nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0102-3

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

John B. Harley[1,2,3,4,5,9*], Xiaoting Chen[1,9], Mario Pujato[1,9], 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*]

Explaining the genetics of many diseases is challenging because most associations localize to incompletely characterized regulatory regions. Using new computational methods, we show that transcription factors (TFs) occupy multiple loci associated with individual complex genetic disorders. Application to 213 phenotypes and 1,544 TF binding datasets identified 2,264 relationships between hundreds of TFs and 94 phenotypes, including androgen receptor in prostate cancer and GATA3 in breast cancer. Strikingly, nearly half of systemic lupus erythematosus risk loci are occupied by the Epstein–Barr virus EBNA2 protein and many coclustering human TFs, showing 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 on EBNA2. Our results nominate mechanisms that operate across risk loci within disease phenotypes, suggesting new models for disease origins.

The mechanisms generating genetic associations have proven difficult to elucidate for most diseases because 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 suggestive relationships between EBV and multiple autoimmune diseases, the underlying molecular mechanisms remain unknown[8,9].

Genome-wide association studies (GWAS) have identified >50 convincing European-ancestry SLE susceptibility loci (Fig. 1a), providing compelling evidence for germline DNA polymorphisms altering SLE risk[10–13]. As in most complex diseases, most SLE loci occur in likely gene regulatory regions[14,15]. We therefore asked whether any of the DNA-interacting proteins encoded by EBV preferentially bind SLE risk loci. Our analyses identified strong associations with an EBV gene product (EBNA2), providing a potential origin of gene–environment interaction, along with a set of human transcription factors and cofactors (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 bound by a particular TF, as determined through chromatin immunoprecipitation and sequencing (ChIP-seq). 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 (Fig. 2a and Supplementary Fig. 1; see 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 showed that AR binding sites in the VCaP prostate cancer cell line 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 the MCF-7 breast cancer cell line significantly intersect breast-cancer-associated variants ($P_c < 10^{-10}$, Table 1), concordant with the established role of GATA3 in this disease[18]. Consistent with EBV contributing to multiple sclerosis (MS)[19–22], RELI showed that the EBV-encoded EBNA2 protein

[1]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [2]Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [3]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [4]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [5]US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA. [6]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [7]Division of Allergy & Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [8]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [9]These authors contributed equally: John B. Harley, Xiaoting Chen and Mario Pujato. *e-mail: john.harley@cchmc.org; leah.kottyan@cchmc.org; matthew.weirauch@cchmc.org

