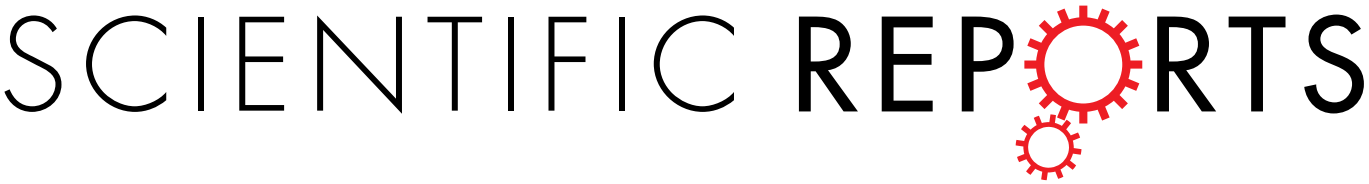

OPEN

ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells

Received: 14 February 2016
Accepted: 12 May 2016
Published: 01 June 2016

Christopher D. Scharer[1,*], Emily L. Blalock[2,*], Benjamin G. Barwick[1], Robert R. Haines[1], Chungwen Wei[2], Ignacio Sanz[2] & Jeremy M. Boss[1]

Biobanking is a widespread practice for storing biological samples for future studies ranging from genotyping to RNA analysis. However, methods that probe the status of the epigenome are lacking. Here, the framework for applying the Assay for Transposase Accessible Sequencing (ATAC-seq) to biobanked specimens is described and was used to examine the accessibility landscape of naïve B cells from Systemic Lupus Erythematosus (SLE) patients undergoing disease flares. An SLE specific chromatin accessibility signature was identified. Changes in accessibility occurred at loci surrounding genes involved in B cell activation and contained motifs for transcription factors that regulate B cell activation and differentiation. These data provide evidence for an altered epigenetic programming in SLE B cells and identify loci and transcription factor networks that potentially impact disease. The ability to determine the chromatin accessibility landscape and identify *cis*-regulatory elements has broad application to studies using biorepositories and offers significant advantages to improve the molecular information obtained from biobanked samples.

Biorepositories are a growing and important source of biological specimens that allow researchers access to large cohorts of samples that would otherwise be unobtainable. Protocols for extraction and molecular phenotyping of DNA and RNA from biobanked specimens have been developed[1]. However, methods examining the epigenetic state of biobanked cells are lacking. Epigenetic information has the potential to reveal details about the molecular programming of cells, including the location and status of *cis*-regulatory elements. For example, the mapping of intergenic regulatory elements combined with traditional GWAS studies could improve the functional understanding of non-coding polymorphisms. However, it is not known whether the biobanking process preserves chromatin structure, thereby facilitating or inhibiting such analyses.

The Assay for Transposase Accessible Chromatin (ATAC-seq) utilizes a sequencing adapter-coupled Tn5 transposase to simultaneously tag and fragment native chromatin, thereby generating a high-resolution map of accessible loci from cells[2]. ATAC-seq is highly efficient, requires fewer cells than other epigenetic profiling assays, such as ChIP-seq, and can be used as a readout to predict epigenetic states. Here, the ATAC-seq assay was applied to both biobanked and freshly processed specimens and an indistinguishable chromatin accessibility pattern was observed. To validate the use of ATAC-seq on clinically biobanked specimens, the chromatin accessibility landscape was determined for naïve B cells isolated from an existing biorepository of Systemic Lupus Erythematosus (SLE) samples. Differentially accessible loci suggested a unique accessibility signature of SLE B cells and highlight transcription factor networks and loci that may contribute to disease.

Results

B cell complexity is preserved through biobanking. To facilitate the storage and sharing of clinical samples within and between institutions of the Autoimmunity Centers of Excellence, a robust PBMC biobanking protocol was established. Following thawing and preparation for FACS sorting, a near identical cellular viability

[1]Department of Microbiology and Immunology, Emory University School of Medicine, Atlanta, GA 30322, USA. [2]Department of Rheumatology, Emory University School of Medicine, Atlanta, GA 30322, USA. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to I.S. (email: ignacio.sanz@emory.edu) or J.M.B. (email: jmboss@emory.edu)

