HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2018 October 16.

Published in final edited form as:

Nat Genet. 2018 May ; 50(5): 699–707. doi:10.1038/s41588-018-0102-3.

Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity

John B. Harley[1,2,3,4,5,*,^], **Xiaoting Chen**[1,^], **Mario Pujato**[1,^], **Daniel Miller**[1], **Avery Maddox**[1], **Carmy Forney**[1], **Albert F. Magnusen**[1], **Arthur Lynch**[1], **Kashish Chetal**[6], **Masashi Yukawa**[7], **Artem Barski**[4,7,8], **Nathan Salomonis**[4,6], **Kenneth M. Kaufman**[1,2,4,5], **Leah C. Kottyan**[1,4,*], and **Matthew T. Weirauch**[1,3,4,6,*]

[1]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[2]Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[3]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[4]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

[5]US Department of Veterans Affairs Medical Center, Cincinnati, Ohio, USA

[6]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[7]Division of Allergy & Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[8]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Abstract

Explaining the genetics of many diseases is challenging because most associations localize to incompletely characterized regulatory regions. We show that transcription factors (TFs) occupy multiple loci of individual complex genetic disorders using novel computational methods. Application to 213 phenotypes and 1,544 TF binding datasets identifies 2,264 relationships between hundreds of TFs and 94 phenotypes, including AR in prostate cancer and GATA3 in breast cancer. Strikingly, nearly half of the systemic lupus erythematosus risk loci are occupied by

[*]To whom correspondence should be addressed: Matthew.Weirauch@cchmc.org, John.Harley@cchmc.org, or Leah.Kottyan@cchmc.org.

[^]These authors contributed equally

Author contributions

The manuscript was written by J.B.H. and M.T.W., with critical feedback from L.C.K., K.M.K., N.S., A.B., X.C., M.P., D.M., and C.F. M.T.W., X.C., M.P., and J.B.H. designed, interpreted, and performed the main computational analyses. K.M.K., N.S., L.C.K., A.M., and K.C. designed, interpreted, and performed additional computational analyses. L.C.K., J.B.H., M.T.W., and A.B. designed and interpreted laboratory experiments. D.M., C.F., A.F.M., A.L., and M.Y. performed the laboratory experiments.

Competing Financial Interests Statement

J.B.H., M.T.W., and L.C.K. have a submitted patent application relating to these findings. A.B. is a co-founder of Datirium, LLC.

the Epstein-Barr virus EBNA2 protein and many co-clustering human TFs, revealing gene-environment interaction. Similar EBNA2-anchored associations exist in multiple sclerosis, rheumatoid arthritis, inflammatory bowel disease, type 1 diabetes, juvenile idiopathic arthritis, and celiac disease. Instances of allele-dependent DNA binding and downstream effects on gene expression at plausibly causal variants support genetic mechanisms dependent upon EBNA2. Our results nominate mechanisms that operate across risk loci within disease phenotypes, suggesting new paradigms for disease origins.

Introduction

The mechanisms generating genetic associations have proven difficult to elucidate for most diseases, since the vast majority of the pertinent variants are presumed to be components of a yet to be sufficiently understood regulome. Gene-environment interactions add another layer of complexity that may help explain the etiology of many autoimmune diseases[1–3]. In particular, Epstein-Barr virus (EBV) infection has been implicated in the autoimmune mechanisms and epidemiology of systemic lupus erythematosus (SLE)[4–7], increasing SLE risk by as much as 50-fold in children[4]. SLE patients also have elevated EBV loads in blood and early lytic viral gene expression[6]. Despite intriguing relationships between EBV and multiple autoimmune diseases, the underlying molecular mechanisms remain unknown[8,9].

Genome wide association studies (GWASs) have identified >50 convincing European ancestry SLE loci (Figure 1a), providing compelling evidence for germline DNA polymorphisms altering SLE risk[10–13]. Like most complex diseases, the great majority of SLE loci occur in likely gene regulatory regions[14,15]. We therefore asked if any of the DNA-interacting proteins encoded by EBV preferentially bind SLE risk loci. Our analyses reveal powerful associations with an EBV gene product (EBNA2), providing a potential origin of gene-environment interaction, along with a set of human transcription factors and co-factors (TFs) in SLE and six other autoimmune diseases. We present allele and EBV-dependent TF binding interactions and gene expression patterns that nominate cell types, molecular participants, and environmental contributions to disease mechanisms for these and 85 other diseases and physiological phenotypes.

Results

Intersection of disease risk loci with TF-DNA binding interactions

To identify TFs that bind a significant number of risk loci for a given disease, we developed the RELI (Regulatory Element Locus Intersection) algorithm. RELI systematically estimates the significance of the intersections of the genomic coordinates of plausibly causal genetic variants and DNA sequences immunoprecipitated (through ChIP-seq) by a particular TF. Observed intersection counts are compared to a null distribution composed of variant sets chosen to match the disease loci in terms of the allele frequency of the lead variant, the number of variants in the linkage disequilibrium (LD) block, and the LD block structure (Figure 2a and Supplementary Figure 1; see Online Methods). RELI is an extension of previous methods such as XGR[16], which estimates the overlap between an input set of

regions and genome-wide annotations, although XGR does not explicitly replicate LD block structure in the null model.

We first gauged the ability of RELI to capture known or suspected relationships between TFs and diseases. The androgen receptor (AR) plays a well-established role in prostate cancer[17], and RELI analysis revealed that AR binding sites in VCaP cells significantly intersect prostate cancer-associated loci (17 of 52 loci, Relative Risk (RR) = 3.7, Bonferroni corrected P-value (P_c)<10^{-6}, Table 1). Similarly, binding sites for GATA3 in MCF7 cells significantly intersect breast cancer variants (P_c<10^{-10}, Table 1), concordant with the established GATA3 disease role[18]. Consistent with EBV contributing to multiple sclerosis (MS)[19–22], RELI reveals that the EBV-encoded EBNA2 protein occupies 44 of the 109 MS loci in Mutu B cells (P_c<10^{-29}, Table 1). Prostate and breast cancer loci do not significantly intersect EBNA2 peaks, nor do the loci of certain inflammatory diseases such as systemic sclerosis (Table 1). Collectively, these observations illustrate that predictions made by RELI are specific and consistent with previously established disease mechanisms.

We assembled 53 European ancestry SLE loci (all with P<$5x10^{-8}$ in case-control studies) with risk allele frequencies >1%, in aggregate constituting 1,359 plausibly causal SLE variants (Supplementary Data Set 1). To explore the possible environmental contribution from EBV, we evaluated the ChIP-seq data from EBV-infected B cells for the EBV gene products EBNA1, EBNA2 (three datasets), EBNA3C, EBNA-LP, and Zta (Supplementary Data Set 2). EBNA2 occupies loci that significantly intersect SLE risk loci in all three available ChIP-seq datasets (Table 1). For example, variants present in 26 of 53 European SLE GWAS loci intersect EBNA2 ChIP-seq peaks from the Mutu B cell line, an almost 6-fold enrichment (P_c<10^{-24}). No association was detected for the other EBV-encoded proteins. To examine the possibility that these results might simply be explained by enrichment of SLE loci in B cell open chromatin regions, we restricted the RELI null model to variants located in DNase hypersensitive regions in EBV-infected B cells. With this higher stringency null model, all of the EBNA2 associations remained significant (Table 1). Thus, the associations we detect between SLE risk loci and EBNA2 cannot simply be explained by the previously established strong co-localization between SLE risk loci and B cell regulatory regions in the genome[23].

We next applied RELI to a large collection of human TF ChIP-seq datasets (1,544 experiments evaluating 344 TFs and 221 cell lines) (Supplementary Data Set 2). In total, 132 ChIP-seq datasets involving 60 unique TFs strongly intersect SLE loci (10^{-53}<P_c<10^{-6}). We chose a stringent corrected P-value cutoff of 10^{-6} based upon results from a simulation procedure aimed at estimating the false positive rate of our approach (see Online Methods). 109 (83%) of the significantly associated ChIP-seq datasets were performed in EBV-infected B cell lines, with impressive fidelity between datasets (Supplementary Data Set 3). Nearly identical results were obtained using a null model that also takes the distance to the nearest gene transcription start site into account (Supplementary Figure 2) and similar results were obtained using the null model employed by the GoShifter[24] method (Supplementary Figure 3). Similar results were also obtained with an expanded set of all 83 SLE risk loci published to date (regardless of ancestry)[10–13] or when separately examining SLE risk loci by ancestry (Supplementary Data Set 3). Strikingly, 20 of these 60 TFs participate in “EBV super-

enhancers", which enable proliferation and survival of EBV-infected B cells[25]. The human TFs in question largely bind the same loci occupied by EBNA2, comprising an optimal cluster of 28 SLE risk loci (Figure 1a).

If EBV is involved in SLE pathogenesis, then the absence of EBV, and hence EBNA2, should diminish the observed associations with SLE risk loci. For eight TFs, ChIP-seq datasets are available in both EBV-infected and EBV negative B cell lines (Supplementary Table 1). Notably, the four TFs with the strongest RELI P-values in EBV-infected B cells (BATF, IRF4, PAX5, and SPI1) have much weaker P-values in EBV negative B cells (Figure 1a bottom left panel, Supplementary Data Set 4), consistent with these TFs occupying many SLE risk loci only in the presence of EBV. Further, all of the datasets for the ten TFs with the strongest RELI P-values were performed in EBV-infected B cells, and none of the other cell types available for these TFs show significant association (Figure 1a, bottom right panel). For example, 22 ChIP-seq datasets are available in EBV-infected B cells for the NFκB subunit RELA. Of these, 20 significantly intersect with SLE risk loci ($10^{-53}<P_c<10^{-17}$), while none of the remaining 14 available RELA datasets in any other cell type have significant intersection. Previous studies have demonstrated that EBV activates the NFκB pathway, supporting the validity of this result[26–28]. Combined with the striking intersection between EBNA2 binding and SLE loci, these data strongly suggest an important role for EBV-infected, EBNA2-expressing B cells in SLE.

