RESEARCH

Open Access

Chromatin activity identifies differential gene regulation across human ancestries

Kade P. Pettie[1], Maxwell Mumbach[2], Amanda J. Lea[3], Julien Ayroles[4], Howard Y. Chang[5,6], Maya Kasowski[7,8] and Hunter B. Fraser[1*]

*Correspondence:
hbfraser@stanford.edu

[1] Department of Biology, Stanford University, Stanford, CA, USA
[2] Department of Genetics, Stanford University, Stanford, CA, USA
[3] Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA
[4] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
[5] Center for Personal Dynamic Regulomes, Stanford University School of Medicine, Stanford, CA, USA
[6] Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA
[7] Sean N. Parker Center for Allergy and Asthma Research, Stanford University School of Medicine, Stanford, CA, USA
[8] Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA

Abstract

Background: Current evidence suggests that *cis*-regulatory elements controlling gene expression may be the predominant target of natural selection in humans and other species. Detecting selection acting on these elements is critical to understanding evolution but remains challenging because we do not know which mutations will affect gene regulation.

Results: To address this, we devise an approach to search for lineage-specific selection on three critical steps in transcriptional regulation: chromatin activity, transcription factor binding, and chromosomal looping. Applying this approach to lymphoblastoid cells from 831 individuals of either European or African descent, we find strong signals of differential chromatin activity linked to gene expression differences between ancestries in numerous contexts, but no evidence of functional differences in chromosomal looping. Moreover, we show that enhancers rather than promoters display the strongest signs of selection associated with sites of differential transcription factor binding.

Conclusions: Overall, our study indicates that some *cis*-regulatory adaptation may be more easily detected at the level of chromatin than DNA sequence. This work provides a vast resource of genomic interaction data from diverse human populations and establishes a novel selection test that will benefit future study of regulatory evolution in humans and other species.

Keywords: HiChIP, Activity-by-Contact, Cis-regulation, Enhancer, bQTL, Gene expression, Directional selection, LCL, Human

Background

Identifying the genetic determinants of complex traits is challenging because their contributions are often diluted across many variants of small effect. Single variants of large effect are simpler to identify and have been well-characterized [1–3], but genome-wide association studies (GWAS), which test millions of variants for statistical association with a trait, have demonstrated that these large-effect loci are rare. Moreover, the vast majority of trait-associated variants are located in non-coding regions [4–7]. For the <

10% of GWAS hits in protein-coding regions, inferences about their evolutionary history and mechanisms of action are often readily available thanks to studies that have focused on these regions. The remaining > 90% of GWAS hits in non-coding regions are thought to affect traits by altering gene expression levels, but causal mechanisms are obscured by a combination of linkage disequilibrium (LD), a genome-wide phenomenon in which nearby variants tend to be inherited together leading to a correlation of their effects [8], and the paucity of information about non-coding relative to coding regions.

Even before the GWAS era variants with highly divergent allele frequencies between populations, measured by estimates of Wright's fixation index (F_{ST}) [9], were found to be enriched in disease-associated genes [10]. Since then, genome-wide scans using associations of allele frequencies with environmental variables as evidence of natural selection have shown signals of positive selection to be somewhat enriched in coding regions [11–15], and even more enriched in cis-regulatory elements [16]. Overall, the preponderance of non-coding variants implicated in human GWAS is paralleled by a similar trend among human genetic variants involved in local environmental adaptation [16].

Intersecting non-coding GWAS hits with information from assays measuring regulatory activity, such as quantitative trait loci (QTL) for molecular-level traits (mol-QTL), has been effective at pinpointing causal variants and molecular mechanisms underlying complex trait variation [17–22]. QTL studies using gene expression as the trait (eQTL) test all variants within a predefined distance (usually one megabase (Mb)) of a gene for an association with that gene's expression, so each eQTL is linked to a target gene [20]. Since transcription factor (TF) proteins bind gene regulatory elements such as enhancers in a sequence-dependent manner to regulate transcription, eQTL can act by altering a TF's binding affinity (i.e., one allele has higher binding affinity than the other, termed a bQTL) [18]. In most cases, increased TF binding is associated with decompaction of chromatin, the DNA-protein complex that packages meters of linear DNA into a nucleus a few microns wide. This opening of the chromatin allows more TFs to bind to previously inaccessible stretches of DNA and to each other in a positive feedback loop of chromatin accessibility. Thus, chromatin accessibility can be used as a proxy for regulatory activity to identify enhancers and their relative activity levels, as is accomplished with Assay for Transpose-Accessible Chromatin (ATAC-seq) [23].

Since enhancers operate in three-dimensional space and can contact target gene promoters (*cis*-regulation) several Mb away, ATAC-seq and high-throughput methods based on Chromatin Conformation Capture (HiC) [24–27] can be combined to identify enhancer-promoter interactions [18, 19, 22, 28]. The activity-by-contact (ABC) model was recently developed to predict enhancer-target gene pairs in a given cell type under the premise that the extent to which an element regulates a gene's expression depends on its strength as an enhancer (activity level), scaled by how often it is near that gene's promoter in 3D space (contact frequency) [29]. HiChIP, which combines HiC with chromatin immunoprecipitation (ChIP) on a protein of interest, is well-suited to generate input for this model, particularly when performed on the histone modification H3K27ac, a hallmark of active chromatin. Since the end product is paired-end reads from H3K27ac-associated long-range interactions, H3K27ac HiChIP provides a simultaneous measure of activity level and contact frequency without the high sequencing depth and cell number required to generate the all-by-all interaction maps of HiC [25, 30]. The ABC model has been shown to attain peak performance with chromatin accessibility and HiChIP data as input and outperforms other enhancer target gene prediction methods [29], making it a powerful metric for hypothesis generation about the mechanisms of noncoding GWAS hits [31].

Additional support for the mechanisms and causality of these hits can come from intersecting molecular-level QTL with putative locally adaptive variants [16]. However, since selection acts on fitness, its impact may be more directly observable at the level of chromatin activity than at the level of DNA sequence, where it is relatively more diluted (Fig. 1a, left). For example, chromatin activity is a better predictor of TF binding than DNA sequence since we do not fully understand the *cis*-regulatory "code" that governs TF binding [32]. This can lead to cases where sequence-level changes, even those disrupting TF binding sites, do not correspond to changes in regulatory function and gene expression when regulatory activity is buffered by the binding of multiple TFs.