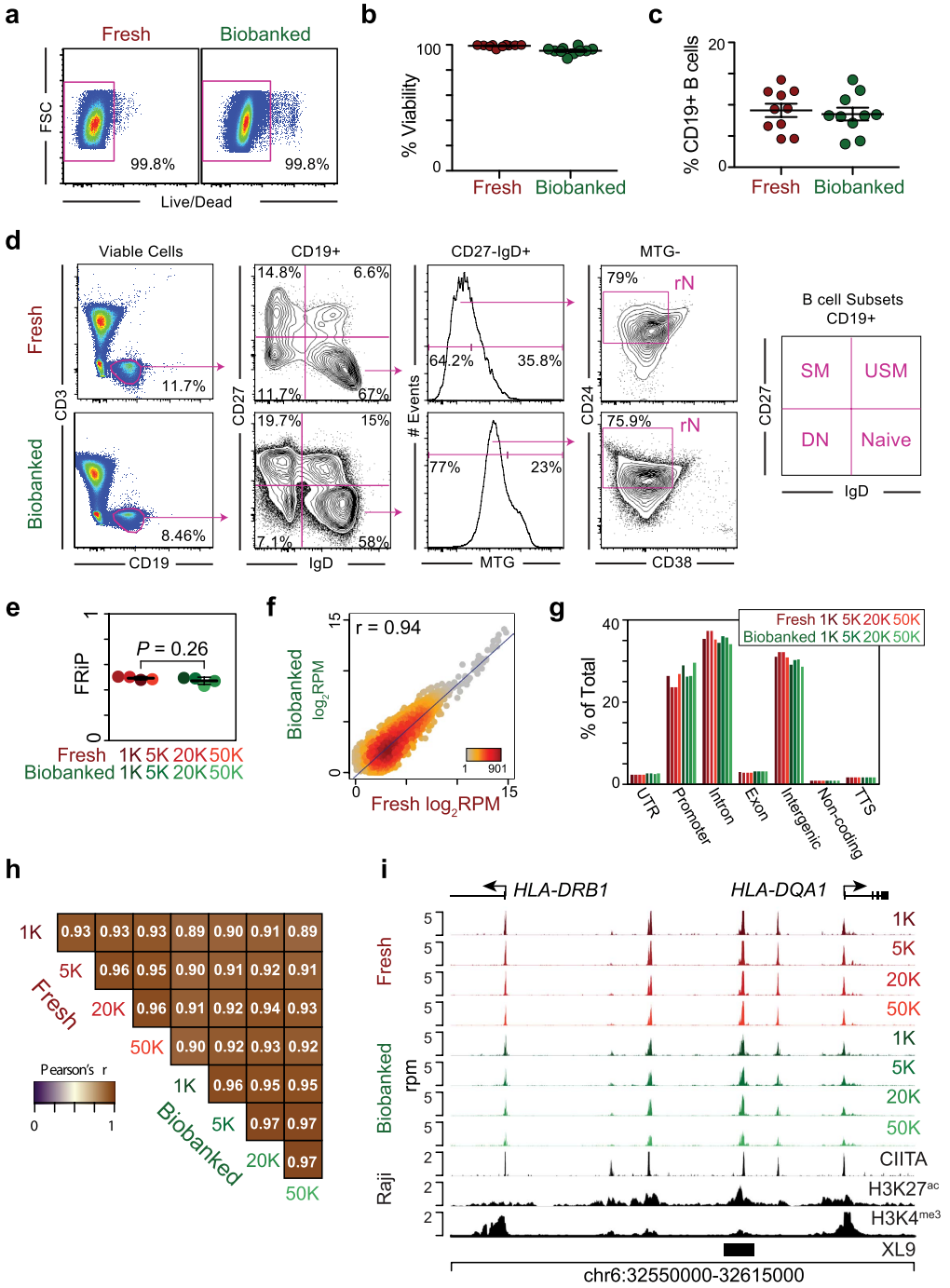

Figure 1. Biobanked samples display indistinguishable chromatin accessibility profiles from freshly processed samples. (**a**) Representative flow cytometry plots of cellular viability for fresh and biobanked specimens. Samples are gated on FSC and SSC prior to viability analysis. (**b**) Dot plot of the percentage viability for ten fresh and ten biobanked specimens. (**c**) Dot plot showing the percentage of $CD19^+$ B cells for ten fresh and ten biobanked samples. (**d**) Flow cytometry analysis showing the phenotype of B cell subsets from a representative fresh and biobanked sample and the gating strategy to isolate naïve B cells (rN). The distinct $CD19^+$ B cell subsets are indicated on the right. SM, switched memory; USM, unswitched memory; DN, double negative. (**e**) The fraction of reads in peaks (FRiP) metric for each sample is plotted. Statistical significance was tested by Student's *t*-test. (**f**) Density scatter plot of the ATAC-seq reads in 76,591 combined accessible peaks from fresh and biobanked samples. Pearson's correlation coefficient r value is indicated along with the scatter plot density. rpm, reads per million. (**g**) Barplot representing the percentage of peaks in each sample overlapping distinct genomic features. TTS, transcription termination site. UTR, 5′ and 3′ UTRs. (**h**) Heatmap of the Pearson's correlation r values for all sample comparisons. The r value for each comparison is indicated. (**i**) Genome plot of the MHCII locus showing the profile for each ATAC-seq sample. Genomic profiles for CIITA, $H3K27^{ac}$, and $H3K4^{me3}$ from Raji B lymphoblastoid cells[6] and the position of the XL9 insulator[5] are plotted. The genomic coordinates, positions of genes, direction of transcription, and exon locations are annotated.

was observed for biobanked specimens compared to freshly isolated samples (Fig. 1a,b). Additionally, the biobanking process maintained B cell complexity as determined by the frequency of peripheral $CD19^+$ B cells (Fig. 1c) and distinct B cell subsets (Fig. 1d and Supplementary Fig. S1a,b). These data demonstrated that biobanked cells are viable and display surface markers similar to freshly processed cells.

Chromatin structure is preserved through biobanking. To determine if the biobanking process preserved chromatin structure, thus facilitating the determination of epigenetic states, a study applying the ATAC-seq assay to fresh and biobanked human B cells was performed. PBMCs from a single healthy donor were split in half and processed fresh or biobanked for one week. Next, $CD19^+$ naive B cells (IgD^+ $CD19^+$ $MTG^-CD27^-CD38^+$ $CD24^+$) were FACS isolated from both fresh and biobanked samples. To determine if there were cell input limitations associated with biobanking, 1,000, 5,000, 20,000, and 50,000 cells were isolated from each sample and ATAC-seq was performed. Accessible peaks of enrichment were determined and the fraction of reads in peaks (FRiP) was calculated. The FRiP metric can be used to assess background in enrichment assays[3]. No difference between the biobanked and fresh samples was observed (Fig. 1e), suggesting identical signal to noise ratios and that tagmentation – the process of tagging and fragmenting accessible chromatin during ATAC-seq – occurred primarily at focal accessible regions. The correlation of accessibility levels in peaks identified across all samples indicated an indistinguishable accessible chromatin landscape (Fig. 1f). Furthermore, the overlap of peaks across a wide range of genomic annotations indicated that either biobanking or reducing the starting cell number did not bias the discovery of certain genomic features (Fig. 1g). Also, the distribution of accessible intergenic, intronic, and promoter regions discovered by ATAC-seq was consistent with previous reports[4]. Finally, the correlation of peak signals both within and between fresh and biobanked samples was high across the large range of starting material (Fig. 1h), indicating that ATAC-seq on biobanked specimens accurately recapitulated that of fresh samples. For example, the major histocompatibility complex class II locus (MHC-II) is actively transcribed in human B cells. The *HLA-DRB1* and *HLA-DQA1* promoters were identified as accessible, as well as intergenic regions representing the XL9 insulator element[5] and CIITA binding sites[6] in a region classified as a super enhancer[7] (Fig. 1i). These data show that chromatin accessibility patterns were preserved during biobanking.