EBNA2-occupied genomic sites intersect autoimmune-associated loci

We applied RELI to 213 diseases and phenotypes obtained from the NHGRI GWAS catalog[29] and other sources (see Online Methods), revealing nine phenotypes displaying strong EBNA2 association in addition to SLE and MS: rheumatoid arthritis (RA), inflammatory bowel disease (IBD), type 1 diabetes (T1D), juvenile idiopathic arthritis (JIA), celiac disease (CelD), chronic lymphocytic leukemia (CLL), Kawasaki disease (KD), ulcerative colitis (UC), and immunoglobulin glycosylation (IgG) (Supplementary Data Set 3). We designate the seven disorders among these with particularly strong EBNA2 associations ($P_c<10^{-8}$) the "EBNA2 disorders." A recent study performed statistical fine-mapping of the variants for six of the seven EBNA2 disorders (IBD was not included)[30]. Of the resulting 1,953 candidate causal variants in that study, 130 overlap with EBNA2 ChIP-seq peaks in Mutu B cells (RR=8.7, $P_c<10^{-132}$). Notably, this represents the second-ranked ChIP-seq dataset out of the 1,544 considered in our study, trailing only POLR2A ChIP-seq performed in CD4+ T cells (Supplementary Data Set 3). Thus, the overlap between EBNA2 ChIP-seq peaks and loci associated with the EBNA2 disorders is even stronger when only considering statistically likely causal variants.

Consistent with the SLE results (Figure 1a), the same TFs tend to cluster with distinguishing loci for each disorder (Figure 1b–g, Supplementary Data Set 5). Further, there is a stronger association in EBV-infected than in EBV negative cells for many TFs, and the 10 most associated TFs consistently intersect more strongly in EBV-infected B cells than in other cell types (Figure 1b–g, Supplementary Data Set 5). Hierarchical clustering identifies a core set of 47 TFs binding to 142 risk loci across the seven EBNA2 disorders (Supplementary Figure

4). RBPJ, an established EBNA2 co-factor[31–33], has the most similar binding profile to EBNA2 across loci, as expected.

NFκB proteins RELA, RELB, REL, NFKB1, and NFKB2 comprise many of the strongest associations with EBNA2 disorder loci (Supplementary Data Set 3). We therefore collected the 348 loci associated with at least one of the EBNA2 disorders, and removed the 179 among these loci that contain at least one disease-associated variant located within a ChIP-seq peak for any NFκB protein in EBV-infected B cells. Among the remaining 169 loci, 19 still contain disease-associated variants falling within EBNA2 ChIP-seq peaks (2.15-fold enrichment, P=0.00012), indicating that many of these loci may be occupied by EBNA2 independent of NFκB involvement.

In order to identify candidate EBNA2 co-factors, we isolated EBNA2 disorder-associated variants located within EBNA2 ChIP-seq peaks and evaluated them using RELI. This analysis confirms the importance of RBPJ, followed by members of the basal transcriptional machinery (TBP and p300), and NFκB subunits (which are involved in EBNA2-mediated gene activation[34]) (Figure 2b). Interestingly, predicted EBNA2 co-factors vary with disease phenotype; for example, EBNA2 and EBNA3C are highly synergistic at the disease loci of three of the EBNA2 disorders (IBD, MS, and CelD), but rarely coincide at loci for the other four diseases (Supplementary Data Set 6).

The particular TFs tend to be shared across the EBNA2 disorders, but the loci they occupy are less frequently shared. No EBNA2-bound locus is associated with all seven EBNA2 disorders; most loci are unique to only one disorder (Figure 2c). Thus, the loci occupied by EBNA2 in each disorder are largely distinct from one another. One counterexample involves the *IKZF3* locus encoding the Aiolos TF, a key regulator in B lymphocyte activation[35], with genetic variants from five different EBNA2 disorders intersecting EBNA2 ChIP-seq peaks (Supplementary Figure 4).

If changes in gene regulation explain these results, then expression quantitative trait loci (eQTLs), ChIP-seq peaks for Pol-II, and histone marks associated with active gene regulatory regions should be relatively concentrated at the risk loci occupied by EBNA2. These predictions are indeed true for each of the seven EBNA2 disorders (Figure 2d and Supplementary Data Set 3). For example, <1% of all common variants in the human genome are eQTLs in EBV-infected B cell lines (Figure 2d). This value rises to 2.3% for common variants located within open chromatin in EBV-infected B cell lines, and rises further to 2.7% for common variants within EBNA2 ChIP-seq peaks (Figure 2d, upper left panel, bars labeled “Common variants”). Thus, there is a slight trend for a common variant located within an EBNA2 ChIP-seq peak to influence gene expression in EBV-infected B cell lines. Strikingly, this relationship is >10-fold increased for EBNA2 disorder-associated variants - 27.8% of EBNA2 disorder variants that are located within EBNA2 ChIP-seq peaks are also eQTLs, a value significantly greater than EBNA2 disorder variants located within open chromatin in EBV-infected B cell lines (20.5%, $P<10^{-5}$, Welch’s one-sided t-test) or EBNA2 disorder variants in general (10.4%, $P<10^{-8}$) (Figure 2d, upper left panel, bars labeled “EBNA2 disorder variants”). Similar trends hold for the other data types examined (Figure

2d). In aggregate, these results hint at the potential magnitude of the environmental influence of EBNA2 upon host gene expression within EBNA2 disorder loci in EBV-infected B cells.

EBNA2 participates in allele-dependent formation of transcription complexes at disease risk loci

The observed associations (Figure 1) are genetic if and only if they are driven by causal allele-dependent differences. Since EBNA2 imitates the binding of NOTCH to RBPJ[36], genetic variants at these loci could alter the binding of RBPJ (or another TF to which EBNA2 binds) or enable allele-dependent binding of a TF that requires the presence of EBNA2 by modulating the local chromatin environment (Figure 3a). Re-analysis of ChIP-seq data provides a means to identify allele-dependent protein binding events on a genome-wide scale - in cases where a given variant is heterozygous in the cell assayed, both alleles are available for the TF to bind, offering a natural control for one another since the only variable that has changed is the allele. We therefore developed the MARIO (Measurement of Allelic Ratios Informatics Operator) pipeline to identify allele-dependent protein binding by weighing imbalance between the number of sequencing reads for each allele of a given genetic variant, the total number of reads available at the variant, and the number and consistency of available experimental replicates (see Online Methods). MARIO is an easy-to-use, modular tool that extends existing methods[37–40] by (1) calculating a score that explicitly reflects reproducibility across experimental replicates; (2) reducing run-time via utilization of multiple computational cores; and (3) allowing the user to directly provide genotyping data as input. To identify heterozygotes for analysis, we genotyped five EBV-infected B cell lines with available ChIP-seq data and performed genome-wide imputation (see Online Methods). We applied MARIO and a related method, ABC[37], to a deeply sequenced (~190 million reads) GM12878 ATAC-seq dataset (GEO accession GSM1155957) and observed strong agreement between the 2,214 resulting scores (Spearman correlation of 0.98 ($P<10^{-15}$)).

We next applied MARIO to 271 ChIP-seq datasets performed in one of the five genotyped cell lines, altogether assessing 98 different molecules. Since EBNA2 binds DNA indirectly as a co-factor, we first asked if the variants displaying EBNA2 allele-dependent binding might coincide with similarly altered binding of other TFs. This analysis revealed strong concordance of allele-dependent binding events both within and across cell types. For example, we identified 68 heterozygous common variants located within allele-dependent EBNA2 GM12878 ChIP-seq peaks. EBF1, whose binding is globally influenced by EBNA2[36], has a coincident ChIP-seq peak favoring the same allele at 39 (57%) of these loci, as opposed to only 8 (11%) on the opposite allele ($P<10^{-4}$, binomial test, Figure 3b). Similar results were obtained when pairing EBNA2 binding in GM12878 with EBNA2 binding in Mutu cells, with established partners SPI1 and RBPJ, or with ATAC-seq chromatin accessibility data (Figure 3b). Analogous results are obtained for EBNA2 ChIP-seq data in Mutu and IB4 cell lines (Supplementary Figure 5). In total, MARIO confidently identified 21 variants associated with 15 different autoimmune diseases displaying allele-dependent EBNA2 binding in at least one cell type (Table 2, Supplementary Data Set 7). We note that the number of heterozygous autoimmune variants for which EBNA2 prefers one allele over the other is not significantly more than expected by chance (see Online Methods). We also

note that several variants might involve the *HLA* genes, and the current view is that coding alleles in the HLA class II in general are likely (though not certainly) causal for autoimmune diseases. Nevertheless, most of these variants also involve allele-dependent host protein binding, chromatin accessibility, or presence of histone marks such as H3K27ac (Supplementary Data Set 8). Together, these results suggest that many autoimmune-associated variants may act by modifying host gene regulatory programs via altered binding of EBNA2 and additional proteins.

To detect potential downstream effects of allele-dependent EBNA2 binding, we measured genome-wide gene expression levels by RNA-seq in Ramos, an EBV negative B cell line that can support an EBV infection. We confirmed the expected presence or absence of EBNA2 by sequencing (Online Methods) and western blot (Supplementary Figure 6). We identified a total of 80 genes with significant EBV-dependent alterations in gene expression (Supplementary Data Set 9), confirming that EBV modulates the expression of human genes. These results are highly consistent with a previous gene expression study and the literature (see Online Methods).

We next searched for autoimmune-associated variants that might impact EBNA2 binding, resulting in allele-dependent expression of a nearby gene. This analysis was dependent on the small subset of genetic variants satisfying four necessary criteria: the variant must be (1) plausibly causal for an autoimmune disorder; (2) immunoprecipitated by EBNA2 antibodies; (3) heterozygous in the cell line assayed; and (4) proximal to a plausible target mRNA that contains a heterozygous variant in Ramos cells (to detect allele-dependent expression). For example, the 21 EBNA2 variants listed in Table 2 satisfy the first three criteria, but only five satisfy the fourth criterion of being within 50kb of a potential target gene containing a heterozygous variant in the Ramos cell line.

