High throughput characterization of genetic effects on DNA:protein binding and gene transcription

Cynthia A. Kalita [1], Christopher D. Brown [2], Andrew Freiman [1], Jenna Isherwood [1], Xiaoquan Wen[3], Roger Pique-Regi [1,4,*], Francesca Luca [1,4,*]

[1]Center for Molecular Medicine and Genetics, Wayne State University

[2]Department of Genetics, University of Pennsylvania

[3]Department of Biostatistics, University of Michigan

[4]Department of Obstetrics and Gynecology, Wayne State University

[*]To whom correspondence should be addressed: rpique@wayne.edu, fluca@wayne.edu.

Many variants associated with complex traits are in non-coding regions, and contribute to phenotypes by disrupting regulatory sequences. To characterize these variants, we developed a streamlined protocol for a high-throughput reporter assay, BiT-STARR-seq (Biallelic Targeted STARR-seq), that identifies allele-specific expression (ASE) while accounting for PCR duplicates through unique molecular identifiers. We tested 75,501 oligos (43,500 SNPs) and identified 2,720 SNPs with significant ASE (FDR 10%). To validate disruption of binding as one of the mechanisms underlying ASE, we performed a high throughput binding assay for NFKB-p50. We identified 2,951 SNPs with allele-specific binding (ASB) (FDR 10%); 173 of these SNPs also had ASE (OR=1.97, p-value=0.0006). Of variants associated with complex traits, 1,531 resulted in ASE and 1,662 showed ASB. For example, we characterized that the Crohn's disease risk variant for rs3810936 increases NFKB binding and results in altered gene expression.

Genome wide association studies (GWAS) have identified thousands of common genetic variants associated with complex traits, including normal traits and common diseases. Many GWAS hits are in non-coding regions, so the underlying mechanism leading to specific phenotypes is likely through disruption of gene regulatory sequence. Quantitative trait loci (QTLs) for molecular and cellular phenotypes [1], such as gene expression (eQTL) [2, 3, 4, 5, 6], transcription factor binding [7], and DNaseI sensitivity (dsQTL) [8] have been crucial in providing strong evidence and a better understanding of how genetic variants in regulatory sequences can affect gene expression levels [9, 6, 10, 11]. In recent work, we were able to validate 48% of computationally predicted allelic effects on transcription factor binding through traditional reporter assays [12]. However, traditional reporter assays are limited by the time and the cost of testing variants one at a time.

Massively parallel reporter assays (MPRA) have been developed for the simultaneous measurement of the regulatory function of thousands of constructs at once. For MPRA, a pool of synthesized DNA oligos containing a barcode at the 3'UTR of a reporter plasmid is transfected into cells, and transcripts are isolated for RNA-seq. The number of barcode reads in the RNA over the number of barcode reads from the plasmid DNA is used as a quantitative measure of expression driven by the synthesized enhancer region [13, 14, 15, 16, 17]. An alternative to MPRA is STARR-seq (self-transcribing active regulatory region sequencing) [18], whose methods involve fragmenting the genome and cloning the fragments 3'of the reporter gene. The approach is based on the concept that enhancers can function independently of their relative positions, so putative enhancers are placed downstream of a minimal promoter. Active enhancers transcribe themselves, with their strength quantified as the amount of RNA transcripts within the cell. Because they do not use separate barcodes, STARR-seq approaches have streamlined protocols that allow for higher throughput.

Recently, high-throughput assays have been used to assess the enhancer function of genomic regions [18, 19], the allelic effects on gene expression for naturally occurring variation in 104 regulatory regions [20], fine-map variants associated with gene expression in lymphoblastoid cell lines (LCLs) and HepG2 [21], and fine-map variants associated with red blood cell traits in GWAS [22]. In addition to using reporter assays to measure enhancer function on gene expression, there are several methods to directly measure binding affinity of DNA sequences for specific transcription factors. These methods include Spec-seq [23], EMSA-seq (electrophoretic mobility shift assay-sequencing) [24], and BUNDLE-seq (Binding to Designed Library, Extracting, and sequencing) [25]. In these assays, synthesized regions are combined *in vitro* with a purified transcription factor. The bound DNA-factor complexes are then isolated by polyacrylamide gel electrophoresis (PAGE). Extracting the DNA from upper (bound complex) and lower (unbound DNA) bands and sequencing of the derived libraries allows for quantification of the binding strength of regulatory regions. While BUNDLE-seq compared binding and reporter gene expression, and EMSA has been previously used to ascertain allelic effects, none of the high-throughput EMSA methods have been previously used to determine allelic effects on binding.

We have developed a method called BiT-STARR-seq (Biallelic Targeted STARR-seq) to

test for allele specific effects in regulatory regions (Figure 1a). BiT-STARR-seq applies the streamlined protocol of STARR-seq to thousands of synthesized oligos targeting independent genomic regions, to create the simplest experimental protocol for high-throughput reporter assays to date. The method also includes the incorporation of unique molecular identifiers (UMIs) during cDNA synthesis, that allows for the removal of duplicates created during library preparation. We used BiT-STARR-seq to test 43,500 regulatory variants, including variants predicted to disrupt transcription factor binding (CentiSNPs [12]) for 874 transcription factors, as well as other regulatory variants [5, 4, 26, 12]. We then adapted BUNDLE-seq to analyze allele-specific binding (ASB) for NFKB (p50) and validate the molecular mechanism underlying the allele-specific effects measured in the BiT-STARR-seq assay. We denote this new method BiT-BUNDLE-seq (Biallelic Targeted BUNDLE-seq). To the best of our knowledge, this is the first use of any high-throughput EMSA to consider allele-specific binding for regulatory regions. Our results demonstrate that high-throughput EMSA approaches complement allele-specific analyses in MPRA/STARR-seq assays, thus providing an effective strategy to dissect the molecular mechanism linking regulatory variants effects on binding and on expression. Our method is especially well suited to test in parallel thousands of computationally prioritized variants that are also associated with complex traits.

We selected different categories of regulatory variants for this study including eQTLs [5, 4], CentiSNPs [12], ASB SNPs [12], variants associated with complex traits in GWAS [26], and negative ASB controls [12] for a total of 50,609 SNPs. We designed two oligos targeting each of the alleles for a SNP, with inserts 230bp long synthesized by Agilent to contain the regulatory region and the SNP within the first 150bp. We also included the use of unique molecular identifiers (UMIs), added during cDNA synthesis. With these random UMIs we are in effect tagging identifiable replicates of the self-transcribing construct, which improves the analysis of the data by accounting for PCR duplicates. Our protocol also has the advantage of being highly streamlined. Unlike STARR-seq, our method does not require preparation of DNA regions for use in the assay, such as whole genome fragmentation [18], or targeting regions [19, 27], while, similar to STARR-seq, it requires only a single cloning and transformation step. Because the UMIs are inserted after transfection, there are no additional bottleneck issues (due to library complexity) in the cloning and transformation steps.