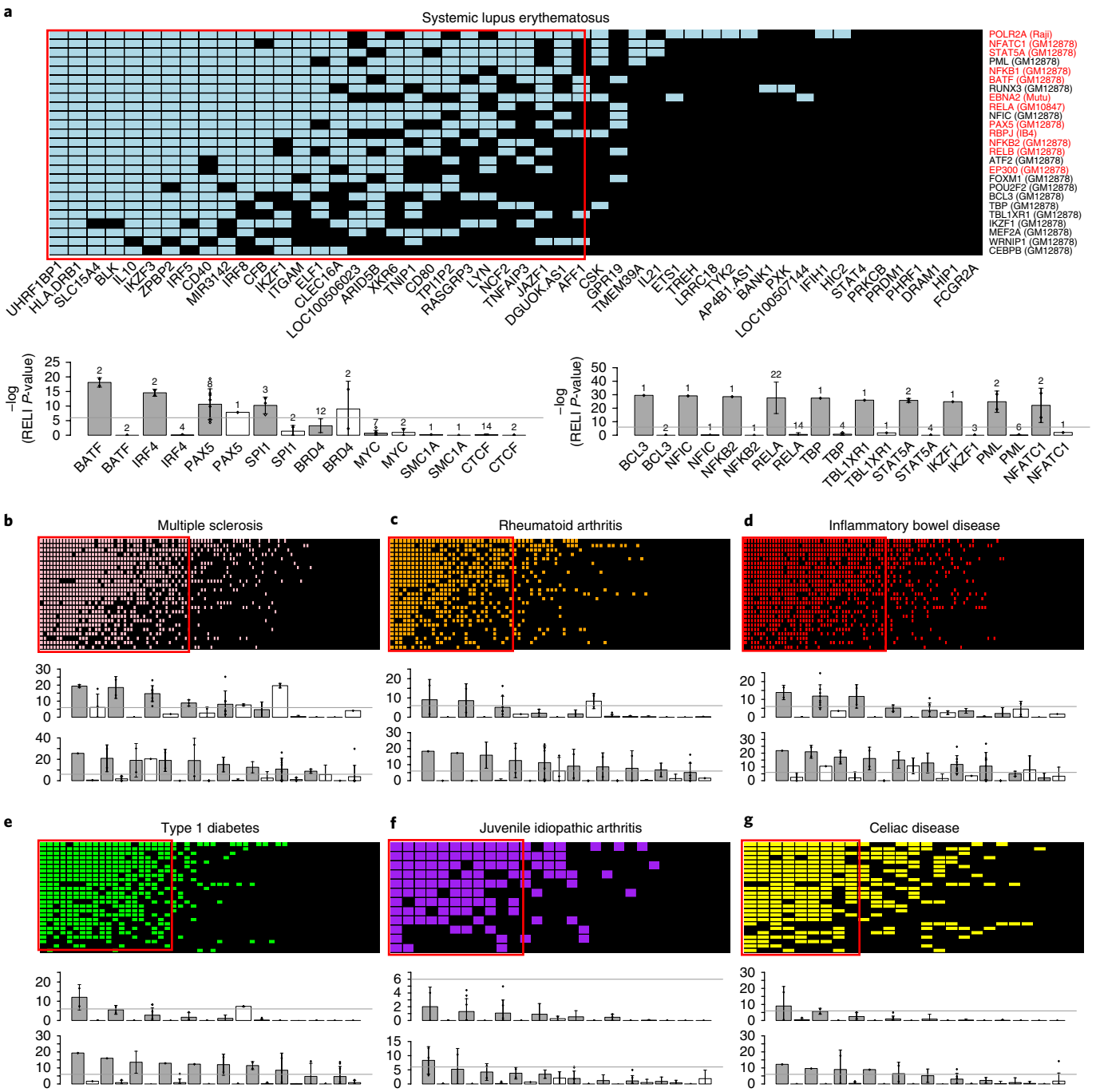

Fig. 1 | Intersection between autoimmune risk loci and TF binding interactions with the genome. a, Results for SLE risk loci. The *x* axis displays SLE-associated loci. The *y* axis displays the top 25 TFs, based on RELI *P*-values, sorted by the number of intersecting loci. A colored box indicates that the given locus contains at least one SLE-associated variant located within a ChIP-seq peak for the given TF (an 'intersection'). The most significant ChIP-seq dataset cell type is indicated in parentheses. Raji, GM12878, Mutu, and IB4 are all EBV-infected B cell lines. TFs that participate in EBNA2 super-enhancers[25] are colored red. The red rectangle identifies those loci and TFs that optimally cluster together (see Methods). Bottom panel, left: comparison of EBV-infected B cell lines (gray bars) to EBV-negative B cells (white bars). The *y* axis shows the distribution of the RELI $-\log(P_c)$ for each of the eight TFs with available data. Bars indicate mean. Error bars indicate s.d. Numbers indicate number of datasets. Horizontal line indicates the $P_c < 10^{-6}$ RELI significance threshold. Bottom panel, right: the top ten TFs (as based on RELI P_c-values) with data available in at least one EBV-infected B cell line (gray 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 Dataset 5.

occupies 44 of the 109 MS risk loci in the Mutu B cell line ($P_c < 10^{-29}$, Table 1). Prostate and breast cancer loci did not significantly intersect EBNA2 peaks, nor did 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 risk loci (all with $P < 5 \times 10^{-8}$ in case/control studies) with risk allele frequencies > 1%, in aggregate constituting 1,359 plausibly causal SLE variants (Supplementary Dataset 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

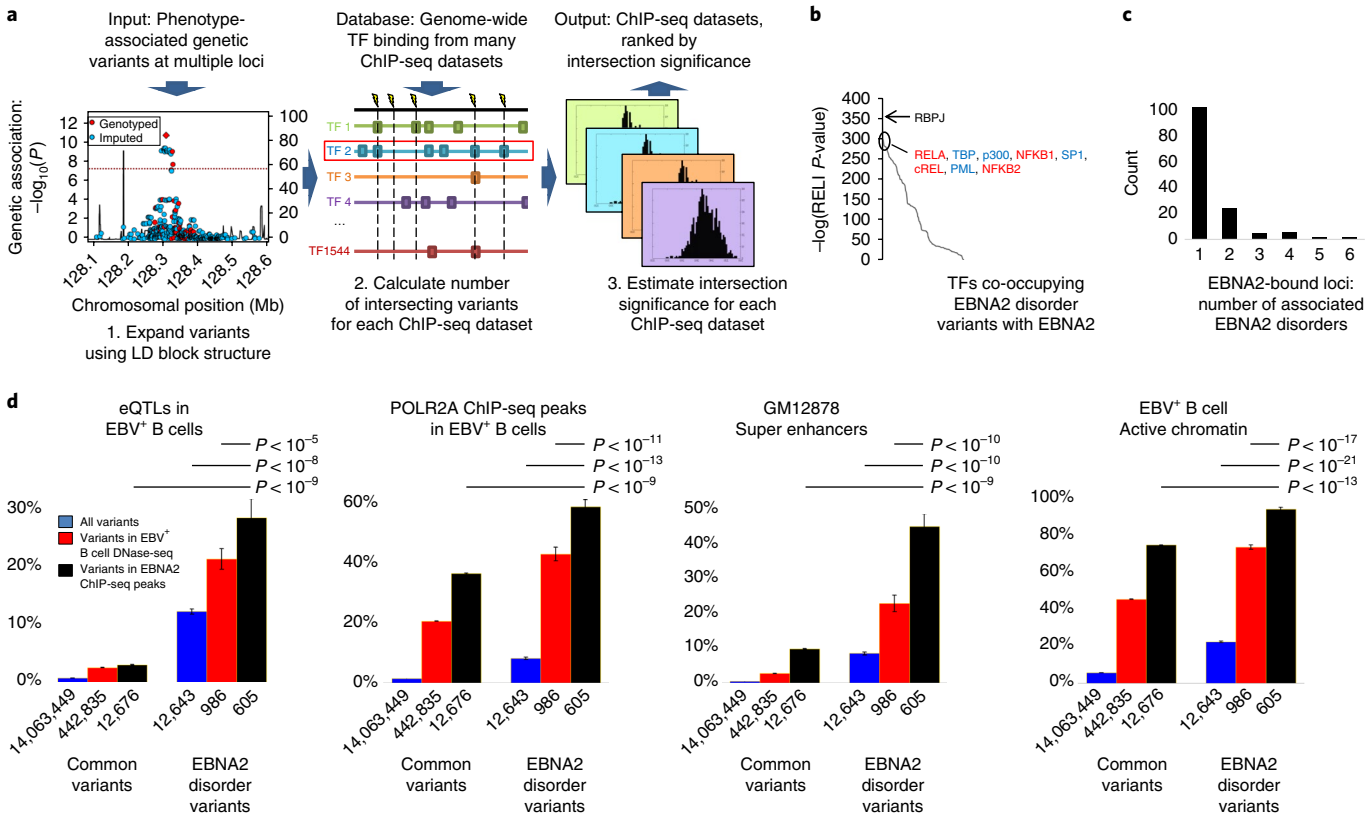

Fig. 2 | Properties of EBNA2-bound autoimmune disease loci. a, Schematic of the RELI algorithm (see Methods for details). In the left panel, the left *y*-axis indicates the strength of the genetic association (shown as data points); the right *y*-axis indicates the recombination rate (shown as blue bars). In the middle panel, lightning bolts represent phenotype-associated genetic variants. **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. NF-κB 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 are categorized by the number of EBNA2 disorders with which the given locus was associated (*x* axis). **d**, 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 percentage of variants in each group that are, for example, eQTLs in EBV-infected B cells (left plot). Error bars indicate the s.d. 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.