During ATAC-seq tagmentation, distinct periodic patterns of chromatin fragmentation are observed as nucleosomes and DNA-binding proteins protect DNA from transposition events[2]. Although the distribution was distinct, the pattern of sequencing read fragment sizes was similar for both fresh and biobanked samples (Fig. 2a). Sequencing reads representing intra-nucleosomal ($<$150 bp) and di-nucleosomal (260–340 bp) fragments were separated and analyzed for their unique distribution pattern at genomic features. The distribution of intra-nucleosomal reads at all human RefSeq transcription start sites (TSS) showed a single peak of enrichment at the nucleosome free region (Fig. 2b). Conversely, di-nucleosomal reads displayed a periodicity surrounding the TSS, identifying the position of the upstream and downstream positioned nucleosomes (Fig. 2c), and indicating that the biobanking process had maintained TSS chromatin structure.

The footprint of mammalian transcription factors were plotted to determine if biobanking affected the ability to resolve the accessibility patterns of DNA-binding proteins. The pattern of intra-nucleosomal and di-nucleosomal reads was computed surrounding the positions of CCCTC binding factor (CTCF) binding motifs calculated from ENCODE data profiling the GM12878 lymphoblastoid cell line[8]. Intra-nucleosomal reads displayed enrichment that peaked at the motif boundaries, identifying the protected footprint where CTCF contacts DNA (Fig. 2d). In contrast, di-nucleosomal reads weakly showed the protected footprint and further identified two additional enriched regions 200 bp surrounding the motif (Fig. 2e). These patterns are similar to the locations of positioned nucleosomes surrounding CTCF binding sites[9]. Additionally, similar transcription factor accessibility footprint patterns were observed at the sequence motifs for other important B cell factors: RFX5, NFYB, CREB1, and PU.1 (Fig. 2f,g). Minimal differences in overall accessibility were observed between fresh and biobanked samples, but this did not influence the ability to observe discrete footprints. Importantly, the distribution of intra-nucleosomal and di-nucleosomal reads surrounding the TSS and transcription factor binding sites were identical in biobanked and fresh samples, indicating biobanking had no global effect on protein-DNA interactions.

Naïve SLE B cells exhibit a unique chromatin architecture. SLE is characterized by increases in autoreactive B cell subsets[10–13]. Genetic predispositions have been identified but there is a strong implication for an epigenetic component that contributes to disease etiology[14,15]. Interestingly, many disease susceptibility polymorphisms, including causal ones, occur in B cell signaling pathways[16,17] and frequently map to non-coding regulatory regions[18]. Recent data revealed that naïve B cells form an underappreciated component of active disease flares[11], suggesting B cells harbor pathogenic alterations at an early stage. Therefore, it was hypothesized that an altered epigenetic program was present in naïve SLE B cells. To test this hypothesis, the ATAC-seq assay was applied to samples isolated from a biorepository of SLE patients undergoing disease flares. Three SLE samples biobanked for two years were processed in combination with one freshly obtained SLE sample. As a comparison, four healthy control (HC) patients were recruited as controls. No difference was observed in the cellular viability post-thawing of the biobanked samples compared to the fresh samples (Fig. 3a) or in the frequency of naïve $CD19^+$ B cells (Fig. 3b). Naïve $CD19^+$ naïve B cells were FACS isolated (Supplementary Fig. S1C) and the accessible chromatin landscape determined by ATAC-seq for each sample. All samples were highly similar with respect to the fragment size distribution of sequencing reads and the number of peaks identified (Fig. 3c).

Differentially accessible regions between SLE and HC were identified and 602 loci demonstrated significant increases in accessibility in SLE B cells while 461 loci were more accessible in HC B cells (Fig. 3d). Differentially accessible loci mapped to 988 distinct genes, including 66 genes that contained more than one differential region. Of these genes, 98% (65/66) displayed concordant changes in accessibility, suggesting coordinated changes in accessibility of potential *cis*-regulatory elements associated with disease. To define the function of SLE or HC specific accessibility changes, differential loci were annotated to the nearest gene and ontology analysis performed