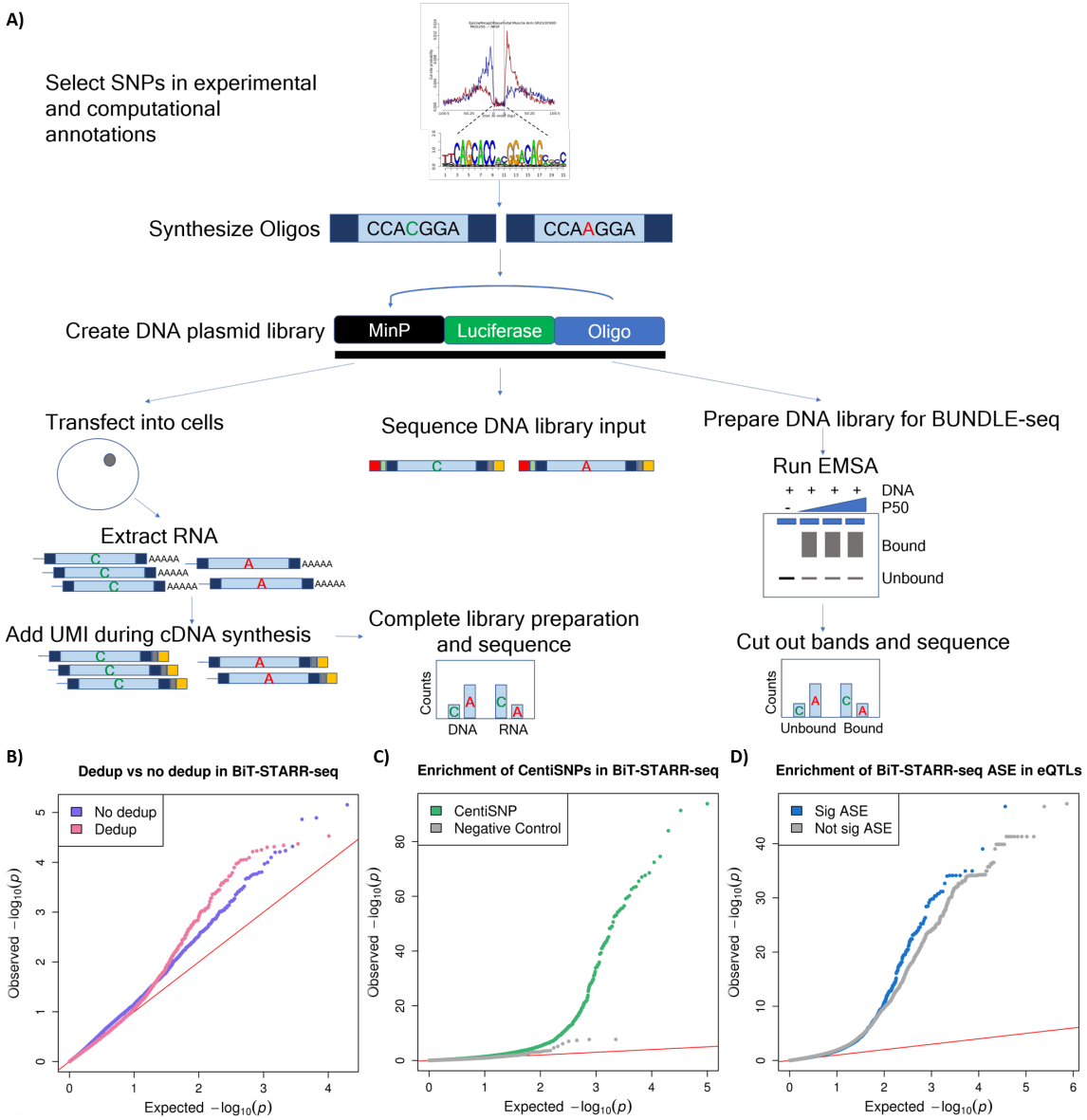

Figure 1: **BiT-STARR-seq and BiT-BUNDLE-seq identify regulatory variants in non-coding regions.** A) Experimental outline. Oligos targeting the regulatory regions of interest (and either reference or alternate alleles) are designed to contain, on their ends, 15bp matching the sequencing primers used for Illumina NGS. The DNA library is used both in the BiT-STARR-seq and BiT-BUNDLE-seq experiments. UMIs are added during cDNA synthesis for the BiT-STARR-seq RNA-seq library and prior to PAGE in the BiT-BUNDLE-seq protocol. B) QQplot depicting the p-value distributions from QuASAR-MPRA for a single experimental replicate processed without removing duplicates (purple) or after removing duplicates using the UMIs (pink). C) QQplot depicting the p-value distributions from the ASE test performed using QuASAR-MPRA on all replicates after removing duplicates. CentiSNPs are in (green)[12] while SNPs in the negative control group are in (grey). D) QQplot depicting the p-value distributions for eQTLs from [28]. SNPs with significant ASE (FDR 10%) are in (blue) or not significant ASE are in (grey).

We generated 7 replicates of the DNA library, which were highly and significantly correlated (Figure S1 Spearman′s $\rho = (0.97, 0.98)$, p-value<0.01). The DNA library was then transfected in LCLs (9 biological replicates) and we were able to recover a total of 43,500 testable SNPs (see methods for RNA counts filter). Read counts for the 9 biological replicates were highly correlated (Figure S2 Spearman's $\rho = (0.35, 0.72)$, p-value<0.01). To identify SNPs with allele-specific effects, we applied QuASAR-MPRA [29]. For each SNP, the reference and alternate allele counts were compared to the DNA proportion in the plasmid library. We identified a total of 2,720 SNPs with ASE from the combined replicates (FDR 10%) (Table S6). To investigate the importance of UMIs in this experimental approach, we re-analyzed our data without removing duplicates. For the combined replicates, inflation is greatly increased (from 1.10 to 1.73). Using UMIs in STARR-seq approaches is particularly relevant when a small number of replicates is used. For example, in the extreme case when a single replicate is considered, inflation increases slightly (from 0.92 to 1.18) and the number of detected ASE is about 4x reduced (Figure 1b).

SNPs with ASE are significantly enriched for variants predicted to impact transcription factor binding (CentiSNPs) [12] (Figure 1c, Table S1a, Fisher′s exact test p-value=4.55×10^{-06}, OR=2.49). Additionally, SNPs with ASE are enriched for low p-values in an eQTL mapping study performed in immune cells [28] (Figure 1d, Table S1c), thus confirming that our synthetic oligos can reproduce allele-specific regulatory effects observed in a native (non-episomal) cellular context.

The CentiSNP annotation is informative of the specific transcription factor motif being disrupted by a SNP. By leveraging this information, we were able to analyze allelic effects for specific transcription factor motifs (Table S7) (Figure 2a). Additionally, by combining the ASE results with the direction of the motif, we can characterize whether the motif is active in both directions or only in one direction. In other words if for a factor ASE tends to occur only in one direction, this would suggest a directional effect on the regulatory function. We found that when both alleles are covered in both directions, the allelic effects on gene expression tend to agree in direction and magnitude. If we categorize these directional allelic effects per motif, we do not observe major differences with the notable exception of CTCF (Figure 2b). Specifically, we find that SNPs in footprints for CTCF are significantly enriched (Fisher′s exact test OR=1.57, Bonferroni p=0.02) for ASE when the direction of transcription of the reporter gene is opposite to that of the motif strand (Figure 2c).