(three datasets), EBNA3C, EBNA-LP and Zta (Supplementary Dataset 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 six-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 colocalization 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 Dataset 2). In total, 132 ChIP-seq datasets involving 60 unique TFs strongly intersected SLE loci ($10^{-53} < P_c < 10^{-6}$). We chose a stringent corrected *P*-value cutoff of 10^{-6} on the basis of results from a simulation procedure aimed at estimating the false positive rate of our approach (see Methods). Notably, 109 (83%) of the significantly associated ChIP-seq datasets were obtained in EBV-infected B cell lines, with high fidelity between datasets (Supplementary Dataset 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 Fig. 2), and analogous results were obtained using the null model employed by the GoShifter[24] method (Supplementary Fig. 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 Dataset 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 (Fig. 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 were 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) had much weaker *P*-values in EBV-negative B cells (Fig. 1a, bottom left, and Supplementary Dataset 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

Table 1 | Intersection of TF ChIP-seq datasets with multiple genetic loci of diseases and phenotypes

Phenotype	Cell line	TF	No.	Fraction	RR	P_c or P[a]
PrCa	VCaP + DHT 18 h	AR	17	0.33	3.70	2.60×10^{-7}
BrCa	MCF-7 + estradiol	GATA3	22	0.36	3.87	7.45×10^{-11}
MS	Mutu	EBNA2	44	0.40	4.66	6.34×10^{-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	1.09×10^{-25}
SLE	IB4	EBNA2	10	0.19	7.46	1.09×10^{-11}
SLE	GM12878	EBNA2	10	0.19	8.57	1.94×10^{-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[a]	Mutu[a]	EBNA2[a]	25[a]	0.63[a]	2.85[a]	1.81×10^{-11a}
SLE[a]	IB4[a]	EBNA2[a]	10[a]	0.25[a]	3.61[a]	2.44×10^{-6a}
SLE[a]	GM12878[a]	EBNA2[a]	10[a]	0.25[a]	4.97[a]	1.22×10^{-9a}

Detailed results are presented in Supplementary Dataset 3. RR, relative risk; P_c, RELI Bonferroni-corrected *P*-value; NS, $P_c > 1 \times 10^{-6}$; DHT, dihydrotestosterone; PrCa, prostate cancer; BrCa, breast cancer; MS, multiple sclerosis; SSc, systemic sclerosis; SLE, systemic lupus erythematosus. All disease studies were performed in subjects of European ancestry. [a]RELI null model limited to EBV-infected B cell line open chromatin regions (see text); *P* values not Bonferroni-corrected.

strongest RELI *P*-values were obtained in EBV-infected B cells, and none of the other cell types available for these TFs showed significant association (Fig. 1a, bottom right). For example, 22 ChIP-seq datasets were available in EBV-infected B cells for the NF-κB subunit RELA. Of these, 20 significantly intersected with SLE risk loci ($10^{-53} < P_c < 10^{-17}$), while none of the remaining 14 available RELA datasets in any other cell type had significant intersection. Previous studies have demonstrated that EBV activates the NF-κB 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 US National Human Genome Research Institute GWAS catalog[29] and other sources (see Methods), and identified nine phenotypes displaying strong EBNA2 association with diseases in addition to SLE and MS: rheumatoid arthritis, inflammatory bowel disease (IBD), type 1 diabetes, juvenile idiopathic arthritis, celiac disease, chronic lymphocytic leukemia, Kawasaki disease, ulcerative colitis and immunoglobulin glycosylation (Supplementary Dataset 3). We designate the seven disorders among these with particularly strong EBNA2 associations ($P_c < 10^{-8}$) the 'EBNA2 disorders' (see Fig. 1). 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 Dataset 3). Thus, the overlap between EBNA2 ChIP-seq peaks and loci associated with the EBNA2 disorders is even stronger when considering only statistically likely causal variants.

Consistent with the SLE results (Fig. 1a), the same TFs tend to cluster with distinguishing loci for each disorder (Fig. 1b–g and Supplementary Dataset 5). Further, there is also a stronger association in EBV-infected than in EBV-negative cells for many TFs, and the ten most associated TFs consistently intersect more strongly in EBV-infected B cells than in other cell types (Fig. 1b–g and Supplementary Dataset 5). Hierarchical clustering identified a core set of 47 TFs binding to 142 risk loci across the seven EBNA2 disorders (Supplementary Fig. 4). RBPJ, an established EBNA2 cofactor[31–33], had the most similar binding profile to EBNA2 across loci, as expected.

NF-κB proteins RELA, RELB, REL, NFKB1 and NFKB2 comprised many of the strongest associations with EBNA2 disorder loci (Supplementary Dataset 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 NF-κB protein in EBV-infected B cells. Among the remaining 169 loci, 19 still contained 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 NF-κB involvement.

To identify candidate EBNA2 cofactors, we isolated EBNA2 disorder-associated variants located within EBNA2 ChIP-seq peaks and evaluated them using RELI. This analysis confirmed the importance of RBPJ, followed by members of the basal transcriptional machinery (TBP and p300) and NF-κB subunits (which are involved in EBNA2-mediated gene activation[34]) (Fig. 2b). Interestingly, predicted EBNA2 cofactors varied with disease phenotype; for example, EBNA2 and EBNA3C were highly synergistic at the disease loci of three of the EBNA2 disorders (IBD, MS and celiac disease), but rarely coincided at loci for the other four diseases (Supplementary Dataset 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 a single disorder (Fig. 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 Fig. 4).

If changes in gene regulation explain these results, then expression quantitative trait loci (eQTLs), ChIP-seq peaks for RNA polymerase 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 (Fig. 2d and Supplementary Dataset 3). For example, <1% of all common variants in the human genome are eQTLs in EBV-infected B cell lines (Fig. 2d, left panel, blue bars). This value rises to 2.3% for common variants located within open chromatin in EBV-infected B cell lines (red bars), and rises further to 2.7% for common variants within EBNA2 ChIP-seq peaks (black bars). 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