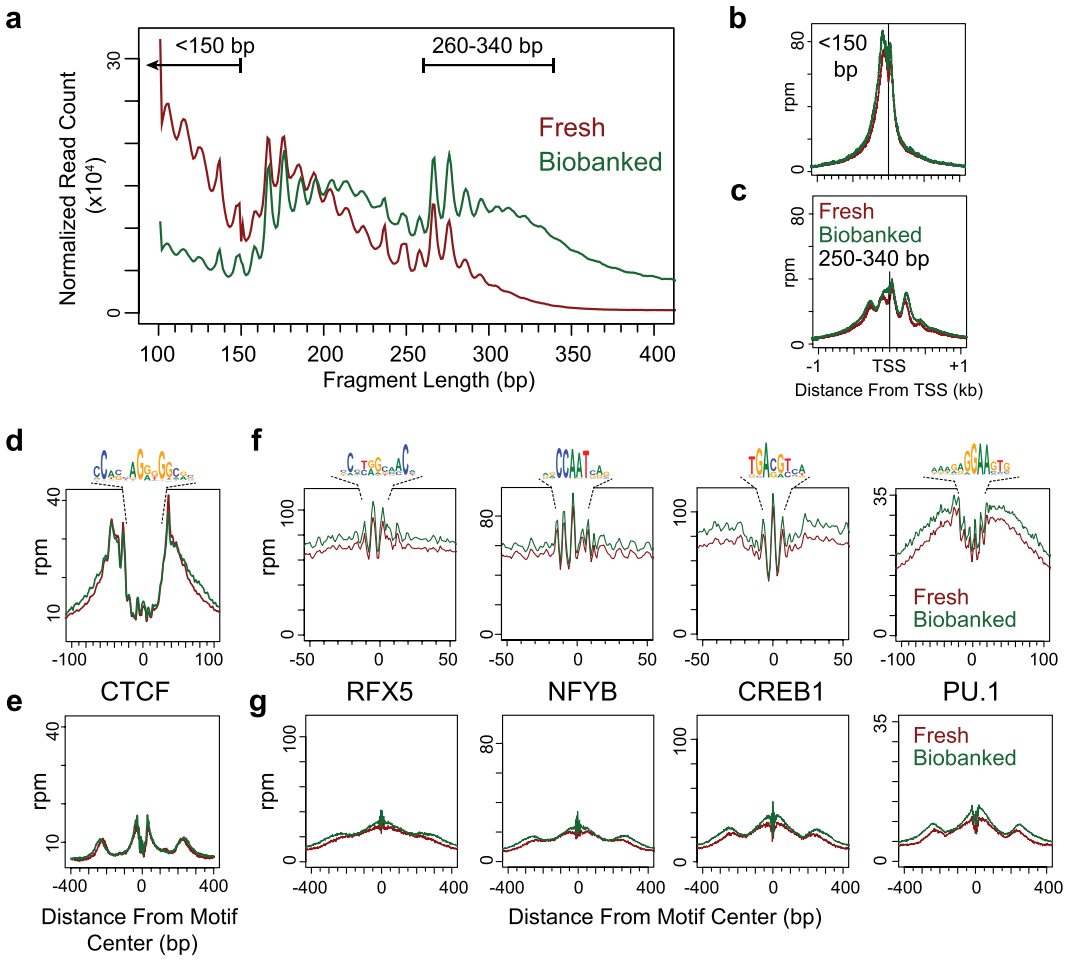

Figure 2. Biobanking preserves protein-DNA interaction structure. (**a**) Histogram of the distribution of fragment lengths in reads from all fresh or biobanked samples. The enriched regions of sub-nucleosomal (<150 bp) and di-nucleosomal (260–340) are indicated. Histograms of fresh and biobanked reads separated by fragment lengths of (**b**) <150 bp and (**c**) 260–340 bp at all hg19 RefSeq transcription start sites (TSS). The vertical bar indicates the position of the TSS. Histograms of fresh and biobanked reads were separated by fragment length of (**d**) <150 bp and (**e**) 260–340 bp at 56,208 CTCF motifs. The CTCF motif used for the analysis is shown above the footprint. (**f**) Histogram comparing fragments corresponding to sub-nucleosomal lengths from fresh and biobanked samples at 11,318 RFX5, 18,094 NYFB, 12,115 CREB1, and 56,420 PU.1 motifs. The motif used for each analysis is indicated. (**g**) Histogram comparing fragments corresponding to di-nucleosomal reads from fresh and biobanked samples at the transcription factor motif locations described in D.

to identify enriched biological processes. Increases in accessibility in SLE B cells were associated with leukocyte differentiation, cellular activation, and B cell activation while HC accessible loci were enriched for processes associated with transcriptional regulation (Fig. 3e). To gain insight into the potential signaling networks programming accessibility changes, the transcription factor motifs enriched in the SLE and HC specific accessible regions were determined. Loci with increased accessibility in HC contained motifs for NRF1, CTCF and STAT5 (Fig. 3f). Contrastingly, SLE specific accessible loci displayed enrichment for transcription factors involved in B cell activation such as NFKB, AP-1, and BATF, as well as B cell differentiation factors IRF4 and PRDM1. The enrichment of these motifs in differentially accessible loci suggests that the binding of NRF1 and BATF impact local chromatin accessibility in HC and SLE, respectively. Indeed, compared to the naïve HC B cells, those from SLE patients displayed increased accessibility in the 200 bp surrounding BATF motifs present at all accessible loci within the genome (Fig. 3g). Conversely, HC B cells demonstrated increased accessibility at NRF1 motifs. No difference in accessibility was observed for a control motif, PAX5, which was not enriched in SLE or HC. These data therefore identified an activation signature in SLE B cells that is manifested in changes in chromatin accessibility surrounding specific transcription factor binding motifs.

Examples of differentially accessible loci include the *STAT4* promoter, which demonstrated higher accessibility in SLE B cells (Fig. 3h). Polymorphisms in *STAT4* are highly associated with autoantibody production[19] and changes in the promoter accessibility of *STAT4* could result from higher IFN-alpha signaling in SLE patients[20] or suggests that SLE B cells are epigenetically predisposed for activation of the *STAT4* pathway. HC specific accessible loci were primarily associated with genes involved in transcriptional regulation. Among the transcriptional