Despite these limitations, our approach identified autoimmune-associated variants displaying allele-dependent EBNA2 binding and allele-dependent expression of a nearby gene. For example, rs3794102, a variant strongly associated with vitiligo ($P<10^{-9}$ for case/control association), has significantly skewed allele-dependent binding of eight proteins - EBNA2, its suspected co-factor EBF1[36], and chromatin accessibility all favor the non-reference ‘G’ vitiligo risk allele (Figure 3c, Table 2, Supplementary Data Set 8). Intriguingly, the proteins favoring the ‘G’ allele are considered activators, whereas the two proteins that prefer the ‘A’ allele are repressors, suggesting that the variant and virus might act synergistically as an allelic switch. rs3794102, which is located within an intron of *SLC1A2* (a gene for which we detect no RNA-seq reads), loops to the promoter of the neighboring *CD44* gene based on Hi-C experiments performed in GM12878 cells (Supplementary Figure 7). rs3794102 is also an established eQTL for *CD44* in EBV-infected B cell lines ($P<10^{-11}$, ‘MRCE’ dataset, RTeQTL database[41]), and particular isoforms of *CD44* are dependent on the presence of EBNA2[42]. In our experiments, *CD44* expression is 6.8-fold higher in EBV-infected Ramos cells compared to uninfected Ramos cells (P = 0.00015, Supplementary Data Set 9). Further, we identified a heterozygous genetic variant (rs8193) in strong LD with rs3794102 ($r^2=0.87$) located within the *CD44* gene body with 12 ‘T’ allele RNA-seq reads and only 5 ‘C’ allele reads in EBV-infected Ramos cells, and no detectable reads in Ramos cells lacking EBV (Supplementary Data Set 10). We

independently confirmed this result with allelic qPCR, observing a significant increase in expression for the T relative to the C allele in EBV-infected Ramos cells, with significantly lower levels of expression in the absence of EBV (Figure 3d). CD44 is a transmembrane glycoprotein involved in B cell migration and activation. Taken together, these results suggest that the 'G' vitiligo risk allele enhances formation of an EBNA2-dependent gene activation complex, resulting in elevated expression of *CD44*, and consequent increased B cell migration and/or activation. We also identified a variant (rs947474) associated with T1D and RA (Table 2) located near *PRKCQ*, another gene with allele- and EBV-dependent expression in our data (Supplementary Data Set 10). Intriguingly, PRKCQ plays an established role in activation of the EBV lytic cycle[43]. Together, these examples establish that multiple autoimmune variants may alter binding events of protein complexes containing EBNA2 and host proteins, resulting in EBV-controlled allele-dependent host gene expression.

Autoimmune-associated genetic mechanisms in EBV-infected B cells

We next used RELI to rank cell types by their relative importance to each of the EBNA2 disorders, based on the intersection between disease-associated variants and likely regulatory regions in that cell type. This procedure revealed a clear enrichment for EBV-infected B cells in SLE. For example, of the 175 H3K27ac ChIP-seq datasets available, the highest ranked 30 datasets are all from EBV-infected B cell lines (Figure 4a). Analogous results are obtained for "active chromatin marks" (a model based on combinations of various histone marks[44]) (Figure 4b), H3K4me3, and H3K4me1, for SLE and virtually all of the seven EBNA2 disorders (Supplementary Data Set 3, Supplementary Data Set 11). Collectively, these results support the EBV-infected B cell being an origin for genetic risk for each of the seven EBNA2 disorders. This analysis also reveals a likely involvement of other immune cell types in these disorders, including T cells, natural killer cells, and monocytes (Supplementary Data Set 3). Although there are limited TF ChIP-seq data available for these cell types, one or more of the EBNA2 disorders are associated with 17 of the available T cell TF ChIP-seq datasets (Supplementary Data Set 3). Further, several EBNA2 disorder loci appear to be specific to T cells. For example, six MS-associated loci are largely T cell-specific, collectively intersecting 67 T cell ChIP-seq datasets, compared to only 12 EBV-infected B cell datasets for these same loci (Supplementary Data Set 12). Together, these results are consistent with multiple shared regulatory mechanisms acting across autoimmune risk loci, some common between cell types (e.g., B and T cells) and others being exclusive to a certain cell type.

RELI identifies relationships between particular TFs and many diseases

Extension of RELI analysis to GWAS data for 213 phenotypes identified 2,264 significant ($P_C<10^{-6}$) TF-disease relationships (Supplementary Data Sets 1 and 3). In addition to the EBNA2-related associations, clustering of these results reveals a large grouping of hematopoietic phenotypes and well-established blood cell regulators such as GATA1 and TAL1 (Figure 4c). Other associations suggest additional mechanisms, many of which are supported by independent lines of evidence from other studies, such as GATA3, FOXA1, and TCF7L2 in breast cancer (Figure 4d), and AR, NR3C1, and EZH2 in prostate cancer (Supplementary Data Set 3). In total, application of these methods produces results

nominating global disease mechanisms for 94 different diseases or phenotypes (Supplementary Data Set 3), providing new directions for understanding their origins.

Discussion

Our efforts to understand the gene-environment interaction of SLE loci with EBV have revealed that EBNA2 and its associated human TFs occupy a significant fraction of autoimmune risk loci. In particular, NFκB subunits such as RELA, RELB, NFKB1, and NFKB2 also strongly intersect many of these loci, suggesting that NFκB is important in the mechanisms that confer risk in these inflammatory diseases. Further analyses suggest that multiple causal autoimmune variants may act through allele-dependent binding of these proteins, resulting in downstream alterations in gene expression. In this scenario, the relevant TFs and gene expression changes must occur in the cell type that alters disease risk. Collectively, our data identify the EBV-infected B cell as a possible site for gene action at select loci in multiple autoimmune diseases, with the caveat that existing data are biased, having been predominantly collected in this cell type.

Notably, four of the top 20 TFs that co-occupy EBNA2 disorder loci with EBNA2 can be targeted by at least one available drug (MED1, p300, NFKB1, and NFKB2)[45], and a recent study shows that the C-terminal domain of the BS69/ZMYND11 protein can bind to and inhibit EBNA2[46]. These results offer promise for the development of future therapies for manipulating the action of these proteins in individuals harboring risk alleles at EBNA2-bound loci.

Our current data nominate particular TFs and cell types for 94 phenotypes, providing mechanisms possibly explaining the molecular and cellular origins of disease risk for experimental verification and exploration. No doubt, as new genetic association and TF binding data are collected, approaches such as ours will reveal additional disease mechanisms.

Online Methods

Collection and processing of datasets

Phenotype-associated genetic variants were largely obtained from the NHGRI GWAS catalog[29]. This catalog does not contain candidate gene studies, including those from the widely-used ImmunoChip platform[60]. Thus, for SLE, MS, SSc, RA, and JIA, peer-reviewed literature was curated (Supplementary Data Set 13). Only genetic associations exceeding genome-wide significance ($P<5x10^{-8}$) were considered. Datasets were separated and annotated by ancestry, except where noted. Only phenotypes with five or more associated loci separated by at least 500 kb were considered, following Farh *et al.*[30]. Loci were anchored by the single most strongly associated variant and expanded to incorporate variants in strong linkage disequilibrium (LD) ($r^2>0.8$) using Plink[61], collectively constituting the *plausibly causal* variants. Final variant lists for each disease and phenotype are provided (Supplementary Data Set 1).

Functional genomics data were obtained from ENCODE[62] (downloaded on 4/14), Roadmap epigenomics[63] (6/15), Cistrome[64] (12/15), PAZAR[65] (4/14), ReMap-ChIP[66] (8/15), and Gene Expression Omnibus[67] (Supplementary Data Set 2). ChIP-seq datasets containing less than 500 peaks were removed. eQTLs were obtained from GTExPortal[49] (1/16), the Pritchard lab eQTL database (http://eqtl.uchicago.edu/) (4/14), and the Harvard eQTL database (https://www.hsph.harvard.edu/liming-liang/software/eqtl/) (4/14). TF binding motif models were obtained from Cis-BP (build 1.02)[68].

Regulatory Element Locus Intersection (RELI)

RELI takes a set of genetic variants as input, expands the set using LD blocks, and calculates the statistical intersection of the resulting loci with every dataset in a compendium (e.g., ChIP-seq datasets) (Figure 2a and Supplementary Figure 1). In Step 1, sequencing data from 1,000 Genomes[69] are used to identify all variants with linkage (r^2>0.8) to any input variant within each major ancestry (European, African, Asian), thereby assigning them to LD blocks. In Step 2, overlapping genomic coordinates determine whether an *observed intersection* is recorded between each LD block and each dataset. In Step 3, the *expected intersection* is estimated between each LD block and each dataset. The most strongly associated variant is chosen as the reference variant for the LD block. A distance vector is generated providing the distance (in bases) of each variant in the LD block from this reference variant. A random genomic variant with approximately matched allele frequencies to the reference variant is then selected from dbSNP[70], and genomic coordinates of *artificial variants* are created that are located at the same relative distances from this random variant using the distance vector. Members of this *artificial LD block* are intersected with each dataset, as was done for the observed intersections. This strategy accounts for the number of variants in the input LD block and their relative distances, while prohibiting 'double counting' due to multiple variants in the block intersecting the same dataset. We repeat this procedure 2,000 times, generating a null distribution with stable P-values. The expected intersection distributions are used to calculate Z-scores and P-values for the observed intersection. The final reported P-values are Bonferroni corrected (P_c) for the 1,544 TF datasets tested. We calculate the relative risk by dividing the *observed intersection* by the *mean expected intersection*. We also considered a higher-stringency null model that only considers variants located within DNase-seq peaks in any of the 22 available EBV-infected B cell line datasets, which controls for the known association of autoimmune variants and B cell regulatory regions[23].

We validated the RELI procedure as follows. First, we compared the Z-score-based P-values produced by RELI to empirically calculated P-values. We selected 187 ChIP-seq datasets with European SLE GWAS RELI corrected P-values that are evenly distributed between 1 and 10^{-7}. An upper bound of 10^{-7} was chosen due to the amount of time required to run the simulations. Overall, we observe very strong concordance between these 187 empirically-derived P-values and the P-values estimated by RELI (Supplementary Figure 8, Panel A), with a Pearson correlation coefficient of 0.82 ($P<10^{-45}$). We also performed 200,000,000 simulations examining the EBNA2 Mutu ChIP-seq vs. European SLE variant relationship. Across these simulations, we observed a maximum of 16 loci intersecting EBNA2 Mutu ChIP-seq peaks (Supplementary Figure 8, Panel B), conservatively setting an empirically-

determined P-value lower bound at 5 x 10^{-9} and further supporting our estimated P-value of $P_c<10^{-24}$ for the 26 observed locus intersections. To validate our choice of 2,000 simulations, we compared the P-values obtained for the 187 datasets when using 2,000 vs. 5,000,000 simulations. Nearly identical P-values were obtained (Supplementary Figure 8, Panel C).