Studies in primates have suggested that directional selection may have contributed to differences in chromatin activity that distinguish each species [33]. For example, sites with decreased chromatin accessibility in human relative to chimpanzee and rhesus macaque white adipose tissue tend to be *cis*-regulatory elements for lipid metabolism-related genes, consistent with humans' greater body fat percentage [34]. Such analysis of chromatin activity divergence has not been conducted on more recent evolutionary timescales within the human lineage, where mechanistic insights could aid understanding of ancestry-dependent disease prevalence [35–37].

Here, we use ATAC-seq and H3K27ac HiChIP, a combined measure of activity and contact frequency [25], to generate ABC scores linking candidate *cis*-regulatory elements (CREs) to candidate target genes (hereinafter "target genes") in eight populations of African or European ancestry. We then decompose these scores into their activity and contact components to identify differential CREs (diff-CREs) for each score between individuals of African and European ancestry (Fig. 1a). Intersecting our diff-CREs with bQTL reveals three transcription factors (NF-κ B, JunD, and PU.1) whose binding sites show signs of lineage-specific selection for differences in binding between the African and European ancestry populations. Our findings illustrate the utility of ABC scores to identify previously unappreciated population-specific activity of CREs, their target genes, and potential mechanisms of gene regulation.

Results

Differential CRE activity is linked to differential expression between ancestries

We previously performed ATAC-seq in lymphoblastoid cell lines (LCLs) from ten different global populations sequenced by the 1000 Genomes Project [19]. This was carried out in a pooled study design, with each population represented by a single pool of ~100 unrelated individuals. We selected the four African (ESN, GWD, LWK, and YRI) and four European (CEU, FIN, IBS, and TSI) ancestry (hereinafter AFR and EUR, respectively) populations for comparison to isolate the effects of any lineage-specific selection on gene regulatory elements that have occurred since the divergence of human populations native to these two continents. The AFR and EUR ancestries were represented by 418 and 413 individuals, respectively. We first identified a common set of CREs by (1) calling peaks on ATAC-seq data combined across the four population pools of each