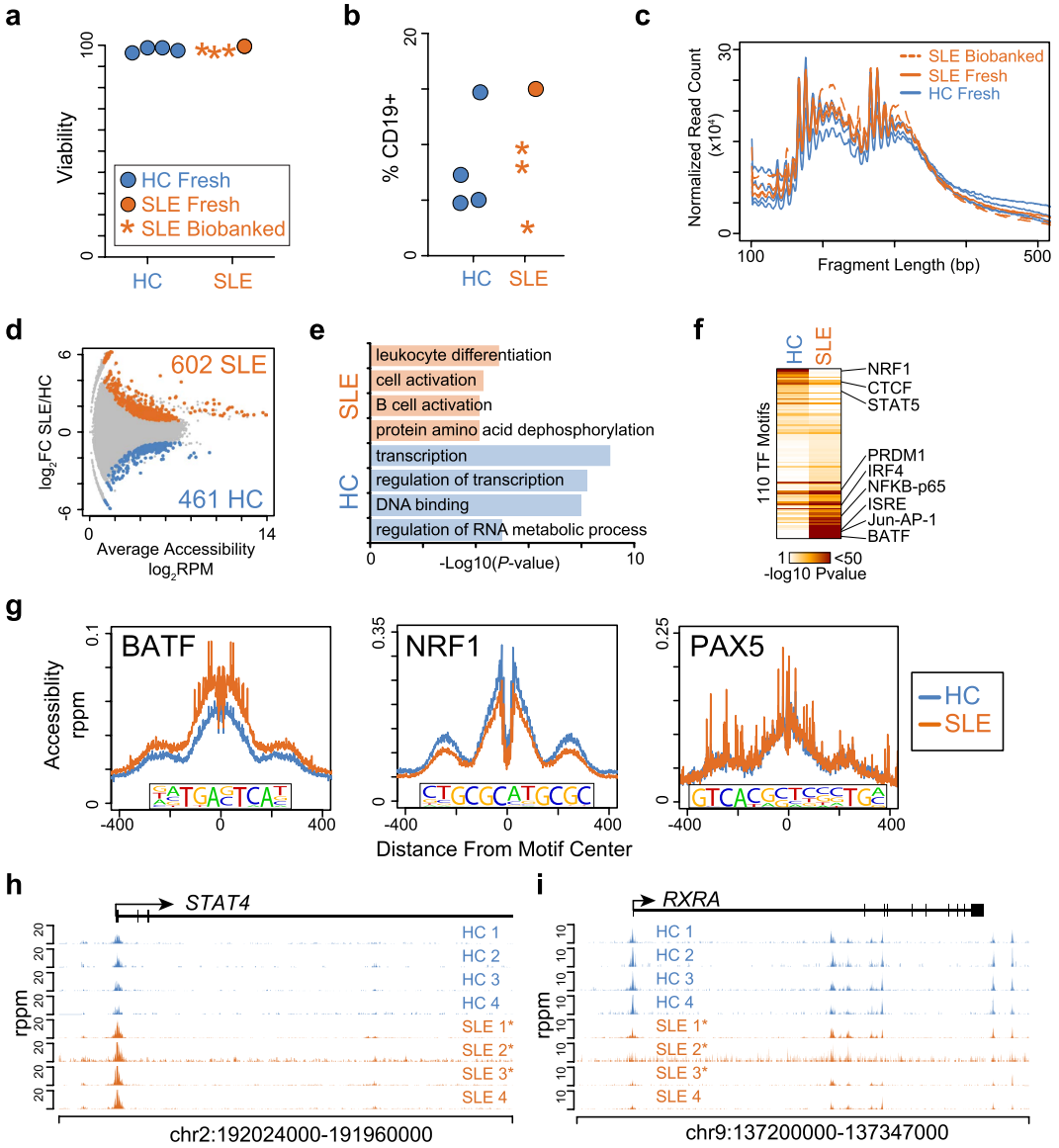

Figure 3. SLE B cells display an altered chromatin accessibility profile. (**a**) Summary of the percentage of viable cells for HC and SLE samples. *P*-value = 0.98. (**b**) Dot plot of the percentage of $CD19^{+}$ naïve B cells for HC and SLE samples. *P*-value = 0.81. (**c**) Histogram of the paired-end fragment lengths of HC and SLE samples. (**d**) Scatter plot of the average accessibility at each peak in HC and SLE versus the log fold change (logFC) in accessibility. Accessible loci that are significantly differentially accessible (FDR <0.05) are highlighted in blue (HC) or orange (SLE) with the number of loci indicated. (**e**) Bar plot of GO Biological Processes enriched in SLE or HC accessible loci. (**f**) Heatmap showing the significance of 110 transcription factor motifs enriched in HC and SLE accessible loci. Motifs are sorted from the most enriched in HC to the most enriched in SLE. The locations of select motifs are highlighted. (**g**) Histogram of the accessibility at 800 bp surrounding BATF, NRF1, and PAX5 motifs identified in all accessible loci in HC and SLE. The motif identified is indicated below each histogram. Data are normalized to reads per peak per million (rppm) as described by equation 3 in Methods. Genome plot depicting the ATAC-seq profiles for HC and SLE samples at the *STAT4* (**h**) and *RXRA* (**i**) genomic loci. The positions of each gene, direction of transcription, and exon locations are indicated. *Indicates biobanked SLE samples.

regulators with increased accessibility in HC B cells was *RXRA* (Fig. 3i). Mice deficient in *RXRA* display increased antibodies to nuclear antigens[21]. Thus, disease-specific changes identified in the accessible chromatin landscape indicate that SLE B cells are epigenetically distinct from HC.

Discussion

Biobanking is routinely used to store clinical samples for future experiments. For long-term studies, biobanking offers significant experimental advantages in that samples can be stored and processed in bulk, thereby reducing technical batch effects due to sample preparation. Additionally, preexisting biorepositories provide access to a vast and diverse number of specimens, thereby avoiding lengthy sample collection studies

and allowing selection of specimens based on outcome data. Rigid biobanking practices are important for long-term sample preservation at the cellular phenotypic and molecular level. Metrics that measure both cellular viability and complexity are important criteria for evaluating biobanking protocols. The data presented herein suggest that measuring chromatin accessibility may be an important molecular metric for determining biobanking success.