We also estimated RELI false positive rates. We first generated a "false library" of 1,544 ChIP-seq datasets that match the "real library" by randomly repositioning each peak within the genome. This random "false library" of ChIP-seq results matches the number of datasets, the number of peaks each dataset contains, and the width of those peaks. Upon running the European SLE variants with RELI using 10 different "false libraries", only one of the 15,440 datasets achieved a P-value less than our $P_c<10^{-6}$ threshold (Supplementary Data Set 3). Further, the P-value for this dataset ($P_c<10^{-8}$) is much less significant than those for EBNA2, RELA, etc. ($P_c \ll 10^{-20}$). We thus estimate our overall false positive rate to be ~ 1/15,440 (~0.006%).

Identification of optimal clusters

We identified *optimal clusters* (red outlines in Figure 1) by comparing the observed number of TF/locus intersections to results from simulations. First, loci (X-axis) and TFs (Y-axis) were sorted in decreasing order of the number of intersections (colored boxes in the heatmap). We then iteratively considered every possible sub-matrix boundary, starting at the upper left corner. In each trial, the total number of intersections is kept fixed, but the locations of the intersecting positions are randomly permuted across loci. A Gaussian null distribution was obtained from 10,000 random trials. P-values were calculated for each sub-matrix by comparing the observed number of intersections within the sub-matrix to the null distribution, using a standard Z-score transformation. The optimal cluster was defined as the sub-matrix with the best P-value.

Cell line genotyping and imputation

We genotyped five EBV-infected B cell lines with available ChIP-seq data (Supplementary Table 2) on Illumina OMNI-5 arrays, as previously described[71]. Genotypes were called using the Gentrain2 algorithm within Illumina Genome Studio. Quality control was performed as previously described[71]. Quality control data cleaning was performed in the context of a larger batch of non-disease controls to allow for the assessment of data quality. Briefly, all cell lines had call rates >99%; only common variants (minor allele frequency >0.01) were included; and all variants were previously shown to be in Hardy-Weinberg equilibrium in control populations at P>0.0001[71]. We performed genome-wide imputation using overlapping 150 kb sections of the genome with IMPUTE2[72] and a composite imputation reference panel of pre-phased integrated haplotypes from 1,000 Genomes (June 2014). Included imputed genotypes met or exceeded a probability threshold of 0.9, an information measure of 0.5, and the same quality-control criteria described above for the genotyped markers.

Detection of allele-dependent sequencing reads using MARIO

We developed the MARIO (Measurement of Allelic Ratios Informatics Operator) pipeline to identify allele-dependent behavior at heterozygous genetic variants in functional genomics datasets. In brief, the pipeline downloads a set of reads, aligns them to the genome, calls peaks using MACS2 (parameters: --nomodel --extsize 147 -g hs -q 0.01), identifies allele-dependent behavior at heterozygotes within peaks (described below), and annotates the results (Supplementary Figure 9).

To estimate the significance of the degree of allelic imbalance of a given dataset at a given heterozygote, we developed the Allelic Reproducibility Score (ARS), which is based on a combination of two *predictive variables*: the total number of reads overlapping the variant and the imbalance between the number of reads for each allele. Other variables tested were uninformative (below). The ARS value also accounts for the number of available experimental replicates and the degree to which they agree. ARS values were calibrated using seven TFs with four replicate ChIP-seq experiments available in the same cell line (GM12878 or K562): SPI1 (set 1), SPI1 (set 2), NRSF, REST, RNF2, YY1 and ZBTB33.

ARS values were calculated as follows:

1. *Determine the number of reads mapping to each allele of each heterozygous variant in each replicate*. We applied our pipeline to each experimental replicate and counted the number of reads for each allele that overlap each heterozygous variant. Insertions and deletions were not considered. All duplicate reads were removed using the “MarkDuplicates” tool from the PICARD software package (https://broadinstitute.github.io/picard/). Before mapping reads using Bowtie2[73] (parameters -N 1 --np 0 --n-ceil 10 --no-unal), we masked all common variants in the GrCh37 (hg19) reference genome to N, which removes bias generated by reads carrying non-reference alleles. We designate the allele with the greater number of reads the *strong allele*, and the other the *weak allele* (Supplementary Figure 10a).

2. *Identify predictive variables of reproducible allele-dependent behavior across replicates.* We collected a set of seven datasets, {D}, with each dataset comprised of four experimental replicates, {R} (Supplementary Figure 10b). Each replicate contains a set of variants {V} that are heterozygous in the given cell type. For each of these variants, we calculated the value of four variables {X}: the ratio between the number of weak and strong allele reads, the total number of reads available at the variant, distance to peak center, and peak width.

   We evaluated the performance of each of these variables using a true-positive set of *reproducible variants.* This set was created by identifying all variants that share the same strong allele across all four replicates (Supplementary Figure 10c). Each variable was assessed based upon its ability to effectively separate reproducible variants from *non-reproducible variants* (all other variants). The reproducible variants are enriched for allele-dependent behavior, whereas the non-reproducible variants are depleted (Supplementary Figure 10d, left-most panel). Of the four variables tested, two were predictive of reproducible allele-

dependent binding: the ratio between the number of weak and strong allele reads (WS_ratio), and the total number of reads available at the variant (num_reads), which we designate the *predictive variables.*

3. *Determine a function mapping the values of the predictive variables to a single ARS value.* Our approach accommodates datasets containing any number of experimental replicates and rewards greater agreement between replicates. Within each of the seven datasets in the set {D}, we consider all possible combinations of one, two, or three replicates. Without loss of generality, we describe the procedure for the case of two replicates, which considers the subsets $\{R_1,R_2\}$, $\{R_1,R_3\}$, $\{R_1,R_4\}$, etc. We first identify the set {H} of reproducible variants (as described above) for each subset. We then threshold the WS_ratio into ranges, {(0 – 0.1), (0 – 0.2), (0 – 0.3), … (0 – 1)}, and for each range, we calculate the fraction of variants that are contained in the reproducible variant set as a function of num_reads (Supplementary Figure 11a). At this stage, this fraction still accounts for all variants, both allele-dependent and non-allele-dependent. We therefore adjust each curve by the normalized cumulative frequency of non-allele-dependent variants within the given range. For example, consider the WS_ratio=0.3 curve (Supplementary Figure 11a). Each point on this curve is divided by a single value representing the normalized cumulative frequency of the non-reproducible variants, which is obtained from the Y-axis at the X=0.3 position in the WS_ratio plot depicted in Supplementary Figure 10d. Before dividing, 1 is added to this value to avoid divide-by-zero errors. Collectively, this approach selectively penalizes non-allele-dependent behavior by accounting for the proportion of non-allele-dependent variants within each curve. These values were averaged across the seven datasets, yielding the final ARS values. This entire procedure is repeated for the cases of one, two, or three available replicates, generating the points shown in Supplementary Figure 11b. Curves were fit to these points using a saturating function:

$$ARS_w = \frac{A_w}{1 + B_w \times r} - A_w,$$

where w is the WS_ratio, r is num_reads, and A_w and B_w are the fitting parameters. The resulting functions yield ARS values for any given heterozygous variant in any dataset, as a function of the number of experimental replicates, the WS_ratio, and num_reads. When multiple replicates are available, we only report an ARS value for a variant if the strong allele is consistent in the majority of cases. A direct interpretation of the ARS values can be seen in the relationship between ARS values and the WS_ratio (Supplementary Figure 11c).

Statistical significance of the number of EBNA2 allele-dependent binding events

We observed a total of 21 cases of allele-dependent EBNA2 binding to an autoimmune risk variant (Table 2). To establish the statistical significance of this observation, we collected the full set of 42 autoimmune-disease associated variants that are (1) located within a ChIP-seq

peak in at least one of the three available EBNA2 datasets and (2) heterozygous in the cell type from which that peak was obtained. This set represents all autoimmune variants for which we could have observed allele-dependent EBNA2 binding. We next created a pool of non-autoimmune-associated variants that also meet the above two requirements (resulting in a total of 4,160 variants). For each of the 42 autoimmune variants, we chose a corresponding non-autoimmune variant from this pool, while approximately matching for the total number of EBNA2 ChIP-seq reads in the peak (within 10% of the read count). This procedure thus creates a matched set of 42 non-autoimmune variants that have an equal chance of resulting in allele-dependent EBNA2 behavior. There were a sufficient number of variants to repeat the above procedure 10 times, without replacement. In total, we observed 256 significant EBNA2 allele-dependent binding events across these matched non-autoimmune variant sets, which is not significantly different from the frequency that we observed with the autoimmune variants.

EBV Infection of Ramos cells

All cells were confirmed to be free of mycoplasma infection using PlasmaTest (InvivoGen, San Diego, CA). Wild-type EBV was prepared from supernatants of B95–8 cells cultured in RPMI medium 1640 supplemented with 10% FBS for two weeks. Briefly, the cells were pelleted and the virus suspension was filtered through 0.45 μM Millipore filters. The concentrated virus stocks were aliquoted and stored at -80ºC.

We infected ~2 x 10^6 Ramos Cells (ATCC CRL-1596) in the presence of growth medium containing 2μg/ml of phytohemagglutinin (PHA) for 4 hours. The infected cells were washed, cultured in growth media, and observed daily for multinuclear giant cell formation and morphological changes characteristic of EBV-infected B cells. After 10 passages, the infection was confirmed by measuring the expression of viral EBNA2 protein levels (Supplementary Figure 6).

RNA-seq

RNA was isolated from Ramos cell lines with and without EBV infection using the mirVANA Isolation Kit (Ambion). RNA sequencing targeting 150 million mappable, 125 base pair reads from paired-end, poly-A enriched libraries was performed at the CCHMC DNA Sequencing and Genotyping Core Facility. Sequencing reads were aligned to the GrCh37 (hg19) build of the human genome using TopHat[74] and Bowtie2[73] with Ensembl[75] RNA transcript annotations as a guide. In parallel, these data were aligned to the EBV genome (NCBI). As expected, 0 reads mapped in the EBV negative dataset, whereas 7,349 reads mapped in the EBV-infected dataset. 82.8% of the sequence reads aligned specifically to the human transcriptome, with a 2.6% increase in the aligned reads in the EBV negative samples. No abnormal quality control (QC) flags were identified following QC analysis with the software FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For allelic analysis, sequencing reads were aligned to the GrCh37 (hg19) build of the human genome using Hisat2[76]. Differential expression analysis was performed using Cufflinks[77].