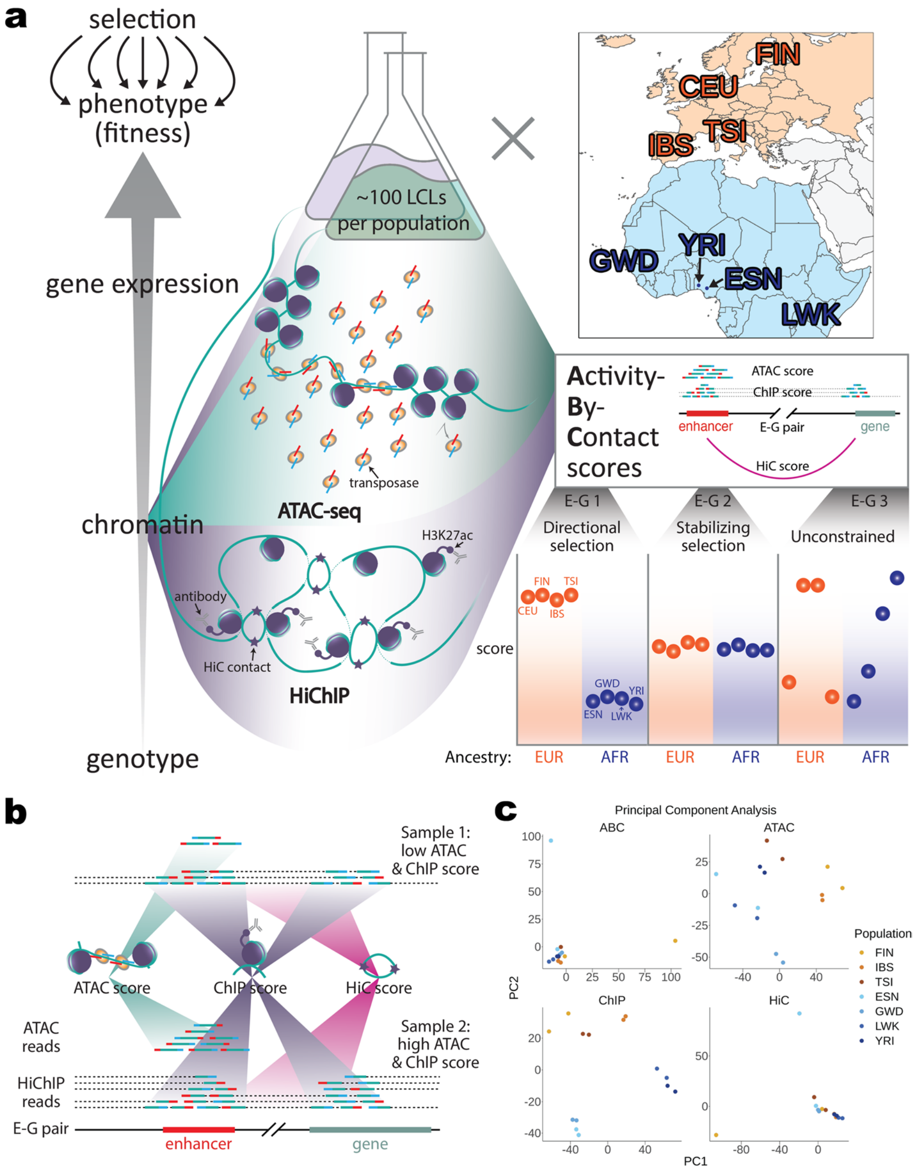


Fig. 1 Identification of candidate enhancer-gene pairs under different selection pressures between ancestries. **a** From left to right: Chromatin accessibility, activity, and contact frequency were assayed in LCLs from four African populations (blue) and four European populations (orange). Enhancer-gene pairs were defined using ABC scores and ATAC, ChIP, and HiC score components quantified. Enhancer-gene pair 1 exemplifies evidence of directional selection (low within- and high between-continental ancestry score variance). **b** Sequencing reads' contributions to each ABC component score are shown for a hypothetical enhancer-gene pair where sample 2 has higher ATAC and ChIP scores, but an equivalent HiC score relative to sample 1. Dotted lines extending to the depicted region border connect to reads whose other paired-end aligns elsewhere in the genome but contributes to the HiChIP score (purple gradient). **c** Principal component analysis results from all enhancer-gene pairs are shown for each score type. Both replicates are shown for each population. CEU is an outlier (Additional file 1: Fig. S5–8) and is thus excluded in these PCAs and in downstream analyses. Abbreviations: ATAC-seq, assay for transpose accessible chromatin with sequencing; E, enhancer; G, gene; H3K27ac, histone 3 lysine 27 acetylation; LCLs, lymphoblastoid cell lines; CEU, Utah residents with North European ancestry; ESN, Esan; FIN, Finnish; GWD, Gambian; IBS, Iberian; LWK, Luhya; TSI, Tuscan; YRI, Yoruban; AFR, African; EUR, European; PC, principal component

ancestry, then (2) resizing them to 500 bp centered on each peak summit to avoid any potential peak width bias, and (3) retaining the top 150,000 by read count ranking. We ensured equal peak contributions between ancestries (see Methods) to balance statistical power and for consistency with how the ABC model was developed [29].

To obtain the additional activity component and the contact component necessary for computing ABC scores, we performed H3K27ac HiChIP, which enriches first on the level of H3K27ac and second on HiC contact frequency of the two interacting regions [25], in two replicates per population of the same pooled LCLs (see the "Methods" section). H3K27ac HiChIP was shown to perform at least as well as H3K27ac ChIP-seq and HiC assayed separately when used in ABC scores adapted for this data type [29]. We mapped reads from each replicate to a common reference. To minimize allelic mapping bias, we retained only reads overlapping variants that mapped to the same unique location after swapping out one allele for the other [38]. Subsequent filtering to reads in valid *cis* interaction pairs yielded ~540 million paired-end reads qualified for use in ABC score computation (see Additional file 1: Fig. S1; Additional file 2).

To calculate ABC scores for each population, we jointly estimated activity level and contact frequency as the product of normalized ATAC-seq reads overlapping a given element and normalized HiChIP reads overlapping that element and the promoter of a given gene at 5 Kb resolution (see the "Methods" section). We identified 50,478 CRE-target gene (enhancer-gene) pairs with nonzero ATAC and HiChIP signal in all samples that passed enhancer-gene pair candidacy thresholds (see the "Methods" section) in at least one sample.

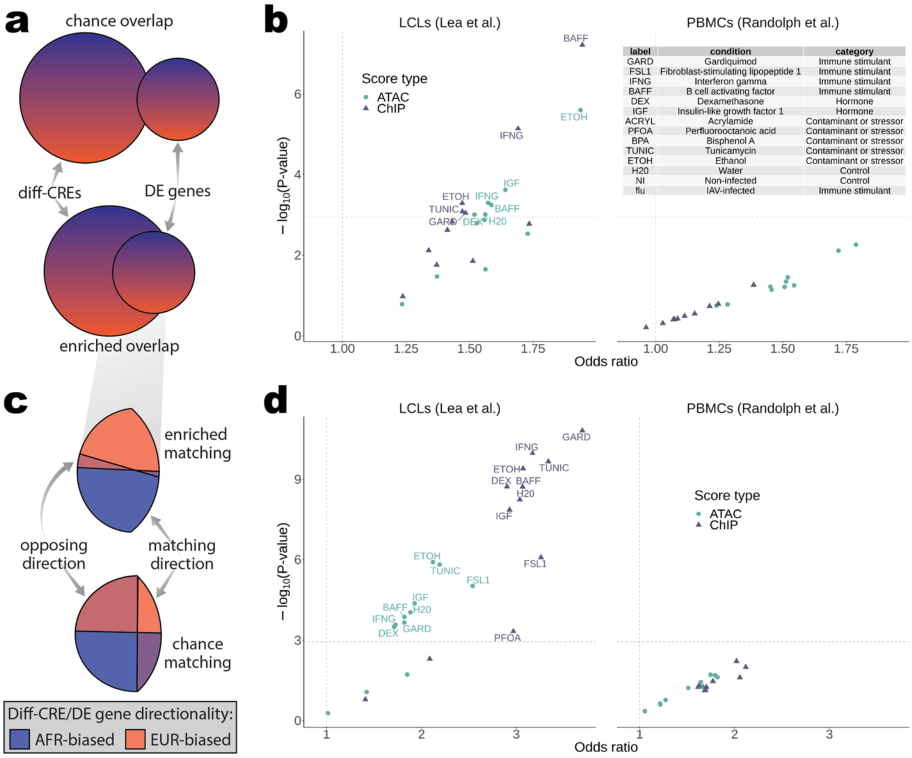
Since ABC scores are designed to identify enhancer-gene pairs, but not the relative expression levels of target genes, we reasoned that decomposing each score into three independent components—ATAC, H3K27ac ChIP, and HiC scores (Fig. 1a–b)—could allow us to search for evidence of selection on each as a distinct mechanism of differential gene expression regulation. Thus, for each enhancer-gene pair defined using ATAC-seq data in combination with our newly generated HiChIP data, ATAC scores represent the chromatin accessibility at the enhancer (Fig. 1b, teal gradients). ChIP scores estimate the enhancer-gene pair's collective H3K27ac signal as the geometric mean of total HiChIP signal at the enhancer and gene promoter (also known as the vanilla coverage square root (VC-sqrt)) (Fig. 1b purple gradients). HiC scores estimate the contact frequency of the enhancer and promoter independent of H3K27ac levels by dividing the HiChIP signal from read pairs specifically connecting the enhancer and promoter (Fig. 1b, magenta gradient) by the VC-sqrt (see the "Methods" section).

