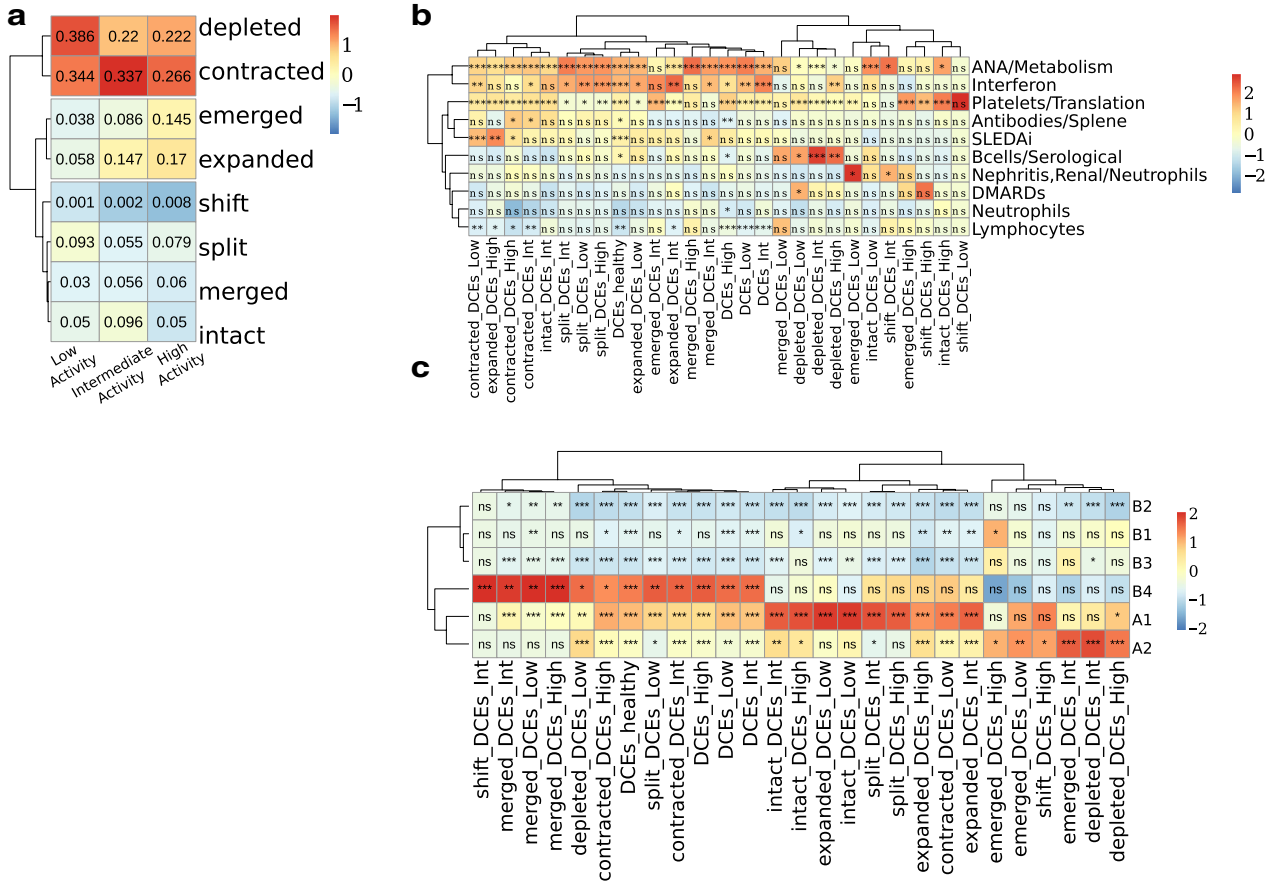

Figure 2. DCEs are extensively fragmented and redistributed in SLE patients and correlate with functional signatures and epigenetic marks. (**a**) Heatmap presenting the different types of DCE reorganization. Numbers inside cells indicate the ratio of the number of DCEs, of the respective type, over the total number of DCEs for each patient group. Colour code is corresponding to column z-score of ratios. (**b**) Heatmap depicting the results of an enrichment test for DCEs in the functionally annotated WGCNA modules. (**c**) Heatmap depicting the results of an enrichment test for DCEs in different genome subcompartments. (**a**–**c**) Scaling and centering has been performed per column. Trees are illustrating the outcome of hierarchical clustering performed on the data. (**b**, **c**) Symbols inside cells demonstrate the significance level of the outcome of each test (*:0.05; **:0.01; ***:0.001). Significance has been assessed by a non-parametric, permutation-based test.