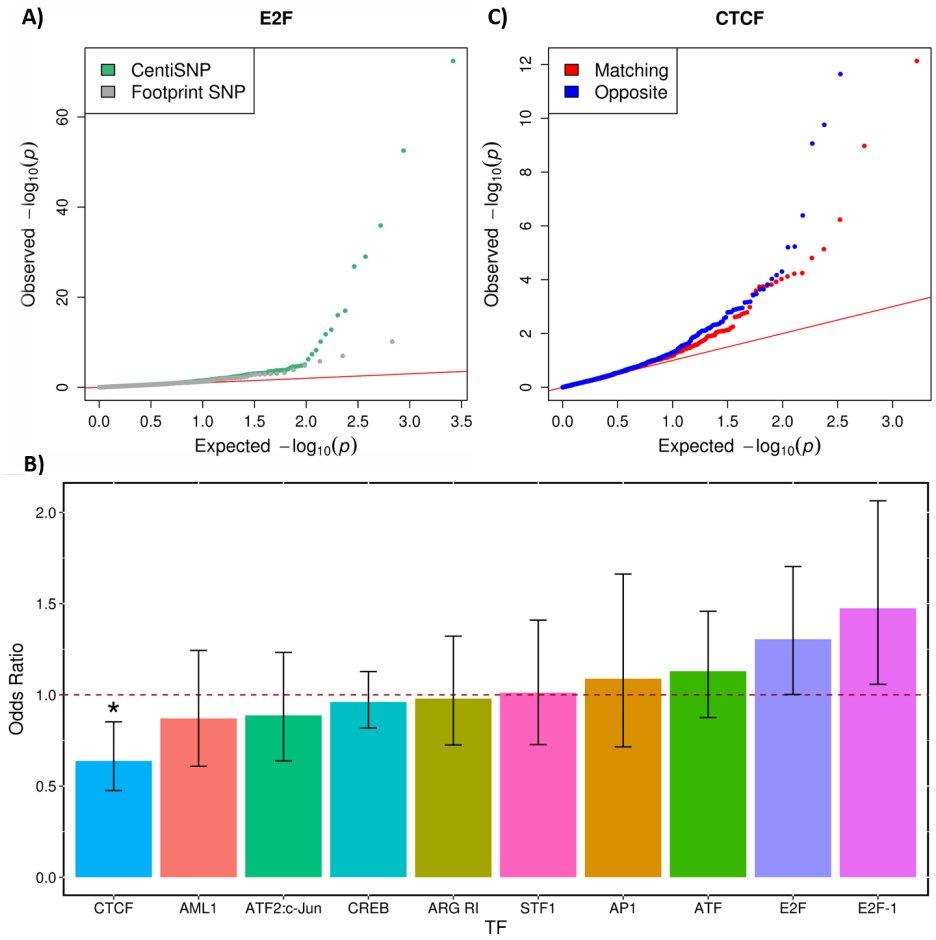

Figure 2: **ASE for individual transcription factors**. A) QQplot depicting the ASE p-value distributions from QuASAR-MPRA, for SNPs overlapping with E2F footprint annotations. SNPs predicted to alter binding (CentiSNPs) are represented in green, while SNPs that are in E2F but predicted to have no effect on binding are in grey. B) Enrichment for ASE in individual transcription factor binding sites calculated when motif strand matched the BiT-STARR-seq oligo transcription direction. Odds ratio (y axis) for each transcription factor tested (x axis) is shown in the barplot, error bars are the 95% CI from the Fisher's exact test. Odds ratios below the dotted line represent enrichment for opposite direction oligo/motif configuration. Stars are shown above significant results (Bonferroni adjusted p-value <0.05). C) QQplot depicting the ASE p-value distributions from QuASAR-MPRA, overlapping with CTCF footprint annotations. In red are the SNPs where the motif strand matches the BiT-STARR-seq oligo direction relative to the TSS, in blue where the motif strand is the opposite of the BiT-STARR-seq direction.

In order to understand the effect of a regulatory variant on complex traits it is necessary to first dissect the molecular function that is impacted by the nucleotide change. The CentiSNP prediction provides specific hypotheses for allelic effects on transcription factor binding that can be directly tested experimentally. Further matching ASB to ASE identified in BiT-STARR-seq would provide a complete molecular mechanism, from predicted binding effects, to experimen-

tally validated binding effects, to validated effects on expression. Due to the enrichment of CentiSNPs among SNPs with ASE in BiT-STARR-seq, we performed BiT-BUNDLE-seq to validate their effect on transcription factor binding. This is a new and efficient extension of high throughput reporter assays, since it uses the same input DNA library. We performed BiT-BUNDLE-seq with purified NFKB-p50 (at three different concentrations), which is an important regulator of the immune response in LCLs and other immune cells [30, 31, 32]. Previous studies have successfully identified ASB from ChIP-seq for NFKB in LCLs [33, 34, 35, 7, 36, 37] and NFKB footprints are induced in response to infection [38]. Additionally, NFKB was found to be 50 fold enriched for reQTLs from response to Listeria and Salmonella [28].

We first analyzed NFKB-p50 binding between the bound and unbound libraries and identified 9,361 significantly (logFC$>$1 and FDR$<$1%) over-represented regions in the bound library (Table S3). As expected, these regions were enriched (OR=11.70,13.75, p-value=7.95×10^{-27},1.23×10^{-15}) for NFKB footprints (Figure 3a, S3), with a higher portion of these regions in the mid concentration of NFKB-p50 as compared to the low or high concentrations (Figure 3b). This is probably because the low concentration doesn′t capture all of the NFKB-p50 binding, whereas the high concentration likely results in non-specific binding. We then used ΔAST [39] to identify ASB in the bound library (as compared to the unbound library), and combined the replicates using Stouffer′s method (See Methods). We successfully identified 386 SNPs at low concentration, 797 SNPs at mid concentration, and 894 SNPs at high concentration with significant ASB at 10% FDR (Figure 3d), for a total of 2,951 SNPs when aggregating all experiments (Table S1b). When we considered these ASB SNPs in combination with the ASE results from the BiT-STARR-seq, we found that SNPs with ASE are enriched for NFKB-p50 ASB (Figure 3e, Fisher′s exact test (OR=2.04, p-value=1.51×10^{-16})). This confirms our hypothesis that disruption of NFKB-p50 binding is one of the mechanisms underlying allele-specific expression in our dataset.