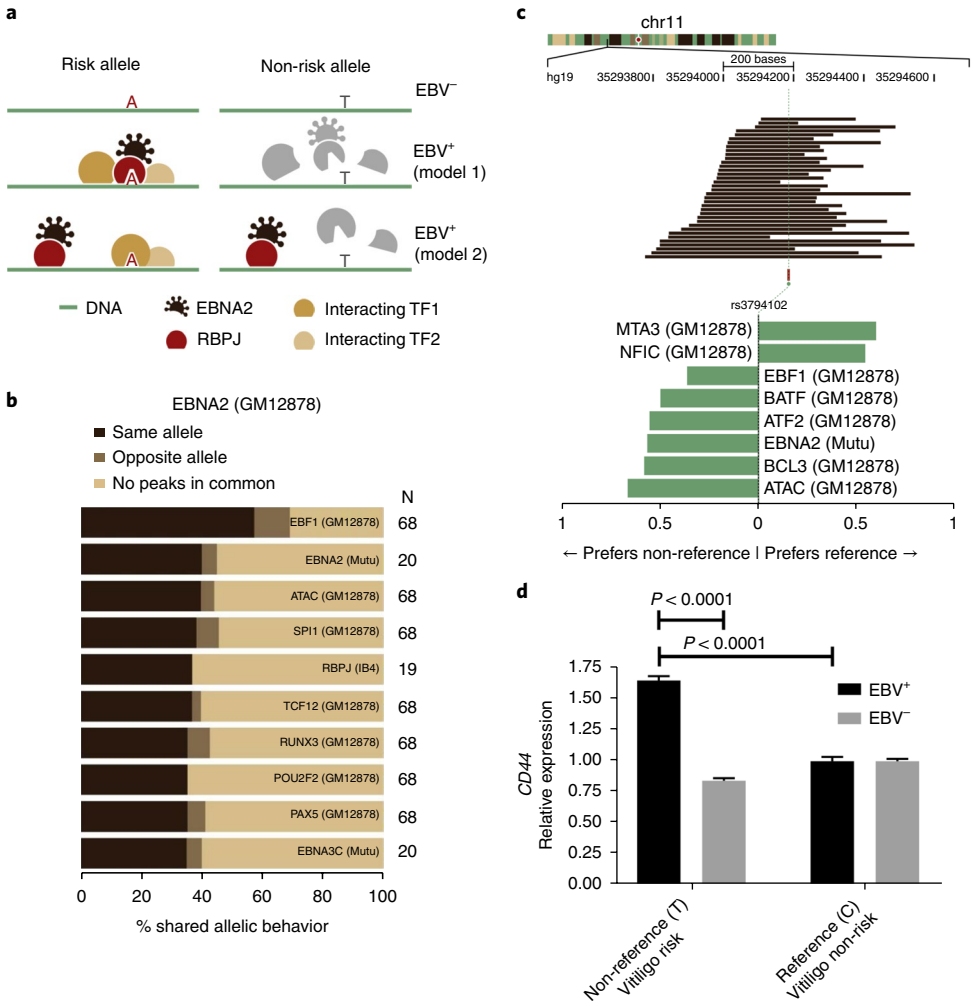

Fig. 3 | Allele-dependent binding of EBNA2 to autoimmune-associated genetic variants. a, Theoretical models presenting possible allele-dependent action of EBNA2 (see text for discussion). **b**, Allele-dependent co-binding of EBNA2 with multiple proteins. ChIP-seq datasets from EBV-infected B cell lines were examined for evidence of allele-dependent binding at heterozygous variants. Datasets are sorted by the proportion of EBNA2 GM12878 B cell line allele-dependent events (MARIO allelic reproducibility score > 0.40; see 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 (multicolored bar), location of EBV-infected B cell line ChIP-seq peaks for various TFs, location of rs3794102 variant, allele-dependent binding events (green bars). The *x* axis indicates the preferred allele, along with a value indicating the strength of the allelic behavior, calculated as 1 minus the ratio of the weak to strong reads (for example, 0.5 indicates the strong allele has approximately twice the reads of the weak allele). **d**, Allelic qPCR of *CD44* expression in EBV-infected and EBV-negative Ramos B cells. Fold-change in expression is provided relative to the C allele. Error bars represent s.d. ($n = 12$: 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.

were also eQTLs (Fig. 2d, right panel, black bars), a value significantly greater than the fraction of EBNA2 disorder variants located within open chromatin in EBV-infected B cell lines (20.5%; $P < 10^{-5}$, Welch's one-sided *t*-test) (red bars) or EBNA2 disorder variants in general (10.4%; $P < 10^{-8}$) (blue bars). Similar trends held for the other data types examined (Fig. 2d). In aggregate, these results hint at the potential magnitude of the environmental influence of EBNA2 on 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 (Fig. 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 modulation of the local chromatin environment by EBNA2 (Fig. 3a). Reanalysis 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 Methods). MARIO is an easy-to-use, modular tool that extends existing methods[37–40] by (i) calculating a score that explicitly reflects reproducibility across experimental replicates, (ii) reducing run time by using multiple computational cores and

Table 2 | Allele-dependent binding of EBNA2 to autoimmune-associated genetic variants

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

All ChIP-seq results are from Mutu cells, except for the *RMI2* locus, which is from GM12878 cells. Additional data are available in Supplementary Dataset 7. ARS, allelic reproducibility score. "Reads (strong)" 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 Dataset 1), with the exception of two additional associations (citations provided). GWAS results are for subjects of European ancestry, except where indicated as East Asian (AS) or multiple (multi) ancestries. CeID, celiac disease; GD, Graves' disease; IBD, inflammatory bowel disease; JIA, juvenile idiopathic arthritis; KD, Kawasaki's disease; MS, multiple sclerosis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; CLL, chronic lymphocytic leukemia; SSc, systemic sclerosis; SjS, Sjögren's syndrome; T1D, type 1 diabetes; UC, ulcerative colitis; V, vitiligo. Each variant was assigned to a gene (column 1) as follows.[a]If the variant is located within the promoter (±5 kb) of a gene expressed in EBV-infected B cells (median RPKM of 2 or more on the basis of GTEx[49] data), we assign it to that gene.[b]Otherwise, if the variant is located within a Hi-C chromatin looping region in GM12878 EBV-infected B cells[50], we assign it to the closest interacting gene that is expressed in EBV-infected B cells.[c]Otherwise, if the variant is located within a Hi-C chromatin looping region in primary B cells[51], we assign it to the closest interacting gene that is expressed in EBV-infected B cells. Otherwise, we assign the variant to the nearest gene that is expressed in EBV-infected B cells.[d]Variants are eQTLs for the indicated gene in at least one EBV-infected B cell dataset[49,52–59].

(iii) 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 Methods). We applied MARIO and a related method, ABC[37], to a deeply sequenced (~190 million reads) GM12878 ATAC-seq (Assay for Transposase-Accessible Chromatin and sequencing) dataset (GEO accession GSM1155957) and observed strong agreement between the 2,214 resulting scores (Spearman correlation 0.98, $P < 10^{-15}$).