able to define gene expression modules and to correlate them to clinical characteristics such as disease activity. Comparison of WGCNA results with clinical characteristics of the cohort samples allowed us to identifiy a "*SLEDAI gene module*", which comprised 224 genes, enriched for innate and adaptive immune pathways, particularly signaling through the Fc-γ and B-cell receptors (BCR). A "*Nephritis module*" (184 genes) and an "*IFN module*" (282 genes) were also identified, the latter being highly associated with anti-nuclear and anti-DNA antibodies (Supplementary Figure 6).

In order to investigate how changes in the gene expression of these modules may be correlated to the patterns of positional co-expression, we analyzed the degree of overlap between the different DCE categories and the WGCNA modules obtained from our dataset (Fig. 2b). The *IFN module* was over-represented in *split* DCEs across all SLE groups and was particularly enriched in DCEs that become *depleted* within the high activity group, implying that increased IFN pathway gene activity may be linked to loss of co-expression structure. Conversely, the *nephritis/neutrophil-specific module* was enriched in *emergent* DCEs from low disease activity genomes. This may be indicative of underlying tendencies in gene deregulation being present even in patients without developed symptoms but who may yet be predisposed to disease flares. Consistent with observations at the level of functional enrichments, the *B-cell module* was enriched in DCEs that are depleted and largely absent from high disease activity DCEs. Taken together, these findings indicate that functional aspects of gene expression pertaining to distinct clinical characteristics are reflected on the genome organization.

DCE dynamics are strongly associated with chromatin accessibility and chromosomal compartments. Transcriptional coordination in self-contained domains is tightly linked to underlying chromatin organization at various levels ranging from topologically associated domains (TADs) to more extended chromosomal compartments. We went on to correlate the dynamics of DCE patterns with underlying genomic features related to chromatin accessibility and chromosomal compartmentalization. By comparing the coordi-

nates of stable and dynamic DCEs against ATAC-Seq peaks defined for B-cells in severe-case SLE against healthy individuals[8], we found *split*, *contracted* and *merged* DCEs (of all disease activity groups) to be enriched in peaks of decreased chromatin accessibility (Supplementary Figure 7). Conversely, *depleted* and *emerged* DCEs of all SLE activity groups were enriched (although with a smaller effect size), almost exclusively, in over-accessible regions. This finding suggests a clear distinction between the DCEs that are locally modified, which tend to be confined in under-accessible regions, and those that are dynamically re-distributed, which are preferentially located in more accessible chromatin.

We performed a similar analysis at the level of chromosomal compartments (at 100kbp resolution) as defined in a B-lymphoblastoid cell line[9]. On a large scale, chromosomes may be organised into two broad compartments labelled A and B, corresponding to active and inactive chromatin, and also bearing other distinct properties. These may be further subdivided to A1 and A2 and B1 to B4[9]. A chromosomal coordinate overlap enrichment analysis showed DCEs to be generally enriched in the euchromatic A compartment (Fig. 2c). When focusing on specific DCE subtypes, we found that regions, belonging to the most dynamic subsets of *emerged* and *depleted* DCEs, were enriched in the A2 subcompartment, which is associated with late-replicating, low GC content DNA, enriched in H3K9me3 and longer gene transcripts[9]. On the other hand, intact DCEs and in general, DCEs that are less dynamic, appear to be more enriched in the gene dense, early-replicating A1 subcompartment. Enrichments in the B4 subcompartments are probably due to the over-representation of particular DCEs in chromosome 19, which hosts the entirety of this very small subcompartment.