As additional QC, we further compared our results to a study examining host gene expression changes to EBV infection in primary B cells[28]. Of the 80 genes whose expression

is significantly altered by the presence of EBV in our study, 18 of them are also significantly differentially expressed in this dataset. Further, among the 80 differentially expressed genes we detect, many of them represent classic host genes whose expression is modulated by EBV. Genes whose expression is concordantly activated by EBV include *CD44*[78], *TNFAIP2*[79], *MX1*[80], and *IFI44*[81]; genes with lowered expression include *VAV3*[82] and *CD99*[83].

Allelic qPCR

gDNA and RNA were extracted from Ramos cells with and without B95.8 EBV infection using the DNeasy Blood & Tissue Kit (Qiagen) and mirVana miRNA Isolation Kit (Invitrogen), respectively. RNA was treated with DNase using the TURBO DNA-free Kit (Ambion) and converted to cDNA using the High-Capacity RNA-to-cDNA Kit (Applied Biosystems). qPCR was performed with a single set of Taqman genotyping primers (Applied Biosystems) to rs8193 using the ABI 7500 PCR system. Fold change of expression was calculated with $2^{-\Delta\Delta CT}$ values, where cDNA was normalized to gDNA.

Statistical analyses

Details on statistical analyses are described in the corresponding sections. For statistical details on RELI and MARIO, see the corresponding sections above. The number of replicates or data points (N) is provided in the Figures and legends. Data are represented as means +/− one standard deviation, unless otherwise noted.

Data availability

RNA-seq data are available in the Gene Expression Omnibus (GEO) database under accession number GSE93709. Full datasets and results, including disease variants (with alleles) and all RELI and MARIO output, are provided in the Supplementary Material.

Code availability

The RELI and MARIO source code, with full documentation and examples, are freely available under the GNU General Public License on the Weirauch Lab GitHub page: https://github.com/WeirauchLab/

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank: J. Lee, C. Schroeder, Y. Huang, X. Lu, Z. Patel, E. Zoller, and The CCHMC DNA Sequencing and Genotyping Core for experimental support; C. Gunawan, K. Ernst, and T. Hong for analytical support; B. Cobb for administrative support; R. Kopan, C. Karp, W. Miller, J. Whitsett, M. Fisher, A. Strauss, S. Hamlin, L. Muglia, H. Singh, J. Oksenberg, I. Chepelev, S. Waggoner, S. Thompson, and H. Moncrieffe for constructive feedback and guidance; Y. Yuan (University of Penn) and D. Thorley-Lawson (Tufts Institute) for generous donation of cell lines (Mutu and IB4, respectively). Thanks to all of our colleagues who have made their data available to us, without which this project and its results would not have been possible. Funding sources: National Institutes of Health (NIH) R01 NS099068, NIH R21 HG008186, Lupus Research Alliance “Novel Approaches”, CCRF Endowed Scholar, CCHMC CpG Pilot study award, and CCHMC Trustee Awards to M.T.W.; NIH R01 AI024717, NIH U01 HG008666, NIH U01 AI130830, NIH P30 AR070549, NIH R24 HL105333, NIH KL2 TR001426, NIH R01

AI031584, NIH P30 DK078392, Kirkland Scholar Award and US Department of Veterans Affairs I01 BX001834 to J.B.H.; NIH R01 DK107502 to L.C.K; NIH DP2 GM119134 to A.B.

References

1. Fujinami RS, von Herrath MG, Christen U, Whitton JL. Molecular mimicry, bystander activation, or viral persistence: infections and autoimmune disease. Clin Microbiol Rev. 2006; 19:80–94. DOI: 10.1128/CMR.19.1.80-94.2006 [PubMed: 16418524]
2. Ercolini AM, Miller SD. The role of infections in autoimmune disease. Clinical and experimental immunology. 2009; 155:1–15. DOI: 10.1111/j.1365-2249.2008.03834.x [PubMed: 19076824]
3. Sener AG, Afsar I. Infection and autoimmune disease. Rheumatol Int. 2012; 32:3331–3338. DOI: 10.1007/s00296-012-2451-z [PubMed: 22811010]
4. James JA, et al. An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. J Clin Invest. 1997; 100:3019–3026. DOI: 10.1172/JCI119856 [PubMed: 9399948]
5. Hanlon P, Avenell A, Aucott L, Vickers MA. Systematic review and meta-analysis of the sero-epidemiological association between Epstein-Barr virus and systemic lupus erythematosus. Arthritis research & therapy. 2014; 16:R3. [PubMed: 24387619]
6. McClain MT, et al. Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry. Nat Med. 2005; 11:85–89. DOI: 10.1038/nm1167 [PubMed: 15619631]
7. Harley JB, James JA. Epstein-Barr virus infection induces lupus autoimmunity. Bulletin of the NYU hospital for joint diseases. 2006; 64:45–50. [PubMed: 17121489]
8. Ascherio A, Munger KL. EBV and Autoimmunity. Curr Top Microbiol Immunol. 2015; 390:365–385. DOI: 10.1007/978-3-319-22822-8_15 [PubMed: 26424654]
9. Draborg AH, Duus K, Houen G. Epstein-Barr virus in systemic autoimmune diseases. Clinical & developmental immunology. 2013; 2013:535738. [PubMed: 24062777]
10. Vaughn SE, Kottyan LC, Munroe ME, Harley JB. Genetic susceptibility to lupus: the biological basis of genetic risk found in B cell signaling pathways. Journal of leukocyte biology. 2012; 92:577–591. DOI: 10.1189/jlb.0212095 [PubMed: 22753952]
11. Alarcon-Riquelme ME, et al. Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. Arthritis Rheumatol. 2016; 68:932–943. DOI: 10.1002/art.39504 [PubMed: 26606652]
12. Bentham J, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015; 47:1457–1464. DOI: 10.1038/ng.3434 [PubMed: 26502338]
13. Sun C, et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. Nat Genet. 2016; 48:323–330. DOI: 10.1038/ng.3496 [PubMed: 26808113]
14. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. DOI: 10.1126/science.1222794 [PubMed: 22955828]
15. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. DOI: 10.1073/pnas.0903103106 [PubMed: 19474294]
16. Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med. 2016; 8:129. [PubMed: 27964755]
17. Schweizer MT, Yu EY. Persistent androgen receptor addiction in castration-resistant prostate cancer. J Hematol Oncol. 2015; 8:128. [PubMed: 26566796]
18. Asch-Kendrick R, Cimino-Mathews A. The role of GATA3 in breast carcinomas: a review. Hum Pathol. 2016; 48:37–47. DOI: 10.1016/j.humpath.2015.09.035 [PubMed: 26772397]
19. Almohmeed YH, Avenell A, Aucott L, Vickers MA. Systematic review and meta-analysis of the sero-epidemiological association between Epstein Barr virus and multiple sclerosis. PLoS One. 2013; 8:e61110. [PubMed: 23585874]

20. Pender MP, Burrows SR. Epstein-Barr virus and multiple sclerosis: potential opportunities for immunotherapy. Clinical & translational immunology. 2014; 3:e27.
21. Marquez AC, Horwitz MS. The Role of Latently Infected B Cells in CNS Autoimmunity. Front Immunol. 2015; 6:544. [PubMed: 26579121]
22. Ricigliano VA, et al. EBNA2 binds to genomic intervals associated with multiple sclerosis and overlaps with vitamin D receptor occupancy. PLoS One. 2015; 10:e0119605. [PubMed: 25853421]
23. Hu X, et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. American journal of human genetics. 2011; 89:496–506. DOI: 10.1016/j.ajhg.2011.09.002 [PubMed: 21963258]
24. Trynka G, et al. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. American journal of human genetics. 2015; 97:139–152. DOI: 10.1016/j.ajhg.2015.05.016 [PubMed: 26140449]
25. Zhou H, et al. Epstein-Barr virus oncoprotein super-enhancers control B cell growth. Cell host & microbe. 2015; 17:205–216. DOI: 10.1016/j.chom.2014.12.013 [PubMed: 25639793]
26. Gewurz BE, et al. Canonical NF-kappaB activation is essential for Epstein-Barr virus latent membrane protein 1 TES2/CTAR2 gene regulation. J Virol. 2011; 85:6764–6773. DOI: 10.1128/JVI.00422-11 [PubMed: 21543491]
27. Ersing I, Bernhardt K, Gewurz BE. NF-kappaB and IRF7 pathway activation by Epstein-Barr virus Latent Membrane Protein 1. Viruses. 2013; 5:1587–1606. DOI: 10.3390/v5061587 [PubMed: 23793113]
28. Price AM, et al. Analysis of Epstein-Barr virus-regulated host gene expression changes through primary B-cell outgrowth reveals delayed kinetics of latent membrane protein 1-mediated NF-kappaB activation. J Virol. 2012; 86:11096–11106. DOI: 10.1128/JVI.01069-12 [PubMed: 22855490]
29. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–1006. DOI: 10.1093/nar/gkt1229 [PubMed: 24316577]
30. Farh KK, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015; 518:337–343. DOI: 10.1038/nature13835 [PubMed: 25363779]
31. Zimber-Strobl U, et al. Epstein-Barr virus nuclear antigen 2 exerts its transactivating function through interaction with recombination signal binding protein RBP-J kappa, the homologue of Drosophila Suppressor of Hairless. EMBO J. 1994; 13:4973–4982. [PubMed: 7957063]
32. Grossman SR, Johannsen E, Tong X, Yalamanchili R, Kieff E. The Epstein-Barr virus nuclear antigen 2 transactivator is directed to response elements by the J kappa recombination signal binding protein. Proc Natl Acad Sci U S A. 1994; 91:7568–7572. [PubMed: 8052621]
33. Henkel T, Ling PD, Hayward SD, Peterson MG. Mediation of Epstein-Barr virus EBNA2 transactivation by recombination signal-binding protein J kappa. Science. 1994; 265:92–95. [PubMed: 8016657]
34. Scala G, et al. Epstein-Barr virus nuclear antigen 2 transactivates the long terminal repeat of human immunodeficiency virus type 1. J Virol. 1993; 67:2853–2861. [PubMed: 8386279]
35. Wang JH, et al. Aiolos regulates B cell activation and maturation to effector state. Immunity. 1998; 9:543–553. [PubMed: 9806640]
36. Lu F, et al. EBNA2 Drives Formation of New Chromosome Binding Sites and Target Genes for B-Cell Master Regulatory Transcription Factors RBP-jkappa and EBF1. PLoS Pathog. 2016; 12:e1005339. [PubMed: 26752713]
37. Bailey SD, Virtanen C, Haibe-Kains B, Lupien M. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. Bioinformatics. 2015; 31:3057–3059. DOI: 10.1093/bioinformatics/btv321 [PubMed: 25995231]
38. Buchkovich ML, et al. Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. BMC medical genomics. 2015; 8:43. [PubMed: 26210163]
39. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2016; 48:206–213. DOI: 10.1038/ng.3467 [PubMed: 26656845]