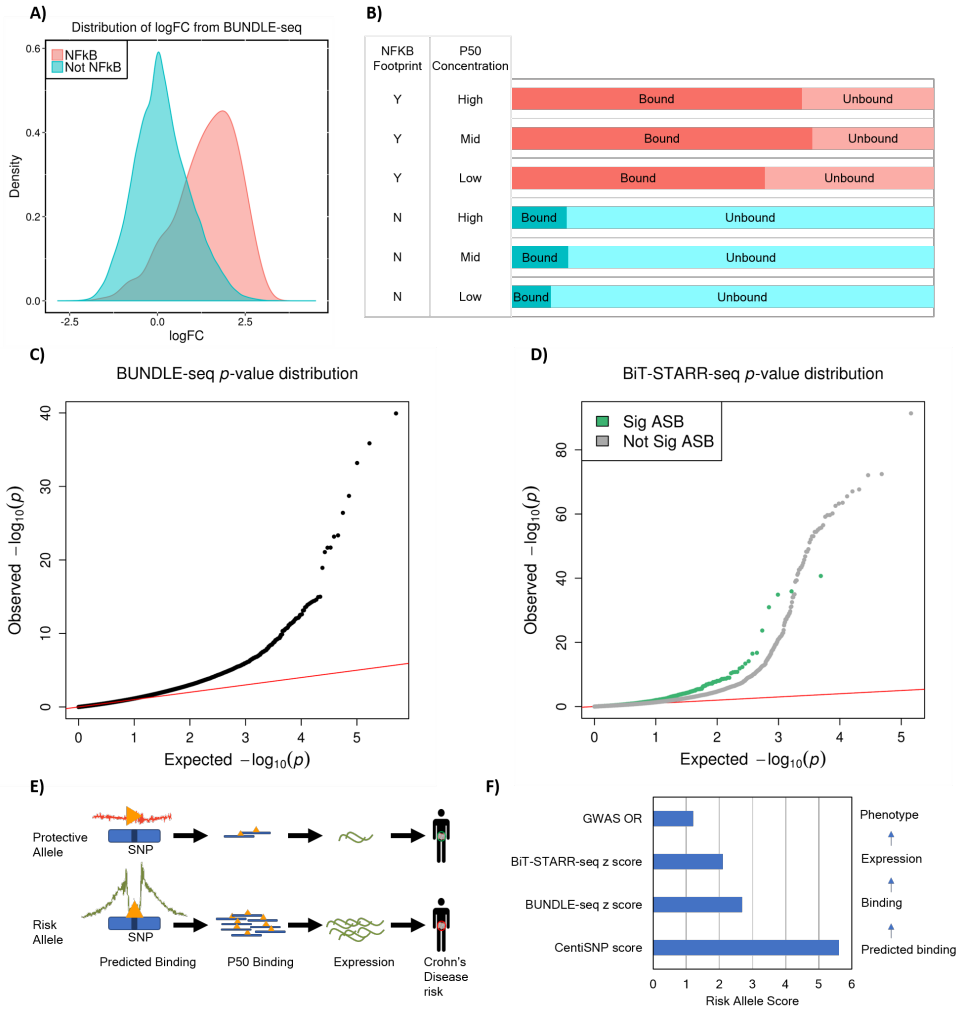

Figure 3: **Allele-specific binding for NFKB-p50**. A) Density plot of the logFC (from DESeq2) between bound and unbound DNA fractions from the BiT-BUNDLE-seq experiment. In red are the regions containing a SNP in a NFKB footprint, in blue the regions containing a SNP in footprints for other transcription factors. B) Barplot representing the number of independent enhancer regions in bound (dark color, DESeq2 logFC>1 and FDR<1%) and unbound (light color) DNA. NFKB-p50 concentration and presence of a NFKB footprint are indicated in the two columns on the left of the panel. C) QQplot depicting the p-value distributions from testing for ASB signal specific to the bound DNA fraction using ΔAST. D) QQplot depicting the ASE p-value distribution from QuASAR-MPRA for SNPs with significant (FDR<10%) (green) or not significant (grey) ASB in the BiT-BUNDLE-seq experiment. E) Integration of prediction, BiT-BUNDLE-seq, BiT-STARR-seq and GWAS results for Crohn's disease risk variant rs3810936. Triangles represent transcription factors. F) Comparing allelic effects from computational prediction to phenotype for rs3810936. Predicted log odds score is the reference prior log odds - alternate log odds from the CentiSNP annotation. BiT-BUNDLE-seq z score is the z score from the meta-analysis of ASB from all 3 concentrations of NFKB-p50. BiT-STARR-seq z score is the z score from meta-analysis of ASE for nine experimental replicates. GWAS OR is the odds ratio from rs3810936 alternate allele with Crohns disease [40, 41, 42, 43]. All scores are signed relative to the risk allele, which is the alternate allele.

We used ASB and ASE in combination with transcription factor binding motifs to assign mechanistic function to putatively causal SNPs linked to complex traits. We found 2,054 CentiSNPs with ASB (p-value$<$0.05) and 1,769 CentiSNPs with ASE (p-value$<$0.05) associated to a complex trait in the GWAS catalog (Table S5,S6) or with a PPA$>$0.1 from fgwas [12]. We were able to dissect the molecular mechanism for rs3810936, a variant associated with risk for Crohn′s disease in multiple populations [40, 41, 42, 43] (Figure 3e,f). This variant is a CentiSNP for the factor Hmx3 (Nkx5-1) and we find ASB for NFKB (p-value=0.006) in the BiT-BUNDLE-seq assay and ASE (p-value=0.034) in both directions in the BiT-STARR-seq. This SNP is a synonymous variant in gene *TNFSF15* (also known as *TL1A*), which encodes for a cytokine that belongs to the tumor necrosis factor (TNF) ligand family. Increased TL1A expression has been reported in inflamed Crohn′s disease tissue compared with non-inflamed areas, and in ulcerative colitis patient serum [44, 45, 46]. TL1A gives costimulatory signals to activated lymphocytes through binding to death-domain receptor 3 (DR3) [47] which induces the secretion of IFN-gamma [48, 49]. This gene modulates Th-1 and Th-17 [45, 50], creating an immunological state that leads to the mucosal inflammation of Crohn′s disease. Stimulation of the TL1A pathway, in monocytes and T cells from patients carrying the disease-associated *TL1A* SNPs, showed higher levels of TL1A expression, therefore aberrant TL1A expression may be a factor driving IBD development [51, 52]. This gene additionally has been found to be downregulated in response to dexamethasone [39], a corticosteroid used to treat many inflammatory and autoimmune conditions. While this variant is not found in ChIP-seq from ENCODE, ENCODE studies used RELA (p65) for NFKB, where our study used p50. We therefore identify a novel variant that disrupts binding of NFKB-p50, where the alternate allele (C) has increased binding. This leads to a increase in gene expression for the alternate allele, which is also the risk allele for Crohn′s disease (OR=1.21, p-value=1×10^{-15}).

The recent adaptation of MPRA to investigate ASE allows for validation of regulatory variants in transcription factor binding sites, which have been shown to be functionally relevant to fine map eQTLs and GWAS signals. However, the use of functional genomics to select relevant regions prior to experimental validation can reduce the number of sites it is necessary to validate. We developed a high throughput reporter assay that synthesizes these selected regions, clones them in 3'of the reporter gene, and includes the addition of a UMI during cDNA synthesis. This is the most streamlined protocol to date, and allows for removal of PCR duplicates which reduces noise in the data for greater power to detect ASE.

Our results show that using existing annotations to prioritize regulatory variants for high-throughput reporter assays is an effective strategy. The CentiSNP annotation in particular contains information that can be used to analyze ASB/ASE for individual transcription factor motifs and investigate potential molecular mechanisms of action. We found that direction is an important factor in the case of CTCF, most likely due to how CTCF functions as an insulator between the enhancer and the promoter when they are in anti-parallel directions. While enhancers can exert their regulatory function in either direction [53], insulators are only active if they can bind the DNA in specific arrangements relative to the promoter [54, 55]. Previous studies have shown that CTCF, a well characterized insulator, has binding sites at the anchors of chromatin loops.

These are arranged in forward-reverse orientations [56, 57, 58, 59, 60, 61], where the relative positions and orientations of the binding sites are important for mechanism of action [56]. In our case, the interaction could be mediated either by the basal transcriptional machinery at the TSS or also an additional weak CTCF binding site (M01259) that is present in the promoter and could help to establish a DNA loop. This would explain why for SNPs in CTCF binding sites, we observe a significant forward-reverse orientation preference and suggests that our episomal assay may be using CTCF to establish a DNA loop or conformation to enhance transcription.

We used our library of oligos also in a BiT-BUNDLE-seq assay for identification of ASB for NFKB-p50. This is a novel approach to combine ASB and ASE identification in high throughput assays using the same sequences. Our results show that this integration is a useful approach to validate the molecular mechanism for specific transcription factors.

Allelic effects on transcription factor binding and gene expression are not always concordant. This is the case, for example, of an allele that increases binding of a factor with repressing activity on gene expression. For example, we identified regulatory variants where there is increased binding for NFKB-p50, but decreased expression. These variants are in regions enriched for the CREB motif, and CREB has been shown to antagonize NFKB binding [62, 63]. These regulatory events are likely to be captured in the BiT-STARR-seq assay, which is performed in LCLs where both CREB and NFKB are active. These results highlight that multiple type of assays are necessary to capture the detailed molecular mechanism of gene regulation. Additionally, integration with GWAS can identify and further characterize the molecular mechanisms linking causal genetic variants with complex traits.

Methods

Oligo selection and design

Table S1 reports the annotations we have considered with their sources. These included: SNPs predicted to alter transcription factor binding in LCLs and HepG2 (CentiSNPs, [64]), LCL eQTLs fine-mapped in [5], liver eQTLs [4], significant fgwas SNPs in transcription factor binding motifs for 18 complex traits [12], significant fgwas SNPs for base models of functional annotations for 18 complex traits [26], ASB SNPs, and strong enhancers with no predicted ASB [64]. CentiSNP is an annotation that we recently developed [12], and that uses the CENTIPEDE framework [65] to integrate DNase-seq footprints with a recalibrated position weight matrix (PWM) model for the sequence to predict the functional impact of SNPs in footprints. SNPs in footprints "footprint-SNPs" are further categorized using CENTIPEDE hierarchical prior for each allele as "CentiSNP" if the prior relative odds for binding are >20. Fasta sequences with a window of 99 (on each side of the SNP) on the bed file were grabbed using seqBedFor2bit, and 15bp matching sequencing primers used for Illumina NGS were added to each end. Each regulatory region was designed to have two oligos: one for each of the alleles. A second list of the fasta sequences without the primer ends was generated to use as a custom reference genome, then converted to fastq using faToFastq. The full SNP list was aligned to the hg19 genome with BWA mem [66], removing the regions with a quality score less than 20. The full SNP list was also aligned to the custom reference genome, and then filtered for a quality score of 190. A total of 39,366 indexes were randomly generated to match this pattern: RDHBVDHBVD. This sequence was chosen to limit the longest possible polyACGT run at any position to 3 nucleotides, and avoid a G in the first and last position (corresponding to a dark cycle on the Illumina NextSeq500).