Together, the differential enrichments of *split* and *contracted* DCEs, compared to the dynamically redistributed *emerged* and *depleted* regions, in terms of chromatin accessibility and genome compartments, indicate an interplay between gene regulation and underlying chromatin environment. Regions of high gene density tend to have highly correlated gene expression, but this pattern changes radically with the splitting of co-expression domains in low disease activity and the emergence of new, probably re-arranged domains in high disease activity SLE patients. We hypothesize that epigenetic changes that increase chromatin accessibility, in particular in A2 genomic compartments, may create a permissive environment for the redistribution of co-regulated genomic domains, which are, moreover, associated with functions characteristic of increased disease activity.

DCE splits disrupt enhancer-promoter interactions of key biological functions. While *split* DCEs represent no more than 5–10% of the total genome coverage, they are highly enriched among differentially expressed genes and in particular with the *IFN gene module*. Given their additional enrichment in low disease activity patients and therefore, their possible implication in further disease progression, we performed a focused analysis of *split* DCEs and the genes lying on their boundaries (see "Methods" section). These were predominantly enriched among the targets of specific transcriptional regulators, a number of which belonged to the broad categories of zinc fingers (SALL1, Ikaros, ZIC3 etc.) and oncogenes (GLI1, ING4) (Fig. 3a). Members of the Ikaros transcriptional regulators have been genetically associated with SLE[2], and interestingly, *IKZF3* lies within a disrupted DCE in all SLE groups.

Based on the differences in the extent of split DCEs between low and high activity genomes, we next assessed their overlaps with the SLE susceptibility and severity gene signatures as previously defined for the same dataset[6]. We found significant differences between the two gene sets with *susceptibility genes* being highly enriched in *split* DCEs in contrast to a depletion of severity genes (Fig. 3b). Genes belonging to the *susceptibility signature* are also enriched in the subset of differentially expressed genes that are found in low disease activity *split* DCE boundaries ($p = 0.0061$). Protein and regulatory interaction network analysis of these genes, performed through STRING-DB[35], revealed an IFN gene signature (Fig. 3c) and interestingly, a set of highly connected genes associated with DAP12 signaling (Fig. 3c, cyan polygon). DAP12 (TYROBP) is a key activator of NK cells, which are reported to have impaired function in SLE patients[36]. Smaller network modules were associated with neutrophils (lime) and B-cells (yellow). We may thus see how, by focusing on split DCE regions we may prioritize genes of the broader susceptibility signature and to investigate their functional connections.

Given the DCE definition as regions with increased regulatory interactions, it is plausible to expect that gene promoters are more likely to be associated with enhancers that are lying within the same region. To test this hypothesis, we obtained cell-type specific promoter-enhancer interactions for CD4, CD8 and CD14 and CD19 cells from Enhancer Atlas[37] and identified genes whose promoter-enhancer pairs were nested within the same DCE in the healthy state but disrupted in SLE. We found that a significant percentage of enhancers-promoter connections that are completely nested in healthy DCEs are disrupted by a DCE *split* or *depletion* in one of the SLE disease activity states. Thus, it seems that the redistribution of gene co-regulation domains in disease may also be disrupting the regulatory links between enhancers and their cognate promoters.

Functional enrichment of the genes, whose enhancer-promoter associations are disrupted in SLE, revealed relevant biological functions (Fig. 3d). More specifically, functions related to the *immune system* are, as expected, highly enriched in all disease activity groups. Others, such as *protein metabolism*, *translation* and *protein turnover* are particularly enriched in high disease activity patients. Interestingly, interleukin-15 (*Il15*) and interleukin-21 (*Il21*) signaling are specifically enriched in high activity patients even though with low effect sizes (Fig. 3d). In the context of SLE, increased Il15 levels may regulate the function of NK cells and also enhance the expression of the costimulatory receptor CD40L (CD154) on T-cells via STAT5[38]. Interleukin-21 is released by CD4+ T follicular helper cells and plays an important role in SLE pathogenesis by promoting the maturation of B-cells into autoantibodies-producing plasma cells[39]. More interestingly, the receptors of Il15 and Il21 share the common gamma chain (γc) subunit (CD132) and mediate intracellular effects through activation of the Janus kinase (JAK)-1 and JAK-3 kinases, which are implicated in SLE and are currently tested as putative therapeutic targets[40].