Here the framework for applying the ATAC-seq assay to biobanked specimens is presented and was applied to a repository of PBMCs biobanked from SLE patients undergoing disease flares. To gain insight into the epigenetic programming of SLE, ATAC-seq was performed on $CD19^+$ naïve B cells from SLE and HC subjects. A unique pattern was observed that indicated increases in genomic accessibility occur both surrounding genes involved in B cell activation and the transcription factor binding sites that regulate B cell activation and differentiation. The transcription factor BATF in particular has an emerging role in B cell activation and function[22], including direct transcriptional activation of IgM and AID[23,24]. Additionally, BATF motifs were previously discovered to be overrepresented in the promoters of autoimmunity susceptibility genes[25]. These data, along with the presence of increased accessibility surrounding BATF motifs in SLE, suggests a previously unknown role for BATF in the etiology of SLE B cells. Currently it is not known if the accessibility signature is an intrinsic feature of SLE B cells or due to external environmental stimuli that results in the activation of signaling networks that drive changes in accessibility. Nevertheless, the finding that alterations in the epigenome converge with genetic data[17] further pinpoint B cell activation as a key dysregulated process in SLE.

The data presented here demonstrate the feasibility of determining the accessible chromatin landscape from biobanked samples. Mechanistically, ATAC-seq facilitates the identification of *cis*-regulatory elements. In addition to the network analyses presented here, ATAC-seq has the potential to impact traditional GWAS studies that seek to relate non-coding, intergenic disease associations to regulatory elements. Therefore, chromatin accessibility profiling is a powerful addition to the toolbox of assays that can be applied to biobanked specimens.

Methods

Biobanking. *PBMC isolation.* Samples were obtained with informed consent in accordance with protocols approved by the Emory University School of Medicine Institutional Review Board. SLE donors fulfilled >4 revised ACR criteria for the classification of SLE[26]. PBMCs were isolated from healthy control (HC) or SLE donors by centrifugation at 1,500× G for 25 min at room temperature (RT) using cell preparation tubes (CPT) containing sodium heparin and Ficoll-Hypaque solution. The plasma layer was removed from CPTs, and PBMCs were transferred into a 50 ml conical tube and topped off to a final volume of 50 ml with PBS (Cellgro). PBMCs were pelleted by centrifugation at 1,300 RPM at RT for 10 min, and PBS was aspirated off of cell pellet. PBMCs were resuspended in 50 mL of PBS and spun at 1,300 RPM at RT for 10 min for a total of 3 washes.

Biobanking of total PBMCs. A biobanking protocol was developed that allowed storage and distribution of samples for the Autoimmunity Centers of Excellence program. Following PBMC isolation, samples to be frozen were slowly resuspended in 1 ml 4 °C freezing medium (heat-inactivated, filtered FBS containing 10% DMSO) at a concentration of 10 million cells/ml, placed into a RT slow-freeze container, moved to −80 °C overnight, and then placed in liquid nitrogen for long-term storage.

Thawing of total PBMCs. Frozen total PBMCs were removed from liquid nitrogen and placed into a 37 °C water bath until thawed (less than 2 minutes). Thawed PBMCs were transferred to a 15 ml conical tube and diluted with PBS drop-wise to 10 ml. The 15 ml conical tube was inverted to mix and spun at 1,300 RPM for 10 min at RT. Freezing media and PBS were aspirated off of the cell pellet. Cells were resuspended in 10 ml of PBS and spun at 1,300 RPM for 10 min at RT. PBMCs were then subjected to FACS.

PBMC staining and FACS sorting of $CD19^+$ B cells. Freshly isolated or thawed total PBMCs were pulsed with 20 nM of MitoTracker Green (Invitrogen, Inc.) in pre-warmed complete media (RPMI 1640 supplemented with 10% FBS and 1% L-glutamine) at 37 °C for 30 min. Cells were pelleted at 1,300 RPM for 10 min at RT, resuspended, and chased with 10 ml of pre-warmed complete media for 30 min at 37 °C. Cells were again spun at 1,300 RPM for 10 min, resuspended at 10^7 cells per 100 μl of PBS containing 0.5% BSA, 5% normal mouse serum, and 5% normal rat serum, and stained for flow cytometry with the following fluorochrome conjugated mouse anti-human monoclonal antibodies: anti-CD3, anti-CD24 (Invitrogen, Inc.), anti-CD19, anti-IgD, anti-CD27, anti-CD38 (BD Biosciences, Inc.). Analysis was performed using a BD LSRII. Sorting was performed on a FACS Aria II. Prior to each sort the FACS AriaII was calibrated with fluorescent beads to achieve a >99% sort purity.

ATAC-seq. ATAC-seq was performed as described previously[6] with adaptations[2,27]. HC and SLE $CD19^+$ naïve B cells were sorted into FACS buffer (PBS containing 10% FBS) and cells centrifuged at 500× g at 4 °C for 10 min. Cells were resuspended in 50 μl of Nuclei Lysis buffer (10 mM Tris-HCl [pH 7.4], 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630, molecular grade H_2O, filter sterilized) and immediately centrifuged at 500× g at 4 °C for 30 min. The supernatant was removed and nuclei resuspended in 25 μl of Tagmentation Reaction mix (2X TD buffer, 1 μl Tagmentation enzyme, molecular grade H_2O, (Illumina, Inc.)). Tagmentation reaction was incubated for 60 min at 37 °C. The tagmentation reaction was stopped with 23 μl of Tagmentation Clean-up buffer (326 mM NaCl, 109 mM EDTA, 0.63% SDS) and 2 μl of 10 mg/ml Proteinase-K and incubated for 30 min at 40 °C. DNA was isolated following a negative SPRI-size selection of 0.3× followed by a positive SPRI-selection of 1.2×. Tagmented DNA was eluted in 28 μl of Tris-HCl (pH 8.0).