Oligo synthesis and amplification

DNA inserts 230bp long, corresponding to 200bp of regulatory sequence, were synthesized by Agilent to contain the regulatory region and the SNP of interest within the first 150bp. We performed a first round of PCR using Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [F_transposase and R_transposase] with cycling conditions: 98°C for 30s, followed by 4 cycles of 98°C for 10s, 50°C for 30s, 72°C for 60s, followed by 6 cycles of 98°C for 10s, 65°C for 30s, 72°C for 60s, followed by 72°C for 5 min. This reaction was used to double strand the oligos and complete the sequencing primers. The PCR product was run on a 2% agarose gel, extracted and purified with the NucleoSpin Gel and PCR Clean-Up Kit (Clontech). A subsequent round of PCR amplified the material using the same reaction as in the first round of PCR, but with cycling conditions: 98°C for 30s, followed by 15 cycles of 98°C for 10s, 65°C for 30s, 72°C for 60s, followed by 72°C for 5min. The PCR product was purified as described above.

Cloning Regulatory regions into pGL4.23

Plasmid pGL4.23 (Promega) was linearized using CloneAmp HiFi PCR Premix (Clontech), primers [STARR_F_SH and STARR_R_SH], and 35 cycles of 98°C for 10s, 60°C for 15s, and 72°C for 5s. The PCR product was purified on a 1% agarose gel as described above. Inserts were cloned into the linear plasmid using standard Infusion (Clontech) cloning protocol. Clones were transformed into XL10-Gold Ultracompetent Cells (Agilent) in a total of 7 reactions. These reactions were pooled and grown overnight in 500ml LB at 37°C in a shaking incubator. DNA was extracted using Endofree maxiprep kit (QIAgen).

Transfection of library

DNA library was transfected into LCLs using standard nucleofection protocol, program DS150, 3μg of DNA and 7.5×10^6 cells. A total of 3 sets of transfections were done in triplicate cuvettes, then pooled. We performed nine biological replicates of the transfection from 7 independent cell growth cultures. After transfection, cells were incubated at 37°C and 5% CO2 in RPMI1640 with 15%FBS and 1% Gentamycin for 24h. Cell pellets were then lysed using RLT lysis buffer (QIAgen), and cryopreserved at -80°C.

Library preparation

RNA-libraries. Thawed lysates were split in three aliquotes and total RNA was isolated using RNeasy Plus Mini Kit (QIAgen). Poly-Adenylated RNA was selected using Dynabeads mRNA Direct Kit (Ambion) using the protocol for total RNA input. RNA was reverse transcribed to cDNA using Superscript III First-Strand Synthesis kit (ThermoFisher) with primer [Nextera_i7_10N] and following the manufacturer's protocol. cDNA technical replicates were pooled and SPRI Select beads (Life Tech) were used for purification and size selection at a ratio of 0.9X. PCR Library Enrichment was performed using a nested PCR protocol. For the first round of PCR we used Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [F_trans_short and Illumina2.1] with cycling conditions: 98°C for 30s, followed by 15 cycles of 98°C for 10s, 72°C for 15s, followed by 72°C for 5 min. PCR product was purified on a 2% agarose gel as described above. The nested PCR used Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [fixed N5xx adapter (Illumina) (unique per each library replicate) and Illumina2.1] with cycling conditions: 98°C for 30s, followed by 5 cycles of 98°C for 10s, 72°C for 15s, followed by 72°C for 5 min. In a side quantitative real-time PCR reaction, 5μL of PCR product, 10X SYBR Green I, and the same primers and master mix were run in conditions: 98°C for 30s, 30 cycles of 98°C for 10s, 63°C for 30s, and 72°C for 60s. To determine the number of PCR cycles needed to reach saturation, we plotted linear Rn versus cycle and determined the cycle number that corresponds to 25% of maximum fluorescent intensity on the side reaction [67]. The PCR product was purified on a 2% agarose gel as described above.

DNA-libraries. We prepared 7 replicates of the DNA library using the PCR protocol as

described in [67] except using primers [fixed N5xx adapter (Illumina) (unique per each library replicate) and Nextera_i7_10N] and 30ng of input plasmid DNA. PCR product was purified on a 2% agarose gel as described above.

BiT-BUNDLE-seq

We used BiT-BUNDLE-seq, a new version of the BUNDLE-seq protocol [25]. Input DNA sequences were extracted from the BiT-STARR-seq DNA plasmid library using the same PCR conditions as in preparing the DNA libraries, followed by purification on a 2% agarose gel as described above. We used N-terminal GST-tagged, recombinant human NFKB-p50 subunit from EMD Millipore. The reaction buffer (0.15 M NaCl, 0.5 mM PMSF [Sigma], 1 mM BZA [Sigma], 0.5X TE, and 0.16 μg/μL PGA [Sigma]) was incubated at room temperature for 2 hours in low binding tubes (ThermoFisher). The tubes were cooled for 30 min at 4°C, and then 0.067 μg/μL BSA (Sigma) was added before adding the NFKB-p50 protein. One hundred nanograms of DNA were then added, and the protein and DNA were incubated for 1 h at 4°C. Experiments were performed in triplicates for each NFKB-p50 concentration. The reaction mix was run with 6μL Ficoll (Sigma) in a 7.5% Mini-PROTEAN TGX Precast 10-well Protein Gel (BIORAD) in cold 0.25X TBE buffer for 2 hours at 100V. The gel was stained for 30 min with 3X GelStar (Lonza). Bound and unbound DNA bands were excised under a blue light transilluminator. The DNA was eluted from the gel using the QIAQuick Gel Extraction Kit with a User-Developed Protocol (QIAgen QQ05). The gel slices were incubated in a diffusion buffer (0.5 M ammonium acetate, 10mM magnesium acetate, 1mM EDTA, ph 8.0 [KD Medical]; 0.1% SDS [Sigma]) at 50°C for 30 minutes. The supernatant was then passed through a disposable plastic column containing packed, siliconized glass wool [Supelco] to remove any residual polyacrylamide. Libraries were then quantified and loaded on the NextSeq500 for sequencing.

Library Sequencing

Pooled RNA and DNA libraries were sequenced on the Illumina Nextseq500 to generate 125 cycles for read 1, 30 cycles for read 2, 8 cycles for the fixed multiplexing index 2 and 10 cycles for index 1 (variable barcode).

Data Processing

Reads were mapped using the Hisat2 aligner [68], using the 1Kgenomes snp index so as to avoid reference bias. First we removed variants whose UMI was not possible to be present, given the UMI pattern selected. We then ran UMItools [69] using standard flags, as well as a q20 filter. We then ran the deduplicated files through mpileup using a bed file of our full SNP list, the -t DP4, -g, and -d 1000000. DNA reads were processed through a counts filter (on the summed replicates) of more than 7 counts per SNP and at least one count for the reference and alternate alleles in either direction. 50,609 SNPs in the DNA library were used as input

to the RNA library. The RNA library was processed following the same procedure as for the DNA library, except that the counts filter required a count of >1 per SNP and at least one count for both reference and alternate alleles. To identify SNPs with allele-specific effects, we applied QuASAR-MPRA [29], where for each SNP the reference and alternate allele counts were compared to the DNA proportion. QuASAR-MPRA results from each replicate were then combined using the fixed effects method, and corrected for multiple tests using BH procedure [70].