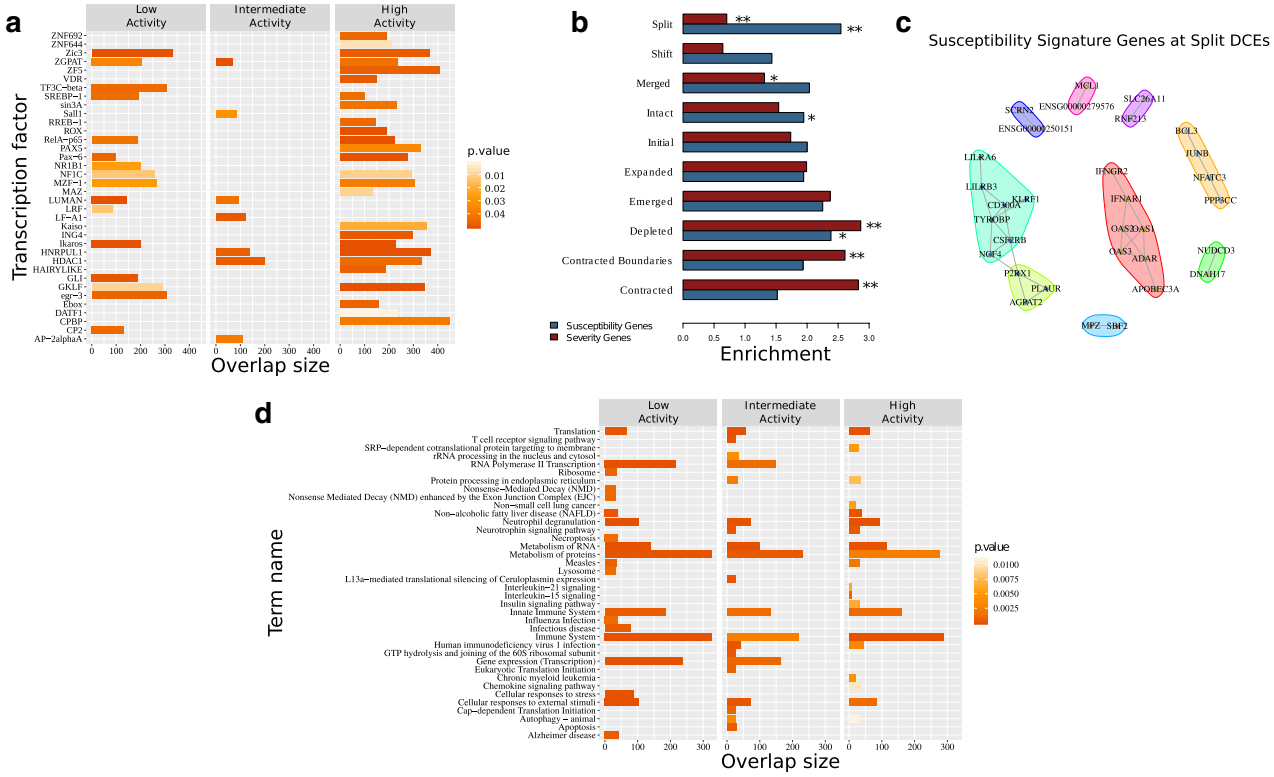

Figure 3. Functional analysis of the disruption events. (**a**) Enrichment analysis of 'Disruptors' in genes that are commonly regulated (suggested by the mutual regulatory motif matches—TRANSFAC database) by transcription factors indicated on y axis. The overlap between the query gene set and the corresponding Pathway members or TF-target genes are displayed on the x axis. The color of each bar illustrates the corrected *p* value of the corresponding enrichment test. (**b**) Average positional enrichments of susceptibility and severity genes[6] against different types of DCEs. Significance levels of one hundred permutations (*:0.05; **:0.01). (**c**) Protein interaction networks for susceptibility signature genes that are found to be differentially expressed and overlapping split DCE boundaries, as obtained from STRING-DB[35]. Genes are grouped on the basis of a modularity analysis. Modules are shown with coloured polygons around genes (red: interferon signature genes, cyan: DAP12 signaling, lime: neutrophil module, green: B-cell module). (**d**) Pathway enrichment analysis of genes which correspond to enhancer-TSS links (CD4+ cells—Enhancer Atlas)[37], that are nested in healthy group DCEs but disrupted in SLE. The top 20 most significant KEGG or/and REACTOME pathways are presented.