To assess how each of these scores captures differences between populations and replicates we performed principal component analysis (PCA) and hierarchical clustering across samples on all enhancer-gene pairs for each score type. Since both CEU replicates were outliers (see Additional file 1: Supplemental text, Fig. S5–8; Additional file 3) [39], we removed this population, redefined enhancer-gene pairs, and computed scores for downstream analyses using the remaining 14 samples. Although FIN rep1 and ESN rep1 are also outliers for HiC scores, and thus also for ABC scores (Additional file 1: Fig. S9–10), this is likely driven by low coverage HiC contacts since these are the two samples with the lowest number of valid HiChIP *cis* interaction pairs (Additional file 1: Fig. S1). For ChIP scores, which quantify the total H3K27ac signal at a CRE (not only that contributed by reads explicitly defining an E-G pair, as in HiC and ABC scores where the aforementioned low coverage effects manifest), these replicates are not outliers, so it is unlikely that coverage or batch effects contribute to any signal differences in this chromatin activity metric.

We then reanalyzed the differences between populations and replicates captured by our score types and quantified their ancestry-associated differential signals. PCA and hierarchical clustering on these scores show that ChIP scores are highly similar between replicates when considering either all enhancer-gene pairs (Fig. 1c) or the 5000 most variable pairs. Clustering by ancestry is apparent when considering the 5000 most variable enhancer CREs, but not promoter CREs (see Additional file 1: Fig. S9; Additional files 4 and 5). To assess the ancestry-associated differential regulatory activity of each ABC score component, we identified differential score (diff-score) enhancer-gene pairs (diff-score $P < 0.05$, see Methods; Additional file 1: Fig. S4). We found little or no differential signal between ancestries for this score type in the diff-HiC scores (FDR = 0.87 at diff $P < 0.05$ relative to FDR = 0.093 and 0.057 for diff-ATAC and diff-ChIP, respectively; see Methods; Additional file 1: Fig. S11; Additional file 6), or in downstream functional analyses.

To determine the extent to which diff-scores are associated with differential gene expression (DE) between African and European ancestry individuals, we analyzed gene expression data from two previous studies. Lea et al. (2022) measured gene expression across 12 cellular conditions (11 exposures and one unexposed control) in many of the same LCLs from African and European populations used in our study. Randolph et al. [40] measured gene expression in non-infected (NI) and IAV-infected (flu) peripheral blood mononuclear cells (PBMCs) at single-cell resolution from a panel of donors with varying degrees of African versus European ancestry. Both studies identified ancestry-associated DE (ancestry DE) genes, Lea et al. by modeling expression as a function of the African or European ancestry of each population, and Randolph et al. by modeling expression as a function of the proportion of African ancestry estimated from whole-genome sequencing. Although the context we assayed our LCLs in to generate ABC scores was closest to Lea et al.'s baseline/unexposed condition, by comparing diff-scores in a baseline (unstimulated) context to DE in other contexts we were able to ask if CREs could be poised for DE regulation upon stimulation and/or in another cell type.

We then asked if chromatin accessibility, H3K27ac levels, HiC contact frequency, and/or the combination of these components in ABC scores were associated with ancestry DE across these 22 combinations of cell type and stimulation conditions. We found six enrichments for ancestry DE in diff-ATAC and five in diff-ChIP genes among the 22 tested contexts (hypergeometric $P < 1.13 \times 10^{-3}$, Fig. 2b). For example, target genes of diff-ChIP CREs were overrepresented among ancestry DE genes in LCLs after four hours of exposure to B-cell-activating factor (BAFF, odds ratio (OR) = 1.94, $P = 6.1 \times 10^{-8}$), a strong B cell activator and tumor necrosis factor family cytokine. As expected based on the lack of signal in our initial FDR analysis, no contexts were enriched for ancestry DE in diff-HiC genes, and only DE genes in LCLs after four hours of exposure to ethanol (labeled "ETOH") were enriched in diff-ABC genes at the same Bonferroni-corrected P-value threshold used for diff-ATAC and ChIP (see Additional file 1: Fig. S12, S13b). This indicates that the inclusion of the contact frequency component in ABC scores



label	condition	category
GARD	Gardiquimod	Immune stimulant
FSL1	Fibroblast-stimulating lipopeptide 1	Immune stimulant
IFNG	Interferon gamma	Immune stimulant
BAFF	B cell activating factor	Immune stimulant
DEX	Dexamethasone	Hormone
IGF	Insulin-like growth factor 1	Hormone
ACRYL	Acrylamide	Contaminant or stressor
PFOA	Perfluorooctanoic acid	Contaminant or stressor
BPA	Bisphenol A	Contaminant or stressor
TUNIC	Tunicamycin	Contaminant or stressor
ETOH	Ethanol	Contaminant or stressor
H20	Water	Control
NI	Non-infected	Control
flu	IAV-infected	Immune stimulant

Fig. 2 Differential enhancer and promoter enrichments for differential expression across conditions and cell types. **a** Schematic depicting a simplified version of the enrichment analysis test results shown in **b**. Any diff-CRE was counted as a "success" overlap in the hypergeometric enrichment test for a given context if it had a target gene that was DE in that context. Color gradients illustrate that each element has a directionality, which is the subject of analysis in **c** and **d**. **b** Results of one-sided Fisher's exact tests for DE gene enrichment in diff-CREs are shown as odds ratios plotted against significance. Vertical dotted lines mark an odds ratio of 1 and horizontal dotted lines mark the Bonferroni-corrected *P*-value threshold. Points above this significance threshold are labeled with their context (see table inset). Score types used to define diff-CREs in each test are indicated by the shape and color of the points. **c** Schematic depicting the enrichment analysis test results shown in **d**. Each element from any overlap in **a** (e.g., gray gradient callout) was split into AFR- and EUR-biased directionality and any diff-CRE with a DE target gene in each context was counted as a "success" overlap if the ancestry DE direction matched the diff-score direction of the CRE (e.g., higher expression and higher score in AFR (blue)). For diff-CREs with multiple DE target genes, these targets were required to match direction to be included in each test. **d** Results of one-sided Fisher's exact tests for diff-CRE matching DE directionality are shown as odds ratios plotted against significance. Vertical dotted lines mark an odds ratio of 1 and horizontal dotted lines mark the Bonferroni-corrected P-value threshold. Points above this significance threshold are labeled with their context. For a more detailed illustration of these tests, see Additional file 1: Fig. S13a,c

weakens the association of the activity components with DE. Overall, the strength of the associations of differential chromatin accessibility (diff-ATAC) and H3K27ac levels (diff-ChIP) with DE across several contexts suggests CREs could be poised for DE regulation upon stimulation or differentiation to another cell type. Importantly, although we observe little-to-no differential HiC signal between ancestries, this component was critical in defining enhancer-gene pairs to test, as each pair must have at least one HiChIP read connecting the two elements to have a non-zero ABC score.