PCR amplification was performed by combining 28 μl of tagmented DNA with 5 μl each of i5 and i7 dual indexing primers (Nextera Indexing Kit, Illumina, Inc.), 10 μl of 5× KAPA HiFi GC Buffer with $MgCl_2$ (KAPA Biosystems),

1 μl of 10 mM dNTPs, and 1 μl of KAPA HiFi HotStart Polymerase (KAPA Biosystems). An initial amplification was performed using the following cycle conditions: 1 cycle of 72 °C for 3 min, 5 cycles of 98 °C for 10 sec, 63 °C for 30 sec, and 72 °C for 30 sec. To estimate the number of PCR cycles required, 2 μl of each reaction was diluted 1:100 with 0.05% Tween-20 and quantitated using the KAPA Library Quantification qPCR Kit (KAPA Biosystems) according to the manufacturer's protocol. The number of additional PCR cycles needed was determined by equation (1):

$$C_{Target} = C_{0.5} + \log_2\left(\frac{Vol_{Quant}}{Dilution_{Quant}}\right) - \log_2(Vol_{MaxAmp}) \quad (1)$$

where C_{Target} is the target number of PCR cycles; $C_{0.5}$ is the cycle number at half maximum fluorescence intensity; $Dilution_{Quant}$ is the dilution of material added to quant; Vol_{Quant} is the volume of the library added to quant PCR number of PCR cycles; and Vol_{MaxAmp} is the maximum volume of the undiluted library to be added to the amplification PCR.

The total number of amplification cycles was normalized between samples by setting C_{Target} to the maximum C_{Target} value and calculating the volume of each sample, Vol_{Amp}, to add to the reaction using equation (2):

$$Vol_{Amp} = 2^{(C_{Target} - C_{MaxTarget})} \times Vol_{MaxAmp} \quad (2)$$

where $C_{MaxTarget}$ is the maximum C_{Target} value of all samples. PCR amplification was completed using 'C_{Target}' as calculated above: C_{Target} cycles of 98 °C for 10 sec, 63 °C for 30 sec, 72 °C 30 sec, and 1 cycle of 72 °C for 60 sec.

To remove primers and high molecular weight DNA following PCR amplification, a dual SPRI-size selection was performed with a 0.2× initial selection and a 0.8× final isolation. Amplified library was eluted in 25 μl Tris-HCl (pH 8.0) and quality checked on a Bioanalyzer and qPCR quantitated using the KAPA Library Quantification qPCR Kit (KAPA Biosystems). Libraries were pooled at equimolar ratios and sequenced on a HiSeq2500 using 50 bp paired-end Illumina chemistry.

ATAC-seq Data Processing. All ATAC-seq data has been deposited in the NCBI GEO database under accession GSE71338. Raw fastq reads were checked for nucleotide distribution and read quality using the FASTX-toolkit and mapped the hg19 version of the human genome using Bowtie[28] and the default settings. Only uniquely mapped and non-redundant reads were used for downstream analyses. Peak calling was performed with HOMER software[29] and the '-style dnase' setting. All peaks between fresh and biobanked samples are listed in Supplementary Table S1. Reads per peak per million (rppm) normalization on HC and SLE samples was performed by equation (3):

$$rppm = reads \times \left(\frac{1 \times 10^6}{UniqueReads \times FRiP}\right) \quad (3)$$

Differential Accessibility Analysis. Significantly different accessible regions between HC and SLE B cells were determined by computing the overlap of all HC and SLE peaks using the HOMER 'mergePeaks' function. The raw, non-normalized reads from each sample were annotated for each peak using the 'annotatePeaks.pl' script with the following options '-size given –noadj'. The resulting composite peak file was used as input for edgeR[30] using the 'getDiffExpression.pl' HOMER script with the following options '-peaks HC HC HC HC SLE SLE SLE SLE'. Peaks with an FDR <0.05 were considered significantly differentially accessible between HC and SLE B cells. All significant peaks are listed in Supplementary Table S2. Differentially accessible peaks were annotated to the nearest gene using the 'annotatePeaks.pl' HOMER script and motif enrichment calculated using the 'findMotifsGenome.pl' script. Ontology analysis of genes with increases and decreases in accessibility between HC and SLE was performed using DAVID[31,32].

Fragment Length Analysis. Bam files were parsed and the fragment length analyzed using the Genomic Alignments[33] R/Bioconductor package. Qnames corresponding to reads with a fragment length of <150 bp or between 250 and 340 bp were extracted. The bam files were converted to sam files using the samtools package[34,35] and parsed for reads with desired fragment lengths based on extracted Qnames with custom python scripts. Fragment-length specific sam files were used as input for HOMER to generate tag directories using the 'makeTagDirectory –format sam' script. All custom scripts are available upon request.

Motif Histograms. The 'wgEncodeRegTfbsClusteredV3.bed.gz' file[8,36,37] was downloaded from the UCSC Genome Browser and binding sites for CTCF, PU.1, RFX5, NFYB, CREB1 extracted using custom Perl scripts. Motif positions in peaks were identified using FIMO[38] and position weight matrices for each transcription factor acquired from the JASPAR database[39]. The highest scoring motif in each peak was chosen for further analyses. Motif coordinates were used as input for the HOMER findPeaks.pl script using the '-size 1000 –fragLength 1 –norm 1e6 –hist' options.

References

1. Yong, W. H., Dry, S. M. & Shabihkhani, M. A practical approach to clinical and research biobanking. *Methods Mol Biol* **1180,** 137–162, doi: 10.1007/978-1-4939-1050-2_8 (2014).
2. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10,** 1213–1218, doi: 10.1038/nmeth.2688 (2013).
3. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22,** 1813–1831, doi: 10.1101/gr.136184.111 (2012).