Discussion

Genome organization is intricately linked to gene expression and regulation in health and disease, with differentially expressed genes creating clusters under various conditions. Our study, the first such conducted in SLE, shows that genes are organized in extended domains of coordinated expression but, moreover, that these domains are highly dynamic and extensively reorganized during disease progression. While high activity patient patterns are suggestive of a general re-organization of gene regulation that extends to broader chromosomal domains, increased fragmentation of gene co-expression is observed even in the genomes of patients with low disease activity. This may suggest that the observed disruptive patterns of gene expression are be related to the way initial cellular signals propagate in the genome in order to affect hundreds of abnormally regulated genes. Thus, the more disconnected co-expression in low activity SLE genomes could be linked to mechanisms, with which flares occur even in patients that are in remission.

While, the governing principles of such mechanisms are yet to be resolved, our analyses suggest a key role for the chromatin environment. Differential enrichment of DCE patterns between open and closed chromatin and chromosomal compartments pertaining to early and late-replicating chromatin, are strong indications of epigenetic patterns underlying the fragmentation and re-organization of gene co-expression. Epigenetic effects, downstream of environmental triggers are expected to lie at the basis of SLE aetiopathogenesis, given the limited association of genetic factors reported for the disease. Further investigation of the mechanisms linking chromatin structure and the organization of gene expression in SLE could be assisted by our approach, through the prioritization of chromosomal domains with increased regulatory potential.

Besides epigenetic phenomena, the formation of co-expression domains could occur more transiently as the result of differential expression in any given setting[41], through the clustering of differentially expressed genes, that have been positionally constrained through evolution[20,42]. Such a notion is supported by our data in two ways. First through the association of the observed DCEs with functions that are known to be activated in SLE. Major pathways related to the intensity of the symptoms (such as the IFN signature) are associated with the disruption of co-expression, while downstream effects of SLE, related to the damage of organs (e.g. nephritis) are correlated