Although gene expression can be predicted by promoter activity [41], the contribution of promoter or enhancer activity to ancestry-associated DE remains unknown. Thus, we asked if the associations between differential activity scores and differential expression were driven by genes whose top diff-CRE is a DE promoter. We found some evidence of this among diff-ATAC promoters across the non-infected and flu PBMC cell types [40] (Additional file 1: Fig. S14b), but no enrichments passed correction for multiple tests. We observed similar strengths of enrichment across the remaining contexts and score types for top diff-CRE enhancers and promoters (Additional file 1: Fig. S14a–b); however, given only eleven significant enrichments when testing all CREs together we were likely underpowered to address this question.

Since the activity levels of enhancers and promoters usually increase and decrease with the expression levels of their target genes, we hypothesized that true enhancer-gene pairs would have higher expression in the same ancestry as that of the populations with higher ATAC and ChIP scores (Fig. 2c) and that this matching directionality would hold for pairs that are poised for DE in other contexts. To test this, we asked if among differential genes (diff-score FDRs = 0.093 and 0.057 for ATAC and ChIP, respectively, with differential expression local false sign rate (LFSR) < 0.05) the ancestry direction of the top diff-CRE matched the DE direction of its target gene more often than expected by chance (see Methods). For example, is a gene with higher AFR ancestry expression also likely to have higher ATAC scores in AFR populations? We found that differential gene directionality matched more often than expected by chance in the same contexts in which differential activity and DE genes overlapped more often than expected by chance, as well as in three additional contexts for diff-ATAC (hypergeometric OR = 1.71–2.54, $P < 3.3 \times 10^{-4}$) and five for diff-ChIP (OR = 2.90–3.70, $P < 4.8 \times 10^{-4}$). Five PBMC contexts [40] were nominally enriched for diff-ChIP matching DE (OR = 1.63–2.12, $P < 0.05$), though not significantly after multiple test correction. This was in contrast to genes identified by diff-ATAC CREs, which were only nominally enriched in four PBMC contexts [40] at lower odds ratios (OR = 1.64–1.82, $P < 0.038$, Fig. 2d; see also Additional file 1: Fig. S13d). We found much weaker enrichment for matching DE directionality again among diff-ABC and HiC genes (see Additional file 1: Fig. S13d, S15).

To better ascertain the relative capacities of diff-ATAC and diff-ChIP (H3K27ac) to identify DE genes and their directionality, we compared the odds ratios across all contexts for all CREs and partitioned by promoter and enhancer top diff-CRE status. We found diff-ATAC to enrich better for DE (Wilcoxon $P = 0.0022$), whereas diff-ChIP enriched substantially better for matching DE direction (Wilcoxon $P = 3.5 \times 10^{-6}$) (see Additional file 1: Fig. S16, left). This difference held when considering enrichments derived from genes whose most differential CRE was a promoter but not an enhancer (see Additional file 1: Fig. S16, right and middle). While the diff-ChIP score incorporates H3K27ac levels from each target gene's promoter for enhancer CREs (Fig. 1b, see the "Methods" section), diff-ATAC enhancers, which do not explicitly incorporate promoter accessibility information, performed similarly to diff-ChIP enhancers at identifying DE direction (see Additional file 1: Fig. S14c–d, S16). These results indicate that of the two types of chromatin activity assayed in our study, accessibility is the better indicator of which genes are DE, while H3K27ac levels better identify which ancestry has higher expression of these genes across numerous cellular contexts.

Differential CRE activity is associated with ancestry-divergent variants that affect binding of specific TFs

Although diff-ATAC and diff-ChIP CREs are associated with DE of their target genes, the mechanism behind this association is unclear. To investigate this we sought to link potentially causal genetic variants to the activity of our CREs by intersecting them with QTL for the binding affinity of five transcription factors (bQTL, Fig. 3a) and H3K4me3 levels (H3K4me3 QTL) previously mapped in the same YRI LCLs used in our study [18]. If differential CRE activity were driven in *cis* by any of these QTL types, as opposed to in *trans* by a difference in transcription factor expression level, we would expect strong associations between those QTL and differential activity CREs. We followed the same approach as in our DE analysis, first testing if our diff-CREs were enriched for any of these QTL relative to non-diff-CREs (Fig. 3a, see the "Methods" section). Any diff-CRE was counted as a "success" overlap in hypergeometric enrichment tests if it contained a bQTL for the TF being tested (or H3K4me3 QTL). We found several significant bQTL enrichments across diff-ATAC and diff-ChIP CREs (Fig. 3b). We further asked if these enrichments were driven by bQTL in enhancers or promoters by performing separate tests on these two CRE types. Interestingly, enrichments for bQTL became even stronger when considering diff-ATAC and diff-ChIP enhancers, while diff-promoters showed no enrichments for any TF across all score types (Fig. 3c). This was despite greater coverage at promoters than at enhancers in both ATAC and HiChIP data (Additional file 1: Fig. S2a–b). These results suggest that many ancestry differences in CRE activity could be associated with differences in binding of specific TFs in *cis*.

To investigate the extent to which higher TF binding affinity corresponds to an increase in CRE activity, we asked if the high-affinity bQTL allele was at higher frequency in the ancestry with higher CRE activity (Fig. 3d, "matching direction diff-CRE bQTL"). We also included bQTL for CTCF [42], a protein that mediates chromosomal looping and chromatin, in these tests. The same TFs (JunD, NF-κ B, and PU.1) were enriched for bQTL matching diff-CRE directionality as were enriched in diff-CREs overall, with the addition of STAT1 for matching diff-ATAC direction. Interestingly, PU.1 bQTL were enriched for matching diff-ATAC direction (Fig. 3e, left), but not diff-ChIP direction (Fig. 3e, right). This was in contrast with this TF's overall bQTL enrichment in diff-ChIP CREs over non-diff (Fig. 3b–c), suggesting that this TF's activity could be linked to context-dependent increases and decreases in H3K27ac levels, but is associated with increased chromatin accessibility in both cases.