BiT-BUNDLE-seq data analysis

Counts from both the unbound and bound DNA were combined, and a filter was set so that each SNP direction combination had 5 counts for each allele. This combined count was also used to calculate a reference proportion. Each replicate for the bound and unbound libraries were then run through QuASAR-MPRA using the calculated reference proportion. These were then compared using ΔAST [39] to identify ASB in the bound fraction that is differential relative to the unbound fraction. The replicates were combined using Stouffers method [71] to identify ASB for each NFKB-p50 concentration, and combined again to identify the total ASB. The unbound and bound libraries counts were additionally analyzed with DEseq2 [72] to identify over-represented bound enhancer regions (FDR 1% and logFC>1). To better estimate the dispersion parameters, the DESeq2 model was fit on all sequencing data and without merging the replicate libraries:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i) \quad (1)$$

$$\mu_{ij} = s_j q_{ij} \quad (2)$$

$$\log_2(q_{ij}) = \beta_{i,0} + \beta_{i,\text{C}(j)} + \beta_{i,\text{B}(j)} \quad (3)$$

For each enhancer region i and sample j, the read counts K_{ij} are modeled using a negative binomial distribution with fitted mean μ_{ij} and an enhancer region-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j and a parameter q_{ij} proportional to the expected true concentration of regions for sample j. The coefficient β_0 represents the mean effect intercept, $\beta_{\text{C}(j)}$ represents the lane (NFKB-p50 concentration:replicate) effect, and and $\beta_{\text{B}(j)}$ represents the Bound/Unbound effect for each NFKB-p50 concentration (High, Medium, and Low).

We then contrasted the bound to the unbound for each concentration (i.e., high concentration bound to high concentration unbound) using the default DEseq2 Wald test for each enhancer region $\beta_{\text{B}(j)} \neq= 0$, and a Benjamini-Hochberg (BH) adjusted p-value was calculated with automatic independent filtering (DEseq2 default setting).

GWAS overlap

SNPs nominally significant (p<0.05) for ASB (identified with ΔAST) or ASE (identified with QuASAR-MPRA) that were also annotated as CentiSNP were overlapped with SNPs from the

GWAS catalogue (V6) [73], as well as with SNPs fine-mapped with the fgwas software as in [12] with a PPA>0.1.

Acknowledgement

Funding to support this research was provided by NIH 1R01GM109215-01 (RPR, FL), AHA 14SDG20450118 (FL) and AHA 17PRE33460295 (CK). We would like to thank Wayne State University HPC Grid for computational resources, members of the Luca/Pique group for helpful comments and discussions and Luis Barreiro for making the reQTL data available.

Competing Interests

The authors declare no competing interests in this study.

References

[1] Dermitzakis, E. Cellular genomics for complex traits. *Nature Reviews Genetics* **13**, 215–220 (2012).

[2] Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572–7 (2005).

[3] Stranger, B. E. Population genomics of human gene expression. *Nature Genetics* **39**, 1217–1224 (2007).

[4] Innocenti, F., Cooper, G. & Stanaway, I. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genetics* **7**, 1–16 (2011).

[5] Wen, X., Luca, F. & Pique-Regi, R. Cross-population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLoS Genetics* **11**, 1–29 (2015).

[6] Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

[7] Kasowski, M., Grubert, F. & Heffelfinger, C. Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).

[8] Degner, J. F. *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–4 (2012).

[9] Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* **16**, 197–212 (2015).

[10] Gibbs, J., van der Brug, M. & Hernandez, D. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics* **6**, 1–13 (2010).

[11] Melzer, D., Perry, J., Hernandez, D. & Corsi, A. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics* **4**, 1–10 (2008).

[12] Moyerbrailean, G. A. *et al.* Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLoS Genetics* **12**, e1005875 (2016).

[13] Melnikov, A., Murugan, A., Zhang, X. & Tesileanu, T. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology* **30**, 271–277 (2012).

[14] Kwasnieski, J., Mogno, I. & Myers, C. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *PNAS* **109**, 19498–19503 (2012).

[15] Patwardhan, R., Hiatt, J., Witten, D. & Kim, M. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology* **30**, 265–270 (2012).

[16] Sharon, E., Kalma, Y., Sharp, A. & Raveh-Sadka, T. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **30**, 521–530 (2012).

[17] Kwasnieski, J., Fiore, C., Chaudhari, H. & Cohen, B. High-throughput functional testing of ENCODE segmentation predictions. *Genome research* **24**, 1595–1602 (2014).

[18] Arnold, C., Gerlach, D., Stelzer, C. & Boryń, Ł. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

[19] Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions in human. *bioRxiv* 193136 (2017).

[20] Vockley, C., Guo, C. & Majoros, W. Massively parallel quantification of the regulatory effects of non-coding genetic variation in a human cohort. *Genome research* **25**, 1206–1214 (2015).

[21] Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).

[22] Ulirsch, J. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).

[23] Stormo, G. D., Zuo, Z. & Chang, Y. K. Spec-seq: determining protein-DNA-binding specificity by sequencing. *Briefings in functional genomics* **14**, 30–8 (2015).

[24] Wong, D. *et al.* Extensive characterization of NF-κB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biology* **12**, R70 (2011).

[25] Levo, M. *et al.* Unraveling determinants of transcription factor binding outside the core binding site. *Genome research* **25**, 1018–29 (2015).

[26] Pickrell, J. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics* **94**, 559–573 (2014).

[27] Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature communications* **6**, 6905 (2015).

[28] Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* **167**, 657–669.e21 (2016).

[29] Kalita, C. A. *et al.* QuASAR-MPRA: Accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* btx598 (2017).

[30] Li, Q. & Verma, I. M. NF-κB regulation in the immune system. *Nature Reviews Immunology* **2**, 725–734 (2002).

[31] Beinke, S. & Ley, S. C. Functions of NF-kappaB1 and NF-kappaB2 in immune cell biology. *The Biochemical journal* **382**, 393–409 (2004).

[32] Smale, S. T. Selective Transcription in Response to an Inflammatory Stimulus. *Cell* **140**, 833–844 (2010).

[33] Zhao, B. *et al.* The NF-κB Genomic Landscape in Lymphoblastoid B Cells. *CellReports* **8**, 1595–1606 (2014).

[34] Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).

[35] Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290 (2013).

[36] Lim, C.-A. *et al.* Genome-wide Mapping of RELA(p65) Binding Identifies E2F1 as a Transcriptional Activator Recruited by NF-κB upon TLR4 Activation. *Molecular Cell* **27**, 622–635 (2007).

[37] Martone, R. *et al.* Distribution of NF-kappaB-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12247–52 (2003).

[38] Pacis, A. *et al.* Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome research* **25**, 1801–11 (2015).

[39] Moyerbrailean, G. *et al.* High-throughput allele-specific expression across 250 environmental conditions. *Genome Research* **26** (2016).

[40] Yamazaki, K. *et al.* Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Human Molecular Genetics* **14**, 3499–3506 (2005).

[41] Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* **42**, 1118–1125 (2010).

[42] Lee, Y. J., Kim, K. M., Jang, J. Y. & Song, K. Association of TNFSF15 polymorphisms in Korean children with Crohn's disease. *Pediatrics International* **57**, 1149–1153 (2015).

[43] Baskaran, K., Pugazhendhi, S. & Ramakrishna, B. S. Protective Association of Tumor Necrosis Factor Superfamily 15 (TNFSF15) Polymorphic Haplotype with Ulcerative Colitis and Crohn's Disease in an Indian Population. *PLoS ONE* **9**, e114665 (2014).

[44] Bamias, G. *et al.* High intestinal and systemic levels of decoy receptor 3 (DcR3) and its ligand TL1A in active ulcerative colitis. *Clinical Immunology* **137**, 242–249 (2010).

[45] Bamias, G. *et al.* Expression, Localization, and Functional Activity of TL1A, a Novel Th1-Polarizing Cytokine in Inflammatory Bowel Disease. *The Journal of Immunology* **171**, 4868–4874 (2003).