We next applied MARIO to 271 ChIP-seq datasets obtained in one of the five genotyped cell lines, altogether assessing 98 different molecules. Since EBNA2 binds DNA indirectly as a cofactor, we first asked whether the variants displaying EBNA2 allele-dependent binding might coincide with similarly altered binding of other TFs. This analysis identified 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, Fig. 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 (Fig. 3b). Analogous results were obtained for EBNA2 ChIP-seq data in Mutu and IB4 cell lines (Supplementary Fig. 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 and Supplementary Dataset 7). We note that the number of heterozygous autoimmune variants for which EBNA2 prefers one allele over the other was not significantly more than expected by chance (see 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 are also involved in allele-dependent host protein binding, chromatin accessibility or presence of histone marks such as acetylation of histone H3 on Lys27 (H3K27ac) (Supplementary Dataset 8). Together, these results suggest that many autoimmune-associated variants may act by modifying host gene regulatory programs via altered binding of EBNA2 and other 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 EBV infection. We confirmed the expected presence or absence of EBNA2 by sequencing (Methods) and western blot (Supplementary Fig. 6). We identified 80 genes with significant EBV-dependent alterations in gene expression (Supplementary Dataset 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 Methods).

We next searched for autoimmune-associated variants that might affect 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 (i) plausibly causal for an autoimmune disorder; (ii) immunoprecipitated by EBNA2 antibodies; (iii) heterozygous in the cell line assayed; and (iv) 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 5 satisfy the fourth criterion of being within 50 kb 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), had significantly skewed allele-dependent binding of eight proteins; EBNA2, its suspected cofactor EBF1[36] and chromatin accessibility all favored the non-reference 'G' vitiligo risk allele (Fig. 3c, Table 2 and Supplementary Dataset 8). Intriguingly, the proteins favoring the 'G' allele are considered activators, whereas the two proteins favoring 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 detected no RNA-seq reads), loops to the promoter of the neighboring *CD44* gene as assessed by Hi-C experiments performed in GM12878 cells (Supplementary Fig. 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