If increased TF binding at bQTL is associated with an increase in CRE activity in *cis*, we should see an increase in correspondence between the ancestry with a higher frequency of the high-affinity bQTL allele and the ancestry with higher CRE activity the more extreme the difference in allele frequencies between ancestries. To test this, we asked if enrichments for matching directionality between bQTL and diff-CREs increase when considering only bQTL in the top 5% of F_{ST} among variants in CREs (corresponding to $F_{ST} > 0.1813$). Indeed, for all TFs with direction matching-enriched bQTL under no F_{ST} thresholding we observed an average 2.36-fold increase in odds ratios when applying this F_{ST} threshold (Fig. 3e). These enrichments were again driven by enhancers, as evidenced by the average 3.41-fold increase in odds ratios for the same comparison restricted to this CRE type (see Additional file 1: Fig. S17, top) and lack of directionality

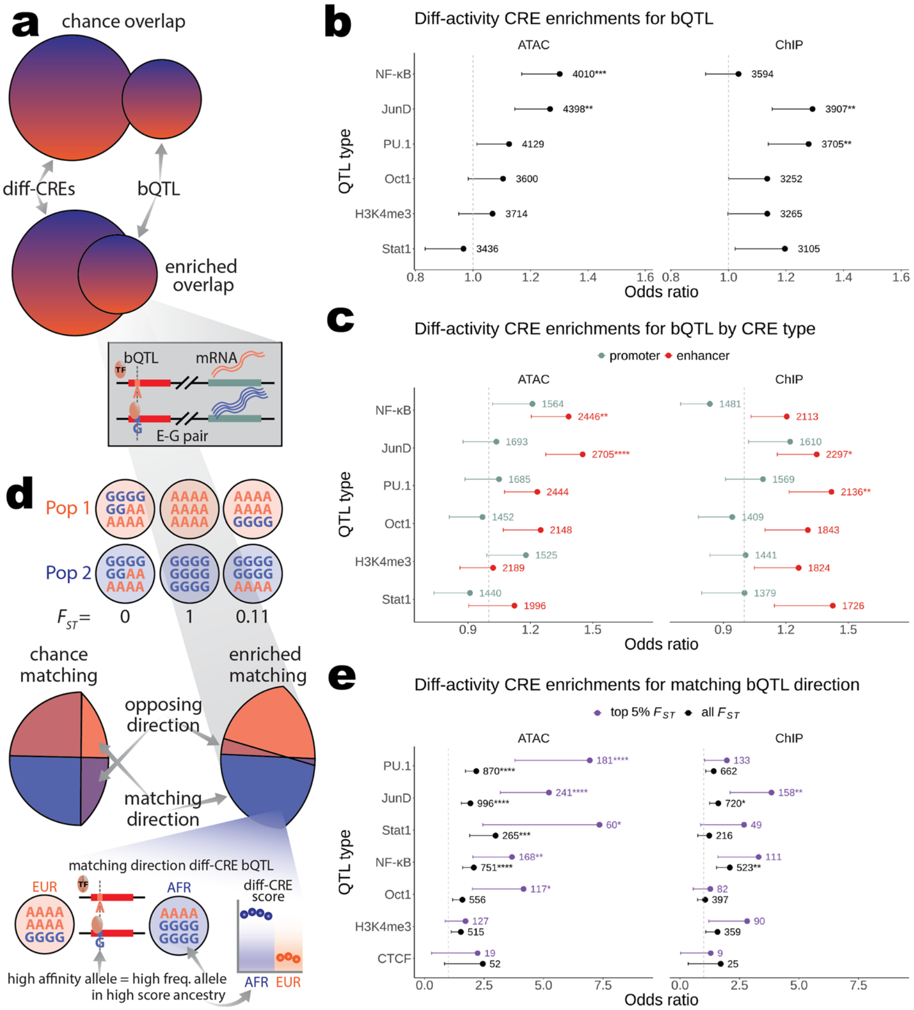


Fig. 3 Enhancer differential activity is linked to sequence-dependent differences in NF-κ B and JunD binding. **a** Schematic depicting a simplified version of the enrichment test results shown in **b**. The boxed gradient callout from the enriched overlap shows a hypothetical example of a bQTL where the "G" allele increases the binding affinity of the TF, which leads to elevated mRNA levels through increased transcription of the target gene. **b** TF bQTL and H3K4me3 QTL enrichments in diff-ATAC, ChIP, and HiC CREs are plotted as Fisher's exact test odds ratios with error bars representing the lower bound of the 95% confidence interval. Since these are one-sided tests, the upper bound (infinity) is not shown. The total number of CREs used in each test is shown to the right of each odds ratio with asterisks indicating if the *P*-value passed multiple test correction (*, **, ***, and **** for Bonferroni-corrected *P*-value < 0.05, 0.005, 5×10^{-4}, and 5×10^{-5}, respectively). **c** Results of tests on the same CREs used in **b** separated by their status as enhancer or promoter CREs. **d** Schematic depicting the directionality matching enrichment test results shown in **e**. Top, three hypothetical variants with "A" and "G" alleles in two populations where the first (leftmost) has no difference in allele frequency between populations or F_{ST}=0, the second (middle) has the maximum possible difference in allele frequency between populations ("A" is fixed in population 1 and "G" is fixed in population 2) or F_{ST}=1, and the third (rightmost) has the "G" allele at intermediate frequency in population 2 or F_{ST}≈0.11 in this case. Middle, diff-CRE bQTL from overlaps in **a**–**c** (gray gradient callout) were tested for enriched matching of ancestry-associated direction. Bottom, schematic of a matching direction diff-CRE bQTL. The blue gradient callout shows an example of a matching direction diff-CRE bQTL where the high affinity "G" allele from the same hypothetical bQTL shown in **a** is at higher frequency in AFR, which is the ancestry that has the higher ATAC or ChIP scores defining the diff-CRE. **e** Diff-CRE bQTL matching direction enrichment test results for diff-CREs defined by each displayed score type are plotted as Fisher's exact test odds ratios with error bars representing the lower bound of the 95% confidence interval. Separate tests were performed on all diff-CRE bQTL (black) and those in the top 5% of F_{ST} values among all variants in CREs genome-wide (purple). Since these are one-sided tests, the upper bound (infinity) is not shown. The total number of CREs used in each test is shown to the right of each odds ratio with asterisks indicating if the *P*-value passed multiple test correction (*, **, ***, and **** for Bonferroni-corrected *P*-value < 0.05, 0.005, 5×10^{-4}, and 5×10^{-5}, respectively)