4. Davie, K. *et al.* Discovery of transcription factors and regulatory regions driving *in vivo* tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS genetics* **11,** e1004994, doi: 10.1371/journal.pgen.1004994 (2015).
5. Majumder, P., Gomez, J. A., Chadwick, B. P. & Boss, J. M. The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *J Exp Med* **205,** 785–798 (2008).
6. Scharer, C. D. *et al.* Genome-wide CIITA-binding profile identifies sequence preferences that dictate function versus recruitment. *Nucleic Acids Res* **43,** 3128–3142, doi: 10.1093/nar/gkv182 (2015).
7. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319, doi: 10.1016/j.cell.2013.03.035 (2013).
8. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22,** 1798–1812, doi: 10.1101/gr.139105.112 (2012).
9. Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS genetics* **8,** e1003036, doi: 10.1371/journal.pgen.1003036 (2012).
10. Wei, C. *et al.* A new population of cells lacking expression of CD27 represents a notable component of the B cell memory compartment in systemic lupus erythematosus. *J Immunol* **178,** 6624–6633 (2007).
11. Tipton, C. M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat Immunol* **16,** 755–765, doi: 10.1038/ni.3175 (2015).
12. Dorner, T., Jacobi, A. M., Lee, J. & Lipsky, P. E. Abnormalities of B cell subsets in patients with systemic lupus erythematosus. *J Immunol Methods* **363,** 187–197, doi: 10.1016/j.jim.2010.06.009 (2011).
13. Cappione, A., 3rd *et al.* Germinal center exclusion of autoreactive B cells is defective in human systemic lupus erythematosus. *J Clin Invest* **115,** 3205–3216, doi: 10.1172/JCI24179 (2005).
14. Javierre, B. M. *et al.* Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res* **20,** 170–179, doi: 10.1101/gr.100289.109 (2010).
15. Absher, D. M. *et al.* Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to $CD4^+$ T-cell populations. *PLoS genetics* **9,** e1003678, doi: 10.1371/journal.pgen.1003678 (2013).
16. Cambier, J. C. Autoimmunity risk alleles: hotspots in B cell regulatory signaling pathways. *J Clin Invest* **123,** 1928–1931, doi: 10.1172/JCI69289 (2013).
17. 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, doi: 10.1189/jlb.0212095 (2012).
18. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343, doi: 10.1038/nature13835 (2015).
19. Chung, S. A. *et al.* Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS genetics* **7,** e1001323, doi: 10.1371/journal.pgen.1001323 (2011).
20. Kariuki, S. N. *et al.* Cutting edge: autoimmune disease risk variant of *STAT4* confers increased sensitivity to IFN-alpha in lupus patients *in vivo*. *J Immunol* **182,** 34–38 (2009).
21. Roszer, T. *et al.* Autoimmune kidney disease and impaired engulfment of apoptotic cells in mice with macrophage peroxisome proliferator-activated receptor gamma or retinoid X receptor alpha deficiency. *J Immunol* **186,** 621–631, doi: 10.4049/jimmunol.1002230 (2011).
22. Murphy, T. L., Tussiwand, R. & Murphy, K. M. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nat Rev Immunol* **13,** 499–509, doi: 10.1038/nri3470 (2013).
23. Ise, W. *et al.* The transcription factor BATF controls the global regulators of class-switch recombination in both B cells and T cells. *Nat Immunol* **12,** 536–543, doi: 10.1038/ni.2037 (2011).
24. Betz, B. C. *et al.* Batf coordinates multiple aspects of B and T cell function required for normal antibody responses. *J Exp Med* **207,** 933–942, doi: 10.1084/jem.20091548 (2010).
25. Dozmorov, M. G., Wren, J. D. & Alarcon-Riquelme, M. E. Epigenomic elements enriched in the promoters of autoimmunity susceptibility genes. *Epigenetics* **9,** 276–285, doi: 10.4161/epi.27021 (2014).
26. Hochberg, M. C. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* **40,** 1725, doi: 10.1002/1529-0131(199709)40:9<1725::AID-ART29>3.0.CO;2-Y (1997).
27. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345,** 943–949, doi: 10.1126/science.1256271 (2014).
28. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25, doi: 10.1186/gb-2009-10-3-r25 (2009).
29. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38,** 576–589, doi: 10.1016/j.molcel.2010.05.004 (2010).
30. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26,** 139–140, doi: 10.1093/bioinformatics/btp616 (2010).
31. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37,** 1–13, doi: 10.1093/nar/gkn923 (2009).
32. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57, doi: 10.1038/nprot.2008.211 (2009).
33. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9,** doi: 10.1371/journal.pcbi.1003118 (2013).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079, doi: 10.1093/bioinformatics/btp352 (2009).
35. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27,** 2987–2993, doi: 10.1093/bioinformatics/btr509 (2011).
36. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100, doi: 10.1038/nature11245 (2012).
37. Wang, J. *et al.* Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* **41,** D171–176, doi: 10.1093/nar/gks1221 (2013).
38. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27,** 1017–1018, doi: 10.1093/bioinformatics/btr064 (2011).
39. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32,** D91–94, doi: 10.1093/nar/gkh012 (2004).

Acknowledgements

We thank the Emory Integrated Genomics Core for Bioanalyzer expertise, the Emory Flow Cytometry Core for cell sorting and analysis, and the NYU Genome Technology Center for Illumina sequencing. Grants from the National Institutes of Health. B.G.B was supported by F31AI112261 and previously by T32GM008490. R.R.H was supported by T32GM008490. This work was funded by U19AI110483 to J.M.B and I.S. and RO1GM47310 to J.M.B.

Author Contributions

C.D.S. and E.L.B. performed the experiments and interpreted the data. C.W. developed the biobanking protocol. C.D.S., B.G.B. and R.R.H. performed the data analysis. I.S. and J.M.B. supervised the experiments and contributed to the interpretation of data. All authors contributed to the writing of the manuscript.

Additional Information

Accession codes: NCBI Gene Expression Omnibus: sequencing data are available under the accession number GSE71338.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Scharer, C. D. *et al.* ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Sci. Rep.* **6**, 27030; doi: 10.1038/srep27030 (2016).