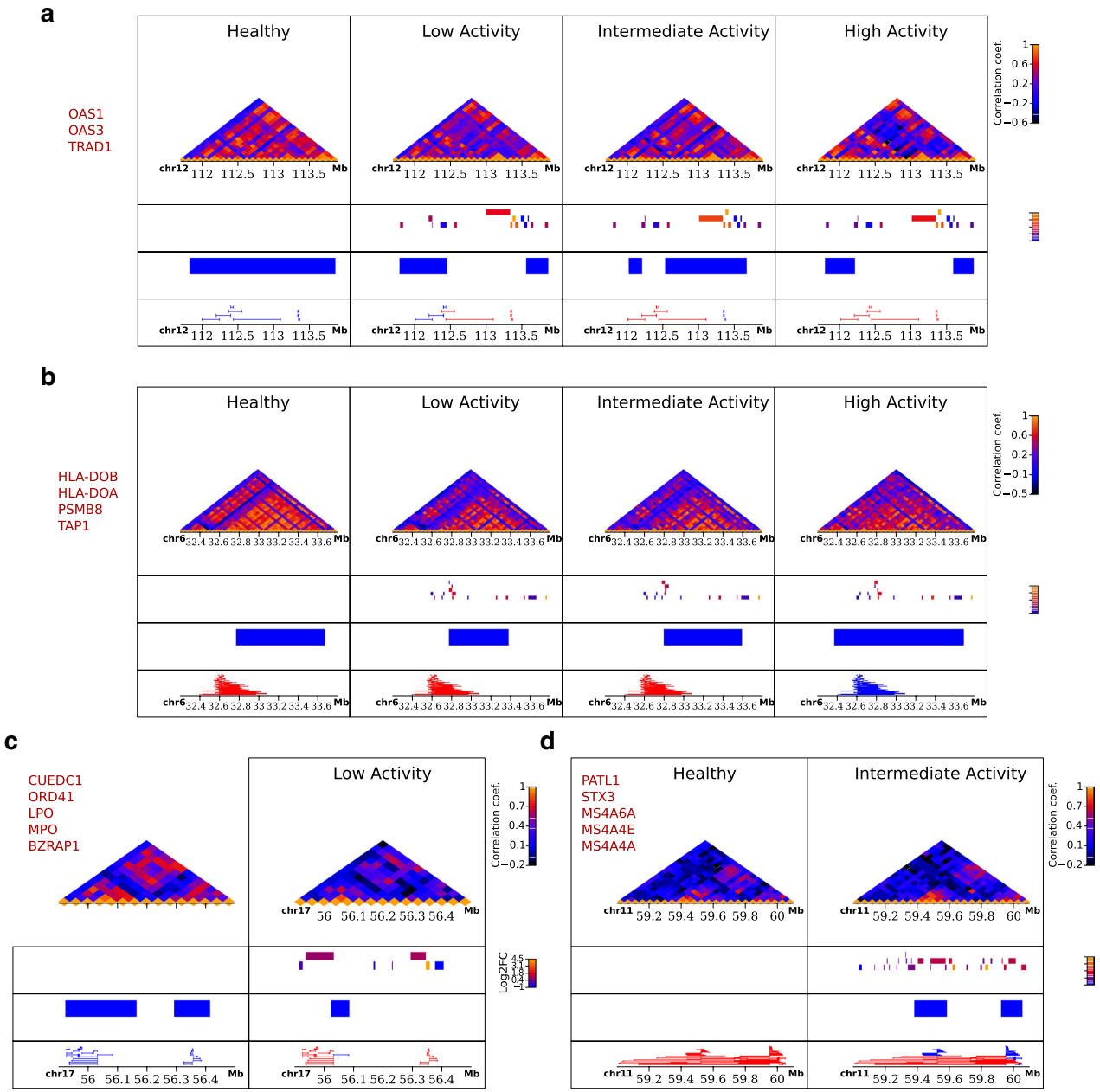

Figure 4. Examples of alterations in the co-expression profile. Heatmaps of expression correlation for selected loci of characteristic cases of disrupted (top), expanded (middle), deleted (bottom left) and emerged DCEs (bottom right). Heatmaps were created with the Sushi package from Bioconductor (https://bioconductor.org/packages/release/bioc/html/Sushi.html). Values in heatmaps correspond to bin signal, while the tracks below them show (from top to bottom) gene positions colour-coded for differential expression as log2(fold-change), DCE coordinates and enhancer-promoter associations that are entirely included in the same DCE (in blue) or not (in red). Names of differentially expressed genes in each locus are shown on the side of each panel.

with the general re-organization of co-expression in emergent domains. In addition gene signatures from both expression and genome-wide association data are enriched in various types of DCEs (Supplementary Figure 8), a strong indication that transcriptomic as well as genetic data may reveal a hidden layer of information when studied through the lens of genome organization.

The dynamics of co-expression clustering are also linked to differential expression, through the tendency of deregulated genes to occur in the boundaries of split DCEs. Inspection of the dynamics of DCE *splits* is, moreover, indicative of the general pattern of fragmentation and redistribution as is showcased in a number of examples where, compared to a contiguous DCE pattern in the healthy state, we observe splits in low disease activity and more generalized reorganization in high disease activity patients (Fig. 4). The fact that split/disrupted regions are more prominent in low disease activity genomes, combined with their proximity to genes belonging to the susceptibility signature, may come as an indication of an underlying hierarchy behind the gene deregulation