matching enrichment for any TF's bQTL in promoters (see Additional file 1: Fig. S17, bottom). As expected, none of the above bQTL enrichment tests were significant for nominally diff-HiC CREs (Additional file 1: Fig. S18).

Having established ancestry-dependent *cis* differences in TF binding as a possible mechanism for ancestry-associated differential CRE activity, specifically in enhancers, we sought to assess the likelihood that these TF bQTL have been under directional selection. We found that JunD, NF-κ B, PU.1, and Oct1 bQTL have higher F_{ST} in diff-ATAC than in non-diff-ATAC enhancers (Wilcoxon $P = 2.4 \times 10^{-9}$, $5.0 \times 10{-}^{4}$, 2.2×10^{-4}, and 6.0×10^{-4}, respectively), consistent with differential binding of these TFs as drivers of differential enhancer activity, as well as the possibility that their binding specifically in differential activity CREs has been subject to selection. While none reached significance in diff-ChIP enhancers after multiple test correction, all of the QTL types except CTCF bQTL were nominally significant (Fig. 4a, top). This is likely due in part to power reduction in the diff-ChIP test due to the combination of lower resolution of HiChIP *cis* interaction pairs relative to ATAC-seq peaks (see the "Methods" section) and only counting the most significant bQTL per diff-score CRE (see Methods). Notably, there were no significant differences between bQTL in diff- versus non-diff promoters after multiple test correction (Fig. 4a, bottom; see Additional file 1: Supplemental text, Fig. S19). To assess evidence for selection on bQTL in diff-enhancers over those in diff-promoters more directly we performed the same test within diff-CREs between enhancers and promoters. Nearly all QTL types had higher median F_{ST} in diff-enhancers than in diff-promoters for ATAC and ChIP although none were significant after multiple test correction (Fig. 4b). Again as expected, there was no difference in F_{ST} between nominally diff- and non-diff-CREs or diff-CRE enhancers and promoters defined by HiC scores (Additional file 1: Fig. S20).

Since F_{ST} can be correlated with allele frequency (i.e., rare alleles introduced by recent mutation have low F_{ST}), we sought to assess whether higher F_{ST} for diff-enhancer bQTL was driven by differences in allele frequencies between CRE types. Performing the same tests in each of ten allele frequency decile bins, we find more enhancer bins than promoter bins with mean F_{ST} greater in diff- versus non-diff CREs (see Additional file 1: Fig. S21–22). Additionally, although binning reduces the power of each test, more of these bins have nominally significant differences in F_{ST} between diff- and non-diff enhancers. These results suggest that greater allele frequency divergence in differential activity enhancers is not dependent on allele frequency differences between the tested CRE types. Overall, these higher F_{ST} values for select bQTL in diff-enhancers are consistent with selection on TF binding sites in our diff-ATAC and diff-ChIP CREs.

Differential CRE activity could be a result of directional selection and/or genetic drift

While these results could reflect directional selection, the underlying divergence in allele frequencies and corresponding ancestry-associated differential CRE activity could still be explained by genetic drift. More convincing evidence of directional selection could result from applying the sign test framework [43, 44] to ask if the high-affinity alleles for bQTL that match diff-CRE direction are at higher frequency in one ancestry over the other more often than expected by chance. The sign test leverages the expectation that under neutrality, where genetic drift is the dominant force operating on allele