40. Shi W, Fornes O, Mathelier A, Wasserman WW. Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Res. 2016; 44:10106–10116. DOI: 10.1093/nar/gkw691 [PubMed: 27492288]
41. Ma B, Huang J, Liang L. RTeQTL: Real-Time Online Engine for Expression Quantitative Trait Loci Analyses. Database : the journal of biological databases and curation. 2014; 2014
42. Kryworuckho M, Diaz-Mitoma F, Kumar A. CD44 isoforms containing exons V6 and V7 are differentially expressed on mitogenically stimulated normal and Epstein-Barr virus-transformed human B cells. Immunology. 1995; 86:41–48. [PubMed: 7590880]
43. Gonnella R, et al. PKC theta and p38 MAPK activate the EBV lytic cycle through autophagy induction. Biochim Biophys Acta. 2015; 1853:1586–1595. DOI: 10.1016/j.bbamcr.2015.03.011 [PubMed: 25827954]
44. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. DOI: 10.1038/nature09906 [PubMed: 21441907]
45. Griffith M, et al. DGIdb: mining the druggable genome. Nature methods. 2013; 10:1209–1210. DOI: 10.1038/nmeth.2689 [PubMed: 24122041]
46. Harter MR, et al. BS69/ZMYND11 C-Terminal Domains Bind and Inhibit EBNA2. PLoS Pathog. 2016; 12:e1005414. [PubMed: 26845565]
47. Li Y, et al. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjogren's syndrome at 7q11.23. Nat Genet. 2013; 45:1361–1365. DOI: 10.1038/ng.2779 [PubMed: 24097066]
48. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014; 506:376–381. DOI: 10.1038/nature12873 [PubMed: 24390342]
49. Consortium GTHuman genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. DOI: 10.1126/science.1262110 [PubMed: 25954001]
50. Mifsud B, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015; 47:598–606. DOI: 10.1038/ng.3286 [PubMed: 25938943]
51. Javierre BM, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016; 167:1369–1384e1319. DOI: 10.1016/j.cell.2016.09.037 [PubMed: 27863249]
52. Liang L, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res. 2013; 23:716–726. DOI: 10.1101/gr.142521.112 [PubMed: 23345460]
53. Stranger BE, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–1224. DOI: 10.1038/ng2142 [PubMed: 17873874]
54. Veyrieras JB, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4:e1000214. [PubMed: 18846210]
55. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. DOI: 10.1038/nature08872 [PubMed: 20220758]
56. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. DOI: 10.1038/nature08903 [PubMed: 20220756]
57. Mangravite LM, et al. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. Nature. 2013; 502:377–380. DOI: 10.1038/nature12508 [PubMed: 23995691]
58. Dimas AS, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009; 325:1246–1250. DOI: 10.1126/science.1174148 [PubMed: 19644074]
59. Gaffney DJ, et al. Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. 2012; 13:R7. [PubMed: 22293038]
60. Trynka G, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011; 43:1193–1201. DOI: 10.1038/ng.998 [PubMed: 22057235]
61. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007; 81:559–575. DOI: 10.1086/519795 [PubMed: 17701901]

62. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. DOI: 10.1038/nature11247 [PubMed: 22955616]
63. Roadmap Epigenomics C, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. DOI: 10.1038/nature14248 [PubMed: 25693563]
64. Liu T, et al. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 2011; 12:R83. [PubMed: 21859476]
65. Portales-Casamar E, et al. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. Nucleic Acids Res. 2009; 37:D54–60. DOI: 10.1093/nar/gkn783 [PubMed: 18971253]
66. Griffon A, et al. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. 2015; 43:e27. [PubMed: 25477382]
67. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013; 41:D991–995. DOI: 10.1093/nar/gks1193 [PubMed: 23193258]
68. Weirauch MT, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158:1431–1443. DOI: 10.1016/j.cell.2014.08.009 [PubMed: 25215497]
69. Genomes Project C, et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. DOI: 10.1038/nature15393 [PubMed: 26432245]
70. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000; 28:352–355. [PubMed: 10592272]
71. Kottyan LC, et al. Genome-wide association analysis of eosinophilic esophagitis provides insight into the tissue specificity of this allergic disease. Nat Genet. 2014; 46:895–900. DOI: 10.1038/ng.3033 [PubMed: 25017104]
72. Verma SS, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Frontiers in genetics. 2014; 5:370. [PubMed: 25566314]
73. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9:357–359. DOI: 10.1038/nmeth.1923 [PubMed: 22388286]
74. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. DOI: 10.1093/bioinformatics/btp120 [PubMed: 19289445]
75. Flicek P, et al. Ensembl 2013. Nucleic Acids Res. 2013; 41:D48–55. DOI: 10.1093/nar/gks1236 [PubMed: 23203987]
76. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature methods. 2015; 12:357–360. DOI: 10.1038/nmeth.3317 [PubMed: 25751142]
77. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. DOI: 10.1038/nbt.1621 [PubMed: 20436464]
78. Birkenbach M, Josefsen K, Yalamanchili R, Lenoir G, Kieff E. Epstein-Barr virus-induced genes: first lymphocyte-specific G protein-coupled peptide receptors. J Virol. 1993; 67:2209–2220. [PubMed: 8383238]
79. Chen CC, et al. NF-kappaB-mediated transcriptional upregulation of TNFAIP2 by the Epstein-Barr virus oncoprotein, LMP1, promotes cell motility in nasopharyngeal carcinoma. Oncogene. 2014; 33:3648–3659. DOI: 10.1038/onc.2013.345 [PubMed: 23975427]
80. Craig FE, et al. Gene expression profiling of Epstein-Barr virus-positive and -negative monomorphic B-cell posttransplant lymphoproliferative disorders. Diagn Mol Pathol. 2007; 16:158–168. DOI: 10.1097/PDM.0b013e31804f54a9 [PubMed: 17721324]
81. Smith N, et al. Induction of interferon-stimulated genes on the IL-4 response axis by Epstein-Barr virus infected human b cells; relevance to cellular transformation. PLoS One. 2013; 8:e64868. [PubMed: 23724103]
82. Portis T, Dyck P, Longnecker R. Epstein-Barr Virus (EBV) LMP2A induces alterations in gene transcription similar to those observed in Reed-Sternberg cells of Hodgkin lymphoma. Blood. 2003; 102:4166–4178. DOI: 10.1182/blood-2003-04-1018 [PubMed: 12907455]
83. Lee IS, Shin YK, Chung DH, Park SH. LMP1-induced downregulation of CD99 molecules in Hodgkin and Reed-Sternberg cells. Leuk Lymphoma. 2001; 42:587–594. DOI: 10.3109/10428190109099318 [PubMed: 11697486]

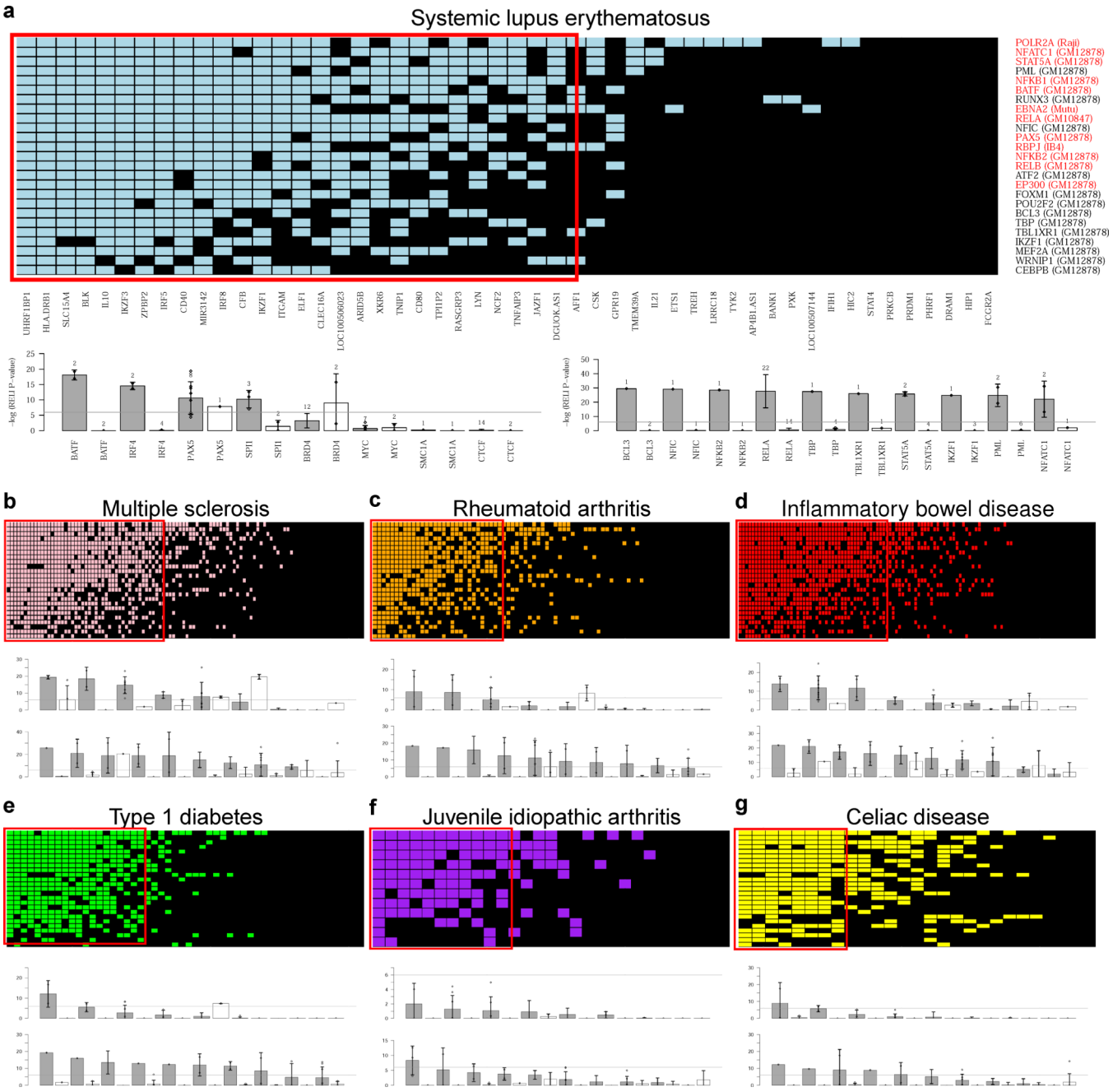

Figure 1. Intersection between autoimmune risk loci and TF binding interactions with the genome

a. Results for SLE risk loci. X-axis displays SLE-associated loci. Y-axis displays the top 25 TFs, based on RELI P-values, sorted by the number of loci. A colored box indicates that the given locus contains at least one SLE-associated variant located within a ChIP-seq peak for the given TF. The most significant ChIP-seq dataset cell type is indicated in parentheses. TFs that participate in "EBNA2 super-enhancers"[25] are colored red. The red rectangle identifies those loci and TFs that optimally cluster together (see Online Methods). Bottom panel, left: comparison of EBV-infected B cell lines (grey bars) to EBV negative B cells (white bars).