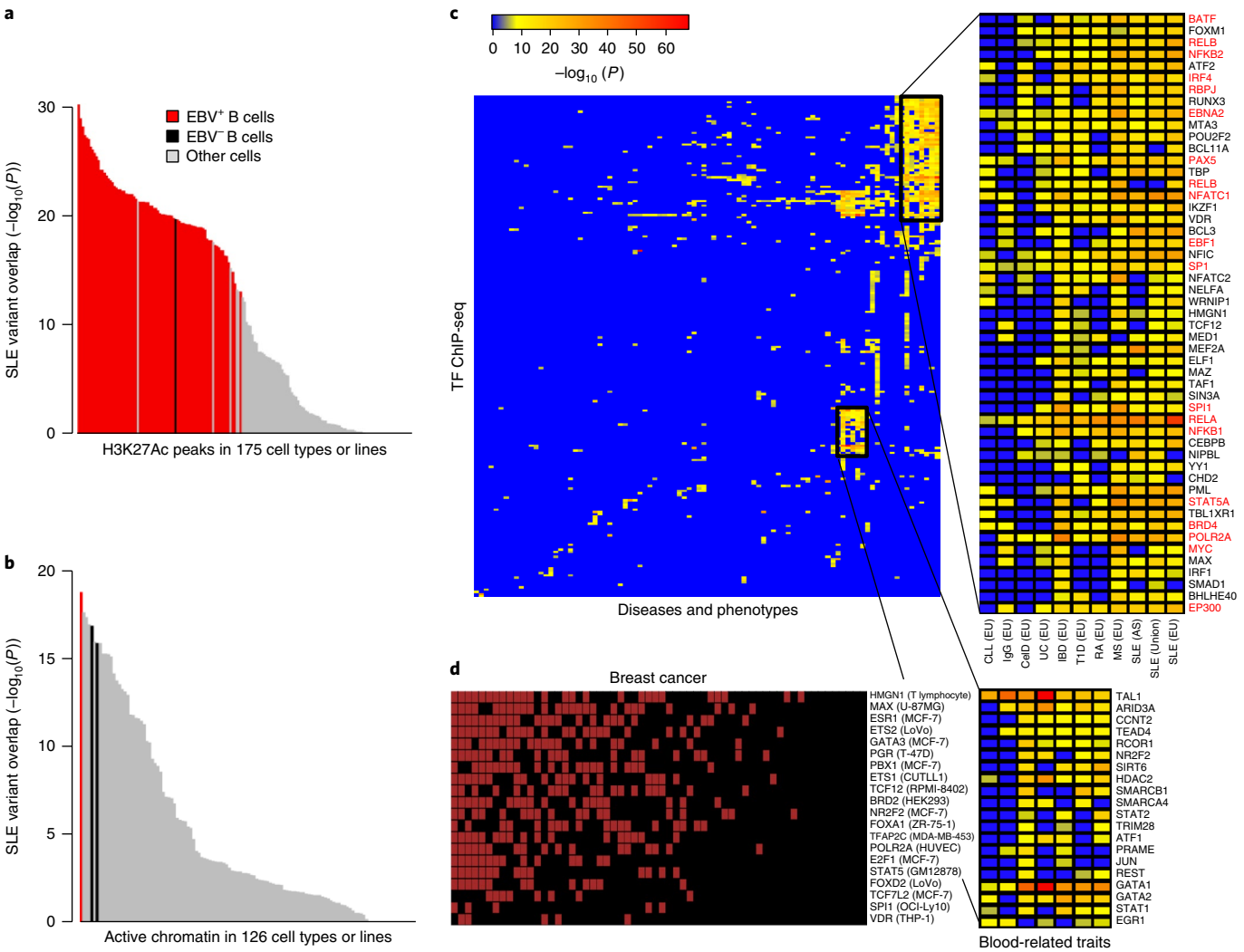

Fig. 4 | Cell types and TFs at disease-associated loci. a, SLE variants significantly intersect H3K27ac-marked regions in EBV-infected B cells. H3K27ac ChIP-seq peaks were collected from 175 different cell lines and types. The *y* axis indicates the negative log of the RELI *P*-value for the intersection of SLE-associated variants with the H3K27ac peaks contained 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 showing all diseases against all TFs. Columns and rows show the 94 phenotypes or diseases and 212 TFs with at least one significant ($P_c<10^{-6}$) RELI result. Color indicates negative log of the RELI *P*-value. TFs that participate in EBNA2super-enhancers[25] are colored red at the right. CLL, chronic lymphocytic leukemia; IgG, immunoglobulin glycosylation; CelD, celiac disease; UC, ulcerative colitis; IBD, inflammatory bowel disease; T1D, type 1 diabetes; RA, rheumatoid arthritis; MS, multiple sclerosis; SLE, systemic lupus erythematosus; EU, European ancestry; AS, Asian ancestry; union, all ancestries. **d**, TFs intersecting breast cancer loci. Intersection between disease loci with TF-bound DNA sequences (as in Fig. 1). However, here the cluster of TFs and risk loci instead may operate largely in ductal epithelial cells, independently of EBNA2. Cell lines: CUTLL1, T cell lymphoma; GM12878, EBV-infected B cell; HEK293, embryonic kidney; HUVEC, umbilical vein endothelial cells; LoVo, colon cancer; MCF-7, breast cancer; MDA-MB-453, breast cancer; OCI-Ly10, B-cell non-Hodgkin lymphoma; RPMI-8402, T cell acute lymphoblastic leukemia; THP-1, leukemic monocyte; T-47D, breast cancer; U-87MG, primary glioblastoma; ZR-75-1, breast cancer. The top 20 TFs are shown. Full results are provided in Supplementary Dataset 3.

presence of EBNA2[42]. In our experiments, *CD44* expression was 6.8-fold higher in EBV-infected Ramos cells compared to uninfected Ramos cells ($P=0.00015$, Supplementary Dataset 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 Dataset 10). We independently confirmed this result with allelic quantitative PCR (qPCR), observing a significantly higher expression for the T relative to the C allele in EBV-infected Ramos cells, with significantly lower expression in the absence of EBV (Fig. 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 type 1 diabetes and rheumatoid arthritis (Table 2) located near *PRKCQ*, another gene with allele- and EBV-dependent expression in our data (Supplementary Dataset 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, on the basis of the intersection between disease-associated variants and likely regulatory regions in that cell type. This procedure identified 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 were all from EBV-infected B cell lines (Fig. 4a). Analogous results were obtained for 'active chromatin marks' (a model based on combinations of various histone marks[44]) (Fig. 4b), H3K4me3 and H3K4me1, for SLE and the other seven EBNA2 disorders (Supplementary Datasets 3 and 11). Collectively, these results support the EBV-infected B cell being an origin of genetic risk for each of the seven EBNA2 disorders. This analysis also highlights a likely involvement of other immune cell types in these disorders, including T cells, natural killer cells and monocytes (Supplementary Dataset 3). Although there are limited TF ChIP-seq data available for these cell types, one or more of the EBNA2 disorders were associated with 17 of the available T cell TF ChIP-seq datasets (Supplementary Dataset 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 Dataset 12). Together, these results are consistent with multiple shared regulatory mechanisms acting across autoimmune risk loci, some common between cell types (for example, 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 Datasets 1 and 3). In addition to the EBNA2-related associations, clustering of these results shows a large grouping of hematopoietic phenotypes and well-established blood cell regulators such as GATA1 and TAL1 (Fig. 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 (Fig. 4d) and AR, NR3C1 and EZH2 in prostate cancer (Supplementary Dataset 3). In total, application of these methods produced results nominating global disease mechanisms for 94 different diseases or phenotypes (Supplementary Dataset 3), providing new directions for understanding their origins.

Discussion

Our efforts to understand the gene–environment interaction between SLE loci and EBV have shown that EBNA2 and its associated human TFs occupy a substantial fraction of autoimmune risk loci. In particular, NF-κB subunits such as RELA, RELB, NFKB1 and NFKB2 also strongly intersect many of these loci, suggesting that NF-κB 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 (Fig. 1), with the caveat that existing data are biased, having been predominantly collected in this cell type.

Notably, 4 of the top 20 TFs that co-occupy EBNA2 disorder loci with EBNA2 (MED1, p300, NFKB1 and NFKB2) can be targeted by at least one available drug[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 for experimental verification and exploration that may explain the molecular and cellular origins of disease risk. As new genetic association and TF binding data are collected, approaches such as this will undoubtedly identify further disease mechanisms.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0102-3.

Received: 17 May 2017; Accepted: 31 January 2018;
Published online: 16 April 2018

References

1. Fujinami, R. S., von Herrath, M. G., Christen, U. & Whitton, J. L. Molecular mimicry, bystander activation, or viral persistence: infections and autoimmune disease. *Clin. Microbiol. Rev.* **19**, 80–94 (2006).
2. Ercolini, A. M. & Miller, S. D. The role of infections in autoimmune disease. *Clin. Exp. Immunol.* **155**, 1–15 (2009).
3. Sener, A. G. & Afsar, I. Infection and autoimmune disease. *Rheumatol. Int.* **32**, 3331–3338 (2012).
4. James, J. A. et al. An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. *J. Clin. Invest.* **100**, 3019–3026 (1997).
5. Hanlon, P., Avenell, A., Aucott, L. & Vickers, M. A. Systematic review and meta-analysis of the sero-epidemiological association between Epstein-Barr virus and systemic lupus erythematosus. *Arthritis Res. Ther.* **16**, R3 (2014).
6. McClain, M. T. et al. Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry. *Nat. Med.* **11**, 85–89 (2005).
7. Harley, J. B. & James, J. A. Epstein-Barr virus infection induces lupus autoimmunity. *Bull. NYU Hosp. Jt. Dis.* **64**, 45–50 (2006).
8. Ascherio, A. & Munger, K. L. EBV and autoimmunity. *Curr. Top. Microbiol. Immunol.* **390**, 365–385 (2015).
9. Draborg, A. H., Duus, K. & Houen, G. Epstein-Barr virus in systemic autoimmune diseases. *Clin. Dev. Immunol.* **2013**, 535738 (2013).
10. Vaughn, S. E., Kottyan, L. C., Munroe, M. E. & Harley, J. B. Genetic susceptibility to lupus: the biological basis of genetic risk found in B cell signaling pathways. *J. Leukoc. Biol.* **92**, 577–591 (2012).
11. Alarcón-Riquelme, M. E. 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.* **68**, 932–943 (2016).
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.* **47**, 1457–1464 (2015).
13. Sun, C. et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat. Genet.* **48**, 323–330 (2016).

14. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
15. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
16. Fang, H., Knezevic, B., Burnham, K. L. & Knight, J. C. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med.* **8**, 129 (2016).
17. Schweizer, M. T. & Yu, E. Y. Persistent androgen receptor addiction in castration-resistant prostate cancer. *J. Hematol. Oncol.* **8**, 128 (2015).
18. Asch-Kendrick, R. & Cimino-Mathews, A. The role of GATA3 in breast carcinomas: a review. *Hum. Pathol.* **48**, 37–47 (2016).
19. Almohmeed, Y. H., Avenell, A., Aucott, L. & Vickers, M. A. Systematic review and meta-analysis of the sero-epidemiological association between Epstein Barr virus and multiple sclerosis. *PLoS One* **8**, e61110 (2013).
20. Pender, M. P. & Burrows, S. R. Epstein-Barr virus and multiple sclerosis: potential opportunities for immunotherapy. *Clin. Transl. Immunology* **3**, e27 (2014).
21. Márquez, A. C. & Horwitz, M. S. The role of latently infected B cells in CNS autoimmunity. *Front. Immunol.* **6**, 544 (2015).
22. Ricigliano, V. A. et al. EBNA2 binds to genomic intervals associated with multiple sclerosis and overlaps with vitamin D receptor occupancy. *PLoS One* **10**, e0119605 (2015).
23. Hu, X. et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* **89**, 496–506 (2011).
24. Trynka, G. et al. Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).
25. Zhou, H. et al. Epstein-Barr virus oncoprotein super-enhancers control B cell growth. *Cell Host Microbe* **17**, 205–216 (2015).
26. Gewurz, B. E. et al. Canonical NF-κB activation is essential for Epstein-Barr virus latent membrane protein 1 TES2/CTAR2 gene regulation. *J. Virol.* **85**, 6764–6773 (2011).
27. Ersing, I., Bernhardt, K. & Gewurz, B. E. NF-κB and IRF7 pathway activation by Epstein-Barr virus Latent Membrane Protein 1. *Viruses* **5**, 1587–1606 (2013).
28. Price, A. M. 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-κB activation. *J. Virol.* **86**, 11096–11106 (2012).
29. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
30. Farh, K. K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
31. Zimber-Strobl, U. et al. Epstein-Barr virus nuclear antigen 2 exerts its transactivating function through interaction with recombination signal binding protein RBP-Jκ, the homologue of *Drosophila* Suppressor of Hairless. *EMBO J.* **13**, 4973–4982 (1994).
32. Grossman, S. R., 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. USA* **91**, 7568–7572 (1994).
33. Henkel, T., Ling, P. D., Hayward, S. D. & Peterson, M. G. Mediation of Epstein-Barr virus EBNA2 transactivation by recombination signal-binding protein J kappa. *Science* **265**, 92–95 (1994).
34. Scala, G. et al. Epstein-Barr virus nuclear antigen 2 transactivates the long terminal repeat of human immunodeficiency virus type 1. *J. Virol.* **67**, 2853–2861 (1993).
35. Wang, J. H. et al. Aiolos regulates B cell activation and maturation to effector state. *Immunity* **9**, 543–553 (1998).
36. Lu, F. et al. EBNA2 drives formation of new chromosome binding sites and target genes for B-cell master regulatory transcription factors RBP-jκ and EBF1. *PLoS Pathog.* **12**, e1005339 (2016).
37. Bailey, S. D., Virtanen, C., Haibe-Kains, B. & Lupien, M. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics* **31**, 3057–3059 (2015).
38. Buchkovich, M. L. et al. Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. *BMC Med. Genomics* **8**, 43 (2015).
39. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
40. Shi, W., Fornes, O., Mathelier, A. & Wasserman, W. W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* **44**, 10106–10116 (2016).
41. Ma, B., Huang, J. & Liang, L. RTeQTL: real-time online engine for expression quantitative trait loci analyses. Database (Oxford) 2014, bau066 https://doi.org/10.1093/database/bau066 (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* **86**, 41–48 (1995).
43. Gonnella, R. et al. PKC theta and p38 MAPK activate the EBV lytic cycle through autophagy induction. *Biochim. Biophys. Acta* **1853**, 1586–1595 (2015).
44. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
45. Griffith, M. et al. DGIdb: mining the druggable genome. *Nat. Methods* **10**, 1209–1210 (2013).
46. Harter, M. R. et al. BS69/ZMYND11 C-terminal domains bind and inhibit EBNA2. *PLoS Pathog.* **12**, e1005414 (2016).
47. Li, Y. et al. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjögren's syndrome at 7q11.23. *Nat. Genet.* **45**, 1361–1365 (2013).
48. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
49. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
50. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
51. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016).
52. Liang, L. et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* **23**, 716–726 (2013).
53. Stranger, B. E. et al. Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
54. Veyrieras, J. B. et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
55. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
56. Montgomery, S. B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
57. Mangravite, L. M. et al. A statin-dependent QTL for *GATM* expression is associated with statin-induced myopathy. *Nature* **502**, 377–380 (2013).
58. Dimas, A. S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
59. Gaffney, D. J. et al. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).

Acknowledgements

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; and Y. Yuan (University of Penn) and D. Thorley-Lawson (Tufts Institute) for generous donation of cell lines (Mutu and IB4, respectively). We also thank 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, 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.

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 interests

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

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0102-3.

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

Correspondence and requests for materials should be addressed to J.B.H. or L.C.K. or M.T.W.

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

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, systemic sclerosis, rheumatoid arthritis and juvenile idiopathic arthritis, peer-reviewed literature was curated (Supplementary Dataset 13). Only genetic associations exceeding genome-wide significance ($P < 5 \times 10^{-8}$) 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 Dataset 1).

Functional genomics data were obtained from ENCODE[62] (downloaded April 2014), Roadmap epigenomics[63] (July 2015), Cistrome[64] (December 2015), PAZAR[65] (April 2014), ReMap-ChIP[66] (August 2015) and Gene Expression Omnibus[67] (Supplementary Dataset 2). ChIP-seq datasets containing less than 500 peaks were removed. eQTLs were obtained from GTExPortal[49] (January 2016), the Pritchard lab eQTL database (http://eqtl.uchicago.edu/) (April 2014) and the Harvard eQTL database (https://www.hsph.harvard.edu/liming-liang/software/eqtl/) (April 2014). 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 (for example, ChIP-seq datasets) (Fig. 2a and Supplementary Fig. 1). In Step 1, sequencing data from 1000 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 Fig. 8a), 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 Fig. 8b), conservatively setting an empirically determined *P*-value lower bound at 5×10^{-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 Fig. 8c).