[46] Prehn, J. L. *et al.* Potential role for TL1A, the new TNF-family member and potent costimulator of IFN-γ, in mucosal inflammation. *Clinical Immunology* **112**, 66–77 (2004).

[47] Migone, T.-S. *et al.* TL1A Is a TNF-like Ligand for DR3 and TR6/DcR3 and Functions as a T Cell Costimulator. *Immunity* **16**, 479–492 (2002).

[48] Papadakis, K. A. *et al.* Dominant role for TL1A/DR3 pathway in IL-12 plus IL-18-induced IFN-gamma production by peripheral blood and mucosal CCR9+ T lymphocytes. *Journal of immunology (Baltimore, Md. : 1950)* **174**, 4985–90 (2005).

[49] Prehn, J. L. *et al.* Potential role for TL1A, the new TNF-family member and potent costimulator of IFN-γ, in mucosal inflammation. *Clinical Immunology* **112**, 66–77 (2004).

[50] Takedatsu, H. *et al.* TL1A (TNFSF15) Regulates the Development of Chronic Colitis by Modulating Both T-Helper 1 and T-Helper 17 Activation. *Gastroenterology* **135**, 552–567.e2 (2008).

[51] Michelsen, K. S. *et al.* IBD-Associated TL1A Gene (TNFSF15) Haplotypes Determine Increased Expression of TL1A Protein. *PLoS ONE* **4**, e4719 (2009).

[52] Kakuta, Y. *et al.* TNFSF15 transcripts from risk haplotype for Crohn's disease are overexpressed in stimulated T cells. *Human Molecular Genetics* **18**, 1089–1098 (2009).

[53] Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

[54] West, A. G., Gaszner, M. & Felsenfeld, G. Insulators: many functions, many mechanisms. *Genes & development* **16**, 271–88 (2002).

[55] Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics* **7**, 703–713 (2006).

[56] Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).

[57] Alt, F., Zhang, Y., Meng, F.-L., Guo, C. & Schwer, B. Mechanisms of Programmed DNA Lesions and Genomic Instability in the Immune System. *Cell* **152**, 417–429 (2013).

[58] Guo, Y. *et al.* CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 21081–6 (2012).

[59] Monahan, K. *et al.* Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 9125–30 (2012).

[60] Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–80 (2014).

[61] VietriRudan, M. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports* **10**, 1297–1309 (2015).

[62] Ollivier, V., Parry, G. C. N., Cobb, R. R., de Prost, D. & Mackman, N. Elevated Cyclic AMP Inhibits NF-κB-mediated Transcription in Human Monocytic Cells and Endothelial Cells. *Journal of Biological Chemistry* **271**, 20828–20835 (1996).

[63] Parry, G. C. & Mackman, N. Role of cyclic AMP response element-binding protein in cyclic AMP inhibition of NF-kappaB-mediated transcription. *Journal of immunology (Baltimore, Md. : 1950)* **159**, 5450–6 (1997).

[64] Moyerbrailean, G. A. *et al.* A high-throughput RNA-seq approach to profile transcriptional responses. *Scientific reports* **5**, 14976 (2015).

[65] Pique-Regi, R., Degner, J., Pai, A. & Gaffney, D. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research* **21**, 447–455 (2011).

[66] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* **1303** (2013).

[67] Buenrostro, J., Giresi, P. & Zaba, L. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213–1218 (2013).

[68] Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360 (2015).

[69] Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27**, 491–499 (2017).

[70] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (* **57**, 289–300 (1995).

[71] STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. & WILLIAMS, R. M. J. The American soldier: Adjustment during army life. *Princeton University Press* **265**, 173–175 (1949).

[72] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

[73] MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* **45**, D896–D901 (2017).

Supplementary Tables

Table S1: **Annotations Used**. SNP annotations used for overlap with BiT-BUNDLE-seq and BiT-STARR-seq. First 4 columns are in the same order for each file (chr, pos, pos1, rsID). A) CentiSNPs. Column 5 contains the transcription factor with a CentiSNP at that location. B) SNPs in complex traits. Column 5 contains the GWAS trait associated with the SNP. C) eQTL SNPs. Column 5 contains the information for whether the eQTL was identified in cells infected with L (*Listeria*), S (*Salmonella*), or NI (not infected). Column 6 contains the gene associated with the eQTL. Column 7 contains the beta for the eQTL association. Column 8 contains the p-value for the eQTL association.

A) `file://centi_supp.txt`
B) `file://gwas_supp.txt`
C) `file://eqtl_supp.txt`

Table S2: **BiT-STARR-seq results**. QuASAR-MPRA results for BiT-STARR-seq.

`file://bitstarr_meta_quasar.txt`

Table S3: **DEseq results**. Differentially bound regions for A) Combined concentrations, B) Low concentration, C) Mid concentration, and D) High concentration. Columns are the same for all 4 files (identifier(rsID_Direction), adjusted p-value, p-value, logFC).

A) `file://EMSA_DEG_stats2_withrepShift.txt`
B) `file://EMSA_DEG_stats2_withrep_concShiftLow.txt`
C) `file://EMSA_DEG_stats2_withrep_concShiftMid.txt`
D) `file://EMSA_DEG_stats2_withrep_concShiftHigh.txt`

Table S4: **BiT-BUNDLE-seq results**. ΔAST results for BiT-BUNDLE-seq. Columns are identifier, z score, p-value, adjusted p-value, rsID

`file://bundleseq_dast_comb.txt`

Table S5: **ASB and complex traits**. ΔAST results for BiT-BUNDLE-seq. SNPs are nominally significant, associated to a complex trait, and are also CentiSNPs. Columns are rsID, direction, p-value, complex trait.

`file://dast_gwas_centi_nomsig.txt`

Table S6: **ASE and complex traits**. QuASAR-MPRA results for BiT-STARR-seq. SNPs are nominally significant, associated to a complex trait, and are also CentiSNPs. Columns are rsID, direction, p-value, complex trait.

`file://bit_gwas_centi_nomsig.txt`

Table S7: **Transcription factors in BiT-STARR-seq**. Number of SNPs in motifs matching the top 10 covered transcription factors in BiT-STARR-seq.

Transcription Factor	Freq
CTCF	4911
E2F-1	2794
E2F	4407
ATF	5567
AML1	3794
ATF2:c-Jun	3651
CREB	12955
AP1	2673
ARG RI	3445
STF1	3561

Table S8: **Primers used in BiT-STARR-seq**

Primer	Sequence
STARR_F_SH	CCGAGCCCACGAGACCTAGAGTCGGGGCGGCCG
STARR_R_SH	TGACGCTGCCGACGAAATTATTACACGGCGATC
F_transposase	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
R_transposase	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
I2.1	CAAGCAGAAGACGGCATACGA
Nextera_i7_10N	CAAGCAGAAGACGGCATACGAGAT**RDHBVDHBVD**GTCTCGTGGGCTCGG