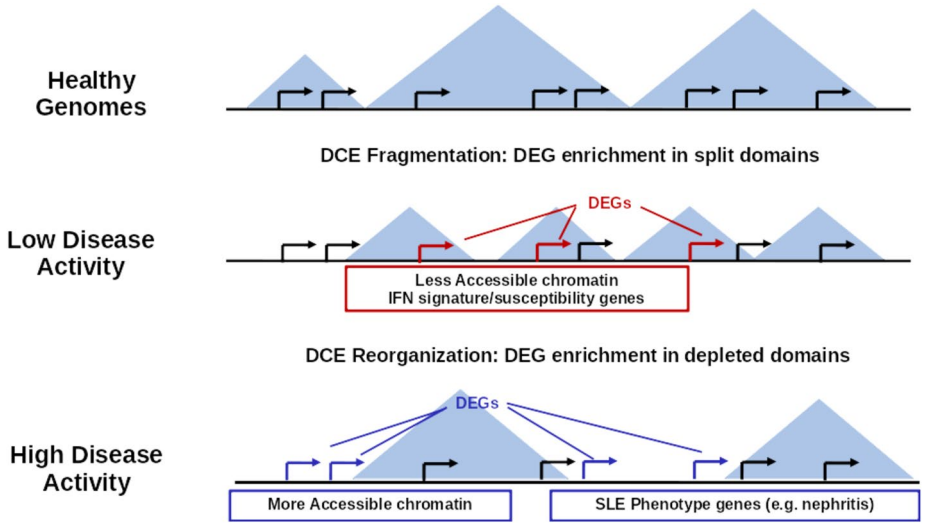

Figure 5. Patterns of gene co-expression in SLE. Graphical representation of the most prominent characteristics of gene co-expression patterns. Healthy genomes have extended domains of co-ordinated gene expression (DCE), but these become shorter and more fragmented in the genomes of patients with low disease activity. A significant number of differentially expressed genes (DEGs) in SLE low-activity genomes are associated with a disease "susceptibility" signature, located in split DCEs, linked to interferon and other signaling pathways and enriched in regions of low chromatin accessibility. In high disease activity genomes DCEs are re-distributed in new regions, where genes linked a SLE "severity" signature and more generalized manifestations of the disease (e.g. nephritis) localize in areas of DCE contraction, depletion and overall increased chromatin accessibility.

program. Indeed, we find enhancer-promoter associations of high relevance to be affected by the disrupted patterns of gene co-expression, which is strongly indicative of DCE splits having a possible multiplicative effect on gene regulation. Overall, our findings suggest that disruption of co-regulation patterns may represent a hallmark of the disease and that, moreover, that this process is constrained by epigenetic factors and the overall chromatin conformation (Fig. 5).

The approach we present here constitutes a first attempt to analyze gene expression at the level of genome organization in a complex disease and points to a number of interesting hypotheses linking the SLE phenotype with the underlying genome structure. Targeted conformation capture experiments on homogeneous cell cultures could be implemented in order to test these hypotheses. At the same time, the implementation of single-cell approaches at both transcriptome and genome conformation levels, could provide a data-rich framework for the application of our approach, with the final aim of obtaining cell-type specific co-expression profiles at increased resolution.

Data availability

Original RNASeq data[6] have been deposited at the European Genome-Phenome Archive (EGA) under the accession number EGAS00001003662. Processed data and original code for all presented analyses may be found at https://github.com/vntasis/SLE_spatial_gene_expression.

Received: 8 April 2020; Accepted: 15 September 2020
Published online: 06 October 2020

References

1. Tsokos, G. C. Systemic lupus erythematosus. *N. Engl. J. Med.* **365**, 2110–2121 (2011).
2. Tsokos, G. C., Lo, M. S., Reis, P. C. & Sullivan, K. E. New insights into the immunopathogenesis of systemic lupus erythematosus. *Nat. Rev. Rheumatol.* **12**, 716–730 (2016).
3. Barturen, G. & Alarcón-Riquelme, M. E. SLE redefined on the basis of molecular pathways. *Best Pract. Res. Clin. Rheumatol.* **31**, 291–305 (2017).
4. Baechler, E. C., Gregersen, P. K. & Behrens, T. W. The emerging role of interferon in human systemic lupus erythematosus. *Curr. Opin. Immunol.* **16**, 801–807 (2004).
5. Bennett, L. *et al.* Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J. Exp. Med.* **197**, 711–723 (2003).
6. Panousis, N. I. *et al.* Combined genetic and transcriptome analysis of patients with SLE: distinct, targetable signatures for susceptibility and severity. *Ann. Rheum. Dis.* **78**, 1079–1089 (2019).
7. Zhao, M. *et al.* Increased 5-hydroxymethylcytosine in CD4+ T cells in systemic lupus erythematosus. *J. Autoimmun.* **69**, 64–73 (2016).
8. Scharer, C. D. *et al.* ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Sci. Rep.* **6**, 27030 (2016).
9. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
10. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).
11. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).