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 Dataset 3). Further, the *P*-value for this dataset ($P_c < 10^{-8}$) is much less significant than those for EBNA2, RELA, etc. ($P_c << 10^{-20}$). 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 Fig. 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 heat map). We then iteratively considered every possible submatrix 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 submatrix by comparing the observed number of intersections within the submatrix to the null distribution, using a standard *z*-score transformation. The optimal cluster was defined as the submatrix 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 nondisease controls to allow 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 1000 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 heterozygous genetic variants within peaks (described below) and annotates the results (Supplementary Fig. 9).

To estimate the significance of the degree of allelic imbalance of a given dataset at a given heterozygous variant, 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 (see 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 Fig. 10a).

(2) Identify predictive variables of reproducible allele-dependent behavior across replicates. We collected a set of seven datasets, {D}, with each dataset comprising four experimental replicates, {R} (Supplementary Fig. 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 Fig. 10c). Each variable was assessed on the basis of 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 Fig. 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). We designate these 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 Fig. 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 Fig. 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 Fig. 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 Fig. 11b. Curves were fit to these points using a saturating function:

$$\mathrm{ARS}_w = \frac{A_w}{1 + B_w \times r} - A_w$$

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 Fig. 11c).

Statistical significance of the number of EBNA2 allele-dependent binding events. We observed 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 (i) located within a ChIP-seq peak in at least one of the three available EBNA2 datasets and (ii) 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 ten 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). Wild-type EBV was prepared from supernatants of B95-8 cells cultured in RPMI medium 1640 supplemented with 10% FBS for 2 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 −80 °C.

We infected ~2×10^6 Ramos cells (ATCC CRL-1596) in the presence of growth medium containing 2 µg/ml of phytohemagglutinin for 4 h. The infected cells were washed, cultured in growth medium and observed daily for multinuclear giant cell formation and morphological changes characteristic of EBV-infected B cells. After ten passages, infection was confirmed by measuring the expression of viral EBNA2 protein levels (Supplementary Fig. 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-bp 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 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 quality control, 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 was 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 detected, 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 is provided in the figures and legends. Data are represented as means ± 1 s.d., unless otherwise noted.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

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 Information.

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 Laboratory GitHub page (https://github.com/WeirauchLab/).

References

60. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
61. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
62. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
63. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
64. Liu, T. et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
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.* **37**, D54–D60 (2009).
66. Griffon, A. et al. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43**, e27 (2015).
67. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
68. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
69. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
70. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).
71. Kottyan, L. C. et al. Genome-wide association analysis of eosinophilic esophagitis provides insight into the tissue specificity of this allergic disease. *Nat. Genet.* **46**, 895–900 (2014).
72. Verma, S. S. et al. Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* **5**, 370 (2014).
73. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
74. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
75. Flicek, P. et al. Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
76. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
77. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
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.* **67**, 2209–2220 (1993).
79. Chen, C. C. et al. NF-κB-mediated transcriptional upregulation of TNFAIP2 by the Epstein-Barr virus oncoprotein, LMP1, promotes cell motility in nasopharyngeal carcinoma. *Oncogene* **33**, 3648–3659 (2014).

80. Craig, F. E. et al. Gene expression profiling of Epstein-Barr virus-positive and -negative monomorphic B-cell posttransplant lymphoproliferative disorders. *Diagn. Mol. Pathol.* **16**, 158–168 (2007).
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* **8**, e64868 (2013). 8.
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* **102**, 4166–4178 (2003).
83. Lee, I. S., Shin, Y. K., Chung, D. H. & Park, S. H. LMP1-induced downregulation of CD99 molecules in Hodgkin and Reed-Sternberg cells. *Leuk. Lymphoma* **42**, 587–594 (2001).

nature research

Corresponding Author: Weirauch

Date: January 18, 2018

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   We determined sample size based upon available data and power calculations. Please see Supplemental Methods: "Collection and Processing of Datasets" "Cell line genotyping and imputation" sections.

2. Data exclusions

   Describe any data exclusions.

   Data was excluded if it did not meet explicit quality control criteria. Please see. Supplemental Methods: "Collection and Processing of Datasets" "Cell line genotyping and imputation" sections. These sections were sent as a word doc with the pdf form.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   The findings were replicated across multiple diseases and cell types (please see Figure 1 and Supplementary Tables).

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   This is not applicable. We used pre-identified genetic variants and showed enrichment of TF binding based on pre-identified ChIP results.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Investigators were blinded to the risk/non-risk alleles during "prefered allele" analysis. Investigators were blinded to risk allele in luciferase studies.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We developed our own code. The source code for RELI and MARIO (the two novel algorithms of this paper) are freely available under the GNU General Public License, with full documentation and examples, on the Weirauch lab GitHub page: https://github.com/WeirauchLab/ Plink 1.9b; Illumina Genome Studio 2.0.3; IMIPUTE2 v2.2.2; MACS2 2.1.0; PICARD 1.89; Bowtie 2.2.0; TopHat 2.0.8b; FastQC v0.11.5; Hisat2 2.0.4; Cufflinks 2.2.1.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were created for this project.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Westerns were performed for EBNA (and a β-actin control) – see Supplementary Figure 6. Antibodies used: EBNA2 (clone PE2-ab90543 (Abcam, Cambridge, MA), anticipated molecular weight of 75kDa); β-actin (ab8227 (Abcam), anticipated molecular weight of 42kDa). Anti-EBNA2 antibody was validated by assessing its specificity to EBV infected cells. Both antibodies were further validated by comparing the band pattern and molecular weight to the manufacturer's validated Western blots.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	GM12878, Corielle Institute for Medical Research; Mutu, Yan Yuan, University of Pennsylvania, School of Dental Medicine, Philadelphia, PA; IB4, David A. Thorley- Lawson, Tufts University School of Medicine, Boston, MA; Ramos, ATCC; Raji, ATCC
b. Describe the method of cell line authentication used.	We genotyped all cells used and confirmed concordance with published ChIP-seq genotypes.
c. Report whether the cell lines were tested for mycoplasma contamination.	Yes, all cell lines were tested for mycoplasma as noted in extended methods.
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used from the list of the commonly mis-identified cell lines database.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals or animal-derived materials were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No direct human subjects research was part of this study.