The Y-axis shows the distribution of the RELI –log (P_Cs) for each of the eight TFs with available data. Bars indicate mean. Error bars indicate standard deviation. Numbers indicate number of datasets. Horizontal line indicates the $P_C<10^{-6}$ RELI significance threshold. Bottom panel, right: The top 10 TFs (based on RELI P_C-values) with data available in at least one EBV-infected B cell line (grey bars) and at least one other cell type (white bars). **b–g. Results for the other six EBNA2 disorders**. Full results are available in Supplementary Data Set 5.

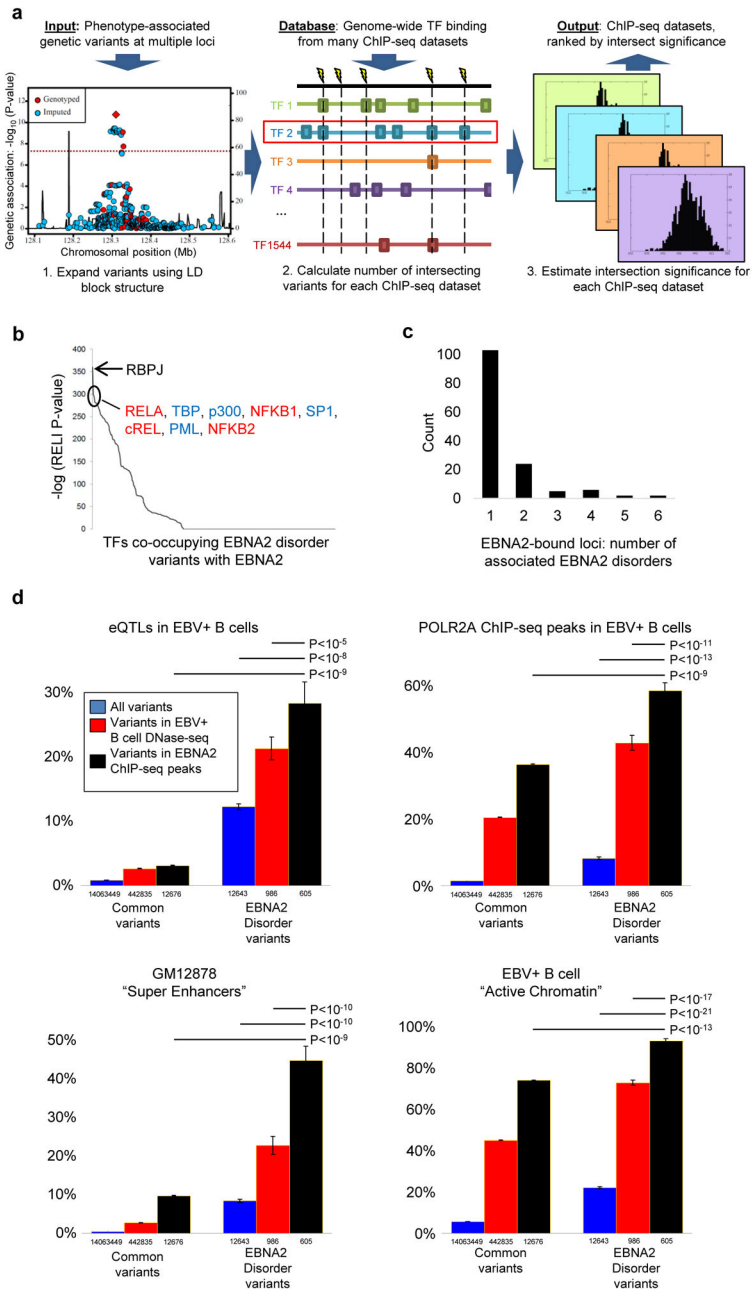

Figure 2. Properties of EBNA2-bound autoimmune disease loci

a. Schematic of the RELI algorithm. See Online Methods for details. **b. TFs intersecting autoimmune risk loci occupied by EBNA2.** RELI was re-executed using EBNA2 disorder variants intersecting EBNA2 ChIP-seq peaks as input. Top TFs are indicated. NFκB subunits are shown in red. Basal transcriptional machinery proteins are shown in blue. **c. Most EBNA2-occupied loci are associated with only a single EBNA2 disorder.** EBNA2-bound loci were categorized by the number of EBNA2 disorders with which the given locus is associated (X-axis). **d. Functional properties of EBNA2 disorder EBNA2-occupied loci.** Functional importance of EBNA2-occupied loci, assessed with four criteria. In each panel,

variants are segregated into two categories – common variants (left bars) and common variants associated with at least one EBNA2 disorder (right bars). Each category is divided into three types of variants (see key). The Y-axis of each plot indicates the percent of variants in each group that are, for example, eQTLs in EBV-infected B cells (top left plot). Error bars indicate the standard deviation obtained from sampling (with replacement) of 50% of the variants. Values below indicate number of variants. Horizontal bars at the top indicate sampling-derived P-values based on Welch's one-sided t-test.

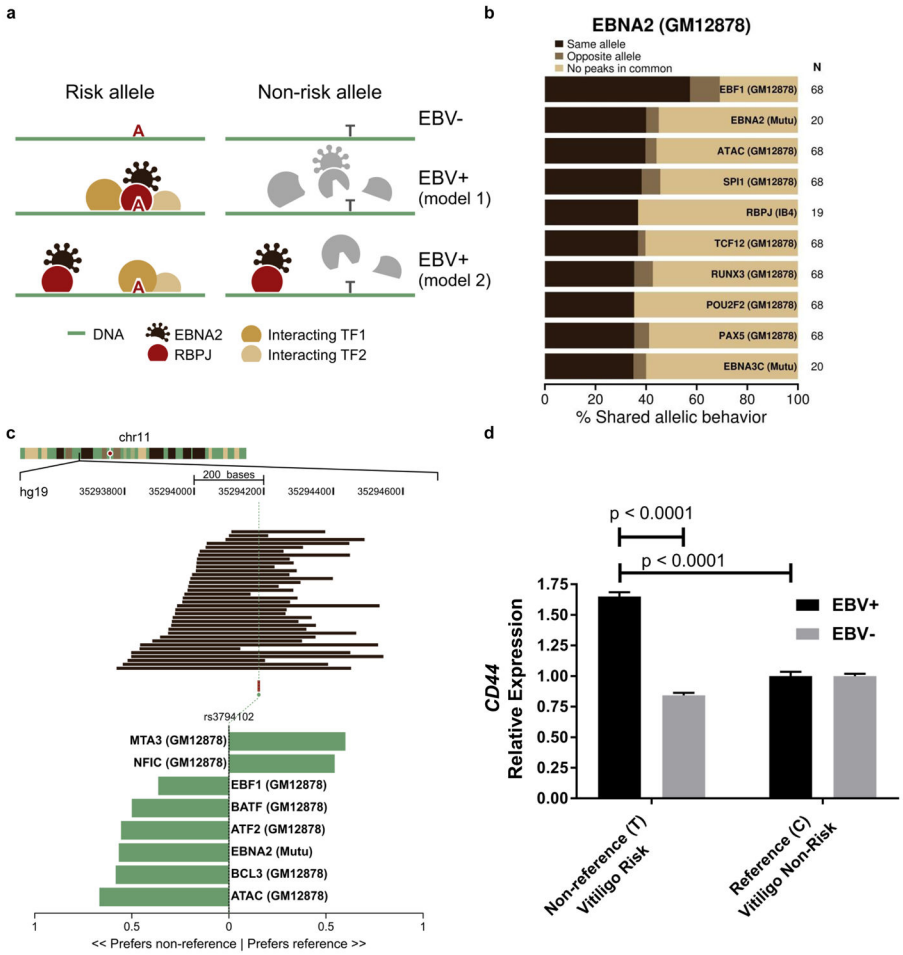

Figure 3. Allele-dependent binding of EBNA2 to autoimmune-associated genetic variants

a. Theoretical models presenting possible allele-dependent action of EBNA2. See text for discussion. **b. Allele-dependent co-binding of EBNA2 with multiple proteins.** ChIP-seq datasets from EBV-infected B cell lines were examined for evidence of allele-dependent binding at heterozygotes. Datasets are sorted by the proportion of EBNA2 GM12878 allele-dependent events (MARIO ARS value > 0.40, see Online Methods) that favor the same allele (X-axis). Values (N) indicate total number of variants. **c. Allele-dependent binding of EBNA2 and human proteins at the *CD44* locus.** Top to bottom: chromosomal band (multi-colored bar), location of EBV-infected B cell line ChIP-seq peaks for various TFs, location of rs3794102 variant, allele-dependent binding events (green bars). X-axis indicates the preferred allele, along with a value indicating the strength of the allelic behavior, calculated as one minus the ratio of the weak to strong reads (e.g., 0.5 indicates the strong allele has approximately twice the reads of the weak allele). **d. Allele and EBV-dependent expression of *CD44*.** Allelic qPCR of *CD44* expression in EBV-infected and EBV negative Ramos B cells (see key). Fold-change in expression is provided relative to the C allele. Error bars represent standard deviation (n=12: three independent experiments of technical quadruplicates). P-values were calculated using a two-way ANOVA with a Tukey post-hoc test. EBV status and variant genotype were used as the two factors.