Supplemental Figures

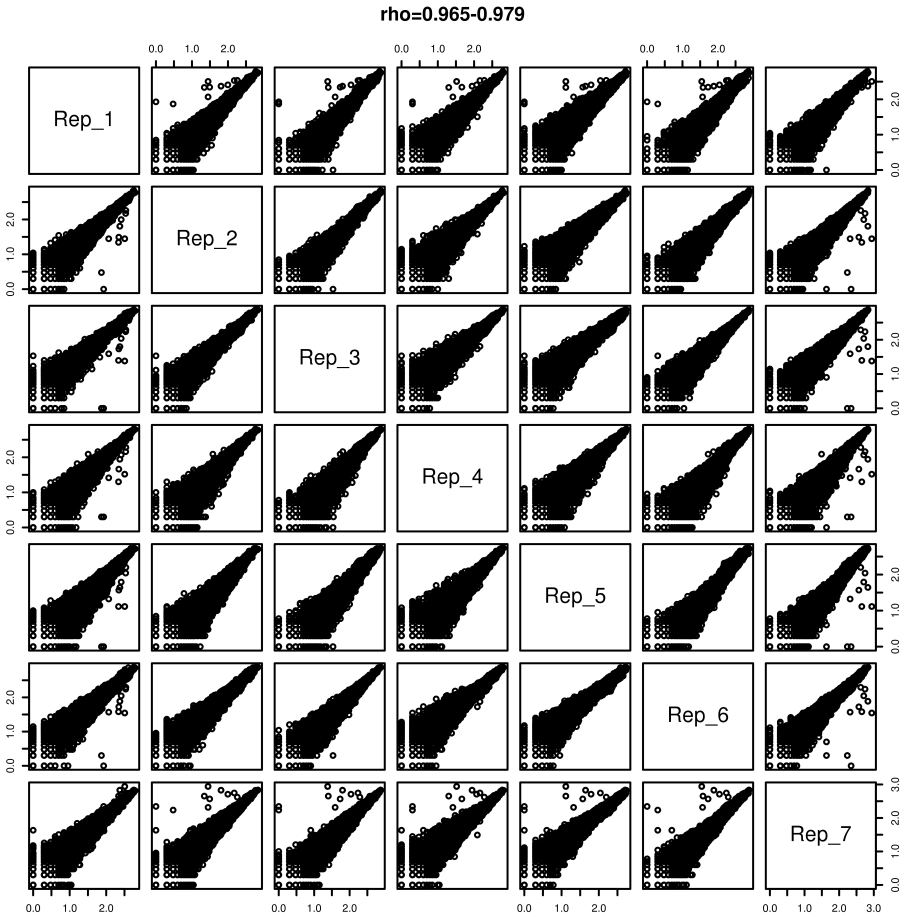

Figure S1: **Correlation of DNA libraries**. Scatterplot of filtered DNA library counts for each replicate plotted against all other replicates. Spearman rho correlation range is stated at the top.

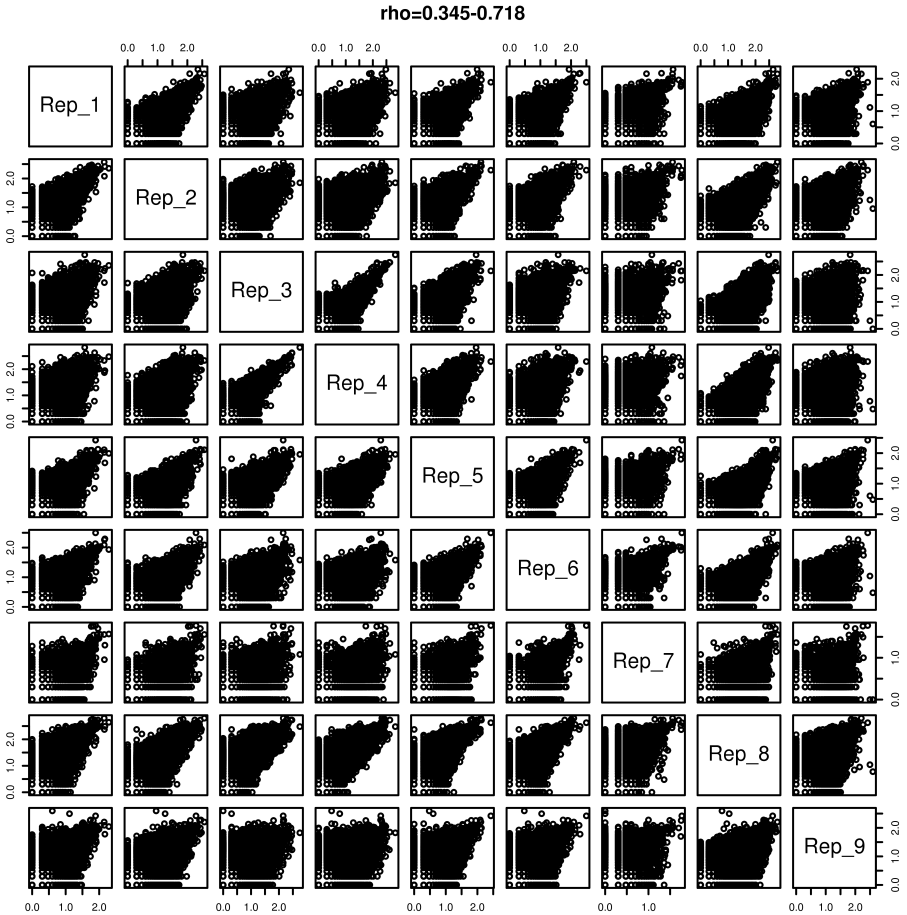

Figure S2: **Correlation of RNA libraries**. Scatterplot of filtered RNA library counts for each replicate plotted against all other replicates. Spearman rho correlation range is stated at the top.

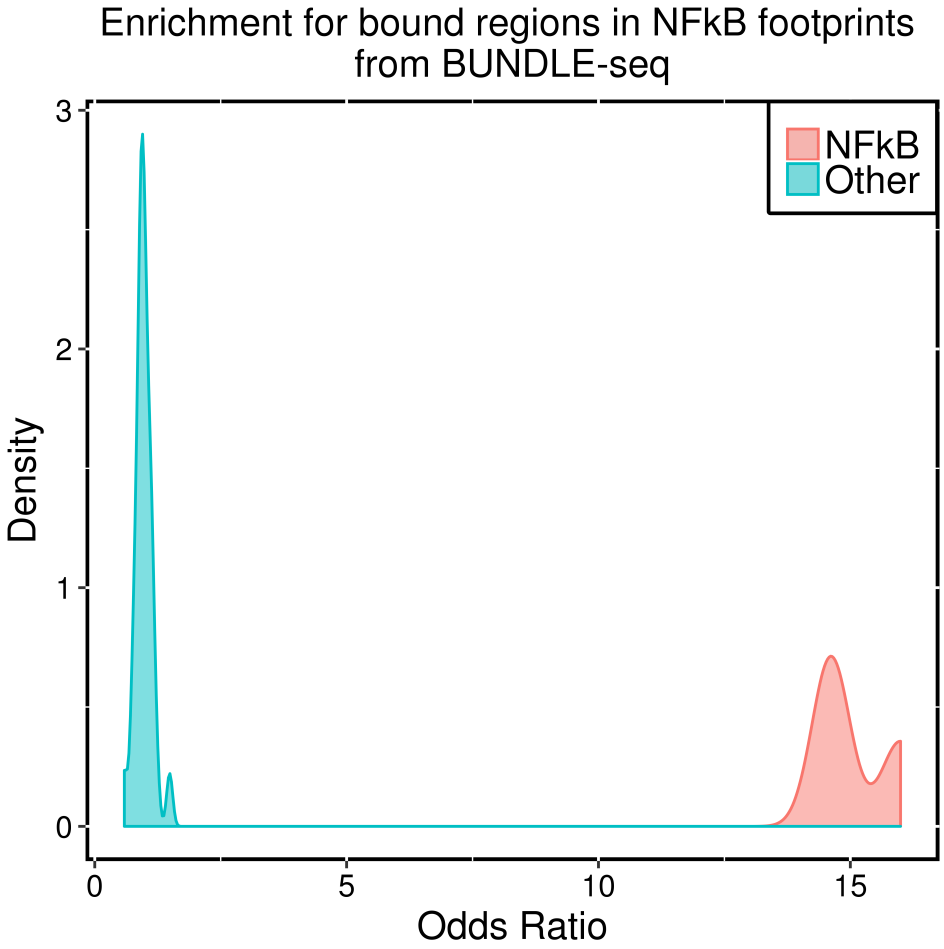

Figure S3: **Enrichment of NFKB footprints in BiT-BUNDLE-seq bound regions**. Fishers exact test was performed to identify enrichment (x axis is the OR) for significant differentially bound regions (logFC>1 and FDR<1%). In red are the regions containing a SNP in a NFKB footprint, in blue the regions containing a SNP in footprints for other transcription factors.