12. Denas, O. *et al.* Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics* **16**, 87 (2015).
13. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science (80-)* **364**, eaat8266 (2019).
14. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
15. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485 (2011).
16. Rennie, S., Dalby, M., van Duin, L. & Andersson, R. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat. Commun.* **9**, 487 (2018).
17. Rada-Iglesias, A., Grosveld, F. G. & Papantonis, A. Forces driving the three-dimensional folding of eukaryotic genomes. *Mol. Syst. Biol.* **14**, e8214 (2018).
18. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.***12** (2015).
19. Quinodoz, S. A. *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744-757.e24 (2018).
20. Tsochatzidou, M., Malliarou, M., Papanikolaou, N., Roca, J. & Nikolaou, C. Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **45**, 5818–5828 (2017).
21. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31**, 180–183 (2002).
22. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* **10**, 1106–1113 (2008).
23. Coolen, M. W. *et al.* Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. *Nat. Cell Biol.* **12**, 235–246 (2010).
24. Letourneau, A. *et al.* Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature* **508**, 345–350 (2014).
25. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
26. Bombardier, C. *et al.* Derivation of the sledai: a disease activity index for lupus patients. *Arthr. Rheum.* **35**, 630–640 (1992).
27. Ran, D. & Daye, Z. J. Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Res.* **45**, e127 (2017).
28. Chaussabel, D. *et al.* A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150–164 (2008).
29. Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).
30. Zyla, J. *et al.* Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms. *Bioinformatics* **35**, 5146–5154 (2019).
31. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
32. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70–e70 (2016).
33. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. in *Methods in Molecular Biology,* vol. 1711. 243–259 (NIH Public Access, 2018).
34. Zaborowski, R. & Wilczyński, B. BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations. *J. Comput. Biol.* **26**, 305–314 (2019).
35. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
36. Toyabe, S. I., Kaneko, U. & Uchiyama, M. Decreased DAP12 expression in natural killer lymphocytes from patients with systemic lupus erythematosus is associated with increased transcript mutations. *J. Autoimmun.* **23**, 371–378 (2004).
37. Gao, T. *et al.* EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**, btw495 (2016).
38. Lowe, R. M., Genin, A., Orgun, N. & Cron, R. Q. IL-15 prolongs CD154 expression on human CD4 T cells via STAT5 binding to the CD154 transcriptional promoter. *Genes Immun.* **15**, 137–144 (2014).
39. Gensous, N., Schmitt, N., Richez, C., Ueno, H. & Blanco, P. T follicular helper cells, interleukin-21 and systemic lupus erythematosus. *Rheumatol. (United Kingdom)* **56**, 516–523 (2017).
40. Schwartz, D. M., Bonelli, M., Gadina, M. & O'Shea, J. J. Type I/II cytokines, JAKs, and new strategies for treating autoimmune diseases. *Nat. Rev. Rheumatol.* **12**, 25–36 (2016).
41. Ahlfors, H. *et al.* Gene expression dysregulation domains are not a specific feature of Down syndrome. *Nat. Commun.* **10**, 2489 (2019).
42. Ghanbarian, A. T. & Hurst, L. D. Neighboring genes show correlated evolution in gene expression. *Mol. Biol. Evol.* **32**, 1748–1766 (2015).

Acknowledgements

We would like to acknowledge the contribution of Halit Ongen, Irini Gergianaki, Maria Trachana, Luciana Romano-Palumbo, Deborah Bielser, Cedric Howald, Cristina Pamfil, Antonis Fanouriakis, Argyro Repa and Prodromos Sidiropoulos in patient screening/enrollment and data generation.

Author contributions

V.F.N. and C.N. conceived of the study, designed and performed analysis, wrote code and modeled data, N.I.P. and E.T.D. performed the initial differential expression analysis, M.G.T., D.T.B. and G.K.B. collected and processed original data, D.T.B., G.K.B., V.F.N. and C.N. wrote the paper, all authors read and approved the final manuscript.

Funding

This work was supported by a Fondation Santé Research Grant to Christoforos Nikolaou.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73654-4.

Correspondence and requests for materials should be addressed to C.N.

Reprints and permissions information is available at www.nature.com/reprints.

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