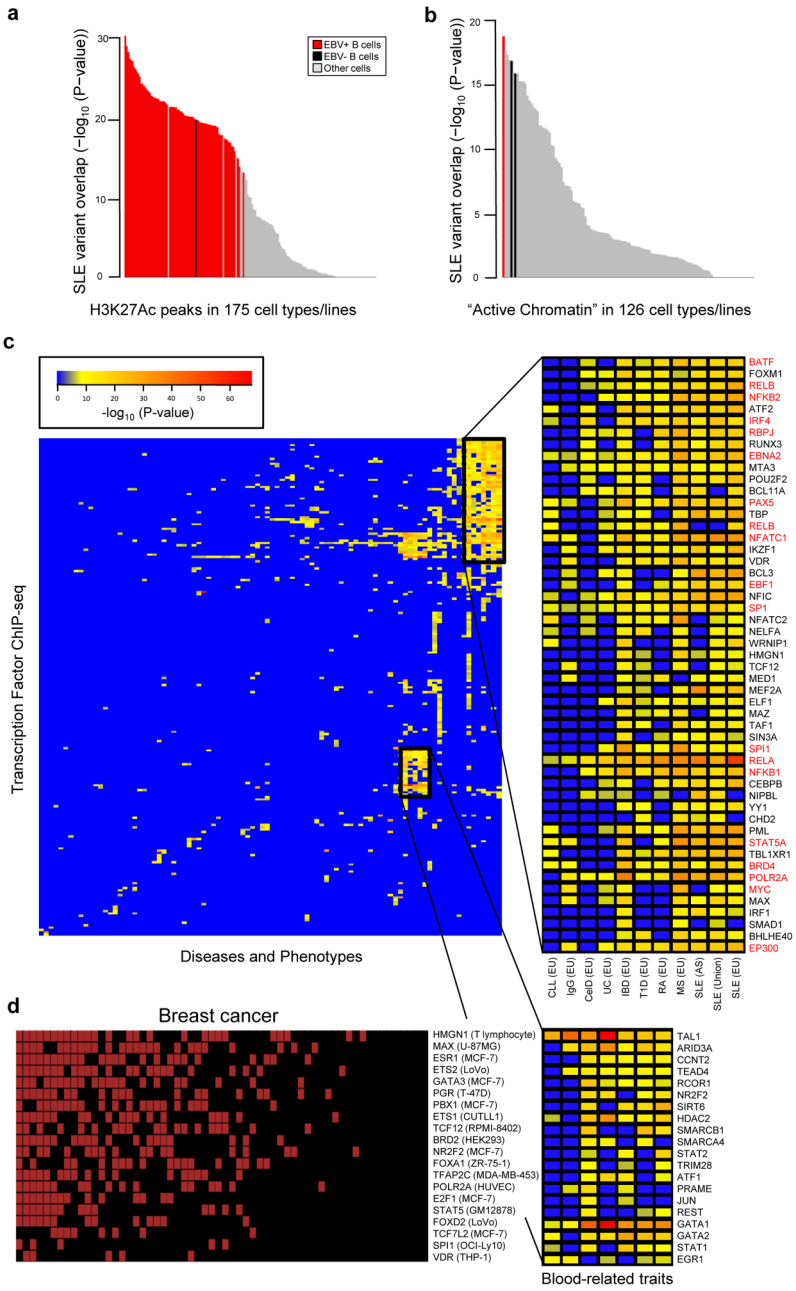

Figure 4. Cell types and TFs at disease-associated loci

a. SLE variants significantly intersect H3K27ac-marked regions in EBV-infected B cells. H3K27ac ChIP-seq peaks were collected from 175 different cell lines and types. The Y-axis indicates the negative log of the RELI P-value for the intersection of SLE-associated variants with H3K27ac peaks in each dataset. **b. SLE variants intersect active chromatin regions in EBV-infected B cells.** Same as (a), but instead using "active chromatin" regions, which are based on combinations of histone marks[44]. **c. Global view of RELI results – all diseases against all TFs.** Columns and rows show the 94 phenotypes/diseases and 212 TFs with at least one significant ($P_c<10^{-6}$) RELI result. Color indicates negative log of the RELI P-value (see key). Disease abbreviations are provided in the main text. **d. TFs intersecting**

breast cancer loci. Intersection between disease loci with TF-bound DNA sequences, as in Figure 1. However, here the cluster of TFs and risk loci instead largely may operate in ductal epithelial cells, independently of EBNA2. The top 20 TFs are shown - full results are provided in Supplementary Data Set 3.

Table 1

Intersection of TF ChIP-seq datasets with multiple genetic loci of diseases and phenotypes.

Phenotype	Cell line	TF	Number	Fraction	RR	P_c & P*
Prostate Ca	VCaP+Dht_18hr	AR	17	0.33	3.70	2.60E-07
Breast Ca	MCF7+Estradiol	GATA3	22	0.36	3.87	7.45E-11
MS	Mutu	EBNA2	44	0.40	4.66	6.34E-30
SSc	Mutu	EBNA2	2	0.10	-	NS
SSc	IB4	EBNA2	1	0.05	-	NS
SSc	GM12878	EBNA2	0	0.00	-	NS
SLE	Mutu	EBNA2	26	0.49	5.96	1.09E-25
SLE	IB4	EBNA2	10	0.19	7.46	1.09E-11
SLE	GM12878	EBNA2	10	0.19	8.57	1.94E-13
SLE	IB4	EBNA-LP	4	0.08	-	NS
SLE	Mutu	EBNA3C	5	0.09	-	NS
SLE	Raji	EBNA1	0	0.00	-	NS
SLE	Akata	Zta	0	0.00	-	NS
SLE*	Mutu*	EBNA2*	25*	0.63*	2.85*	1.81E-11*
SLE*	IB4*	EBNA2*	10*	0.25*	3.61*	2.44E-06*
SLE*	GM12878*	EBNA2*	10*	0.25*	4.97*	1.22E-09*

Detailed results are presented in Supplementary Data Set 3.

*RELI null model limited to EBV-infected B cell line open chromatin regions (see text).

RR = 'relative risk'. P_C = RELI Bonferroni corrected P-value. NS = P_C>10E-6. All disease ancestries are European. Ca = cancer. MS = multiple sclerosis. SSc = systemic sclerosis. SLE = systemic lupus erythematosus.

Table 2

Allele-dependent binding of EBNA2 to autoimmune-associated genetic variants.

Gene(s)	rs ID	ARS	Reads (Str.)	Reads (Weak)	Str. Base	Disease(s)
HLA-DQA1	rs9271693[#]	0.66	27	3	C	IBD, UC, Lung cancer
HLA-DQA1	rs9271588[#]	0.50	22	11	C	SjS[47]
IKZF2[*]	rs996032[#]	0.65	27	6	A	SLE (AS)
RERE[^]	rs2401138	0.63	48	20	C	V
TMBIM1[*]	rs2382818[#]	0.61	31	12	A	IBD
CLEC16A[^^]	rs7198004	0.59	16	0	G	SLE
CLEC16A	rs998592	0.50	10	0	C	SLE
CD44[^^]	rs3794102[#]	0.58	30	13	G	V
CD37[*]	rs1465697[#]	0.57	57	29	C	MS
BLK[^]	rs2736335	0.53	19	8	A	KD, KD (AS), SLE, SLE (AS), SLE (multi)
HLA-DQB1[^^]	rs3129763	0.52	11	0	A	CLL, SSc
PRKCQ	rs947474	0.52	11	0	A	T1D, RA[48]
TNIP1[*]	rs2233287	0.52	17	7	G	SSc
RHOH[^^]	rs13136820	0.52	141	86	T	GD
DQ658414 (MIR3142, MIR164A)[*]	rs73318382	0.50	10	0	A	SLE, SLE (AS), SLE (multi)
RMI2[^]	rs34437200	0.49	10	2	A	CelD, IBD, JIA, MS
ZFP36L1	rs194749[#]	0.47	11	4	C	IBD, T1D
HLA-DQB1[^^]	rs532098[#]	0.41	24	15	G	SLE
HLA-DRB1, HLA-DRB5	rs674313	0.41	24	15	G	CLL, SSc
PPIF[^^]	rs1250567	0.41	8	3	T	MS
TAGAP[*]	rs1738074	0.40	47	32	T	CelD

All ChIP-seq results are from Mutu cells, except for the *RMI2* locus, which is from GM12878 cells. Additional data are available in Supplementary Data Set 7. Each variant was assigned to a gene (column 1) as follows. If the variant is located within the promoter (+/− 5kb) of a gene expressed in EBV-infected B cells (median RPKM of 2 or more based on GTEx[49] data), assign it to that gene (indicated with '*'). Otherwise, if the variant is located within a Hi-C chromatin looping region in GM12878 EBV-infected B cells[50], assign it to the closest interacting gene that is expressed in EBV-infected B cells (indicated with '^^'). Otherwise, if the variant is located within a Hi-C chromatin looping region in primary B cells[51], assign it to the closest interacting gene that is expressed in EBV-infected B cells (indicated with '^'). Otherwise, assign the variant to the nearest gene that is expressed in EBV-infected B cells. Variants marked with '#' are eQTLs for the indicated gene in at least one EBV-infected B cell dataset[49,52–59]. "ARS": Allelic Reproducibility Score. "Reads (Strong (Str.))" and "Reads (Weak)" indicate the number of ChIP-seq reads mapping to the strong and weak allele, respectively. All disease associations are taken from the original disease lists (see Supplementary Data Set 1), with the exception of two additional associations (citations provided). GWAS results are of European ancestry, except as

indicated (East Asian (AS)). Disease abbreviations: MS, multiple sclerosis; IBD, inflammatory bowel disease; UC, ulcerative colitis; SLE, systemic lupus erythematosus; CLL, chronic lymphocytic leukemia; SSc, systemic sclerosis; SjS, Sjögren's syndrome; CelD, celiac disease; V, vitiligo; KD, Kawasaki's disease; T1D, type 1 diabetes; GD, Graves' disease; JIA, juvenile idiopathic arthritis.