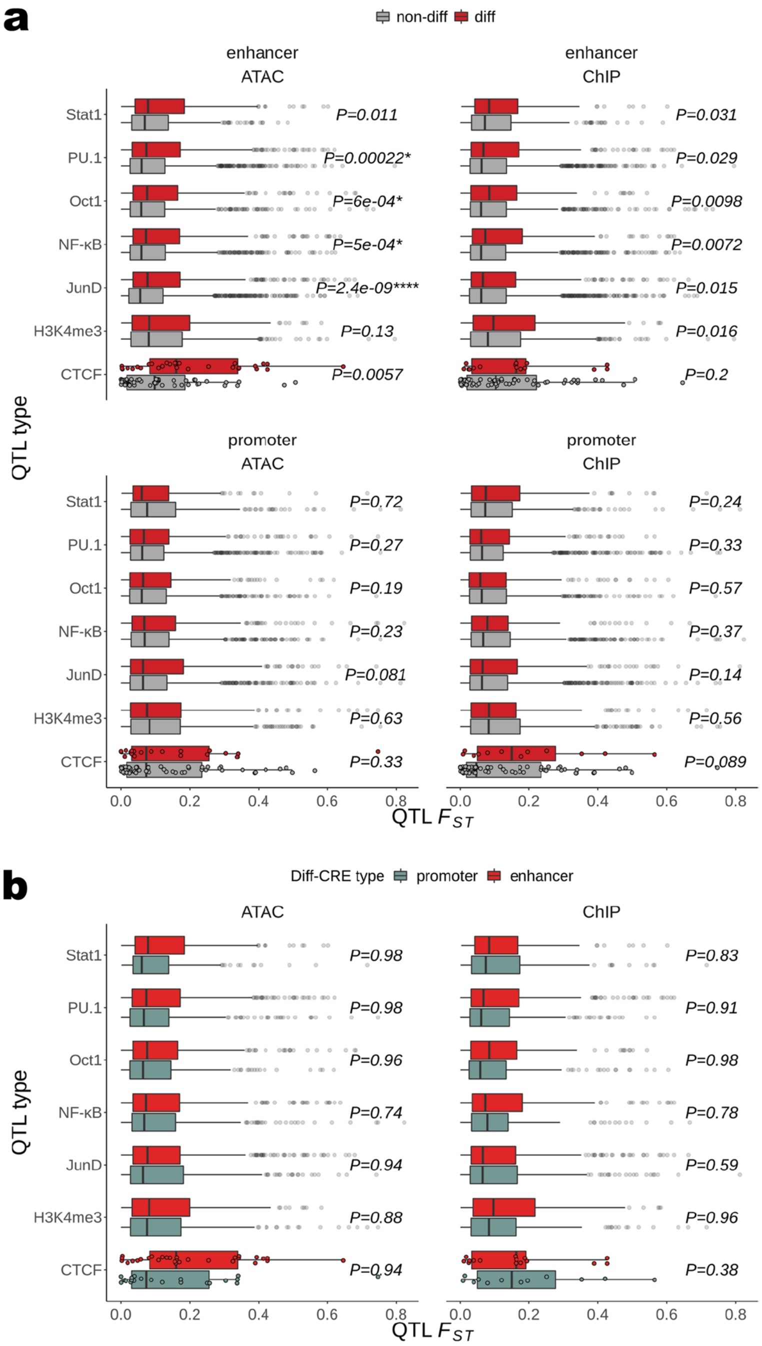


Fig. 4 Evidence for selection on TF binding in enhancers versus promoters. **a** F_{ST} values of TF bQTL and H3K4me3 QTL are shown as boxplots for diff- and non-diff CREs of each displayed score type separated by if the QTL is in an enhancer (top) or promoter (bottom). *P*-values from the one-sided Wilcoxon test on the F_{ST} distributions from diff-enhancer versus diff-promoter are displayed. **b** F_{ST} values of TF bQTL and H3K4me3 QTL are shown as boxplots for enhancer and promoter diff-CREs of each displayed score type. *P*-values from the one-sided Wilcoxon test on the F_{ST} distributions from diff-enhancer versus diff-promoter are displayed. None of the P-values pass multiple test correction, so no asterisks are displayed. For visualization, non-outlier F_{ST} values are only plotted over boxplots for distributions with fewer than 200 points

frequency in populations of both ancestries, the high-affinity alleles matching diff-CRE direction will not be biased toward higher frequencies in either population (Additional file 1: Fig. S23a). Among bQTL matching diff-CRE direction, we found no more population-specific allele frequency bias than expected relative to the background of each TF's bQTL in all CREs (Additional file 1: Fig. S23b). Thus, genetic drift could be responsible for the association with increased ancestry divergence in diff-CREs matching bQTL directionality.

Moving from genotype toward phenotype (Fig. 1a, left), we next sought to identify the functional pathways most closely associated with our diff-score enhancer-gene pairs and their ancestry directionality. If variants in any subset of diff-CREs linked to target genes associated with a particular pathway have been subject to lineage-specific selection, these may not have been detected in our previous analyses. To address this possibility, we used gene set enrichment analysis with the gene ontology (GO) biological processes and MSigDB Hallmark gene sets on genes ranked by the difference in means between ancestries in ABC component scores of their top diff-CREs scaled by a measure of score variance (i.e., ranked from high EUR activity to high AFR activity, see the "Methods" section). Again, under neutrality, we would not expect diff-CREs with target genes in a particular pathway to have higher activity in one ancestry over the other. We did not find any significant enrichments among these gene sets after multiple test correction; however, some immune-related gene sets including interferon gamma (IFNG) response and TNF-α signaling via NF-κ B were among the top nominal enrichments for genes with top diff-ChIP and/or diff-ATAC CREs in the AFR high activity direction (Additional file 1: Fig. S24). These nominal enrichments are consistent with diff-ATAC and diff-ChIP target gene enrichments for DE genes and matching DE directionality in IFNG-exposed LCLs (Fig. 2b,d, left), and Randolph et al.'s [40] finding of TNF-α signaling via NF-κ B enrichment among genes with higher AFR expression in monocytes both before and after flu infection. Thus, although we do not find strong evidence for lineage-specific selection on diff-CREs in aggregate, the possibility that selection has more subtly affected gene regulatory architecture remains.

Discussion

We have presented results from the first genome-wide comparison of chromatin activity and contact frequency between human populations with the goal of identifying CREs under recent selection. Since recent evidence points toward gene expression changes as the dominant force shaping recent human adaptation relative to protein sequence changes [16, 45], this approach has the potential advantage of directly identifying CREs responsible for adaptive gene expression differences.

Using ABC scores to link CREs to target genes and decomposing these scores into their components allowed us to identify genes whose ancestry-associated expression differences across multiple contexts could be identified by the differential activity of their enhancers in the context of LCLs at baseline. This was particularly true for identifying the ancestry-associated direction of DE. Although H3K27ac alone is not required to maintain CRE activity [46], it seems to be a more reliable indicator of expression direction than chromatin accessibility as measured by ATAC-seq in the context of our study. For example, one of many models capable of explaining this difference would be the binding of a transcriptional repressor to a promoter that yields an increase in chromatin accessibility but not in H3K27ac levels. About 25% of ABC-predicted and validated enhancer-gene pairs were found to have repressive effects via CRISPRi-flowFISH [29] and any such effects within the matching DE directionality enrichments from our study could have contributed to the 39% of differential activity pairs that "opposed" DE direction. More generally, the strength of these cross-context enrichments for DE and its direction is consistent with the maintenance of ancestry-associated regulatory differences in contexts beyond those where the target genes are DE. Matching differential CRE activity in LCLs at baseline and DE in many other contexts suggests CRE poising for DE regulation upon stimulation or differentiation to another cell type, or footprints of regulatory activity from a previous cell state remaining after the transition from that state.

Although our bQTL enrichment results suggest that differential activity is a result of cis-regulatory activity, it is possible that transcription factor differential expression in *trans* partially accounts for this. Indeed, JunD and NFKB2 (NF-κ B subunit 2 of 2) show AFR-biased expression in LCLs at baseline (ancestry effect $\beta = -0.26$, LFSR = 0.0026 and ancestry effect $\beta = -0.18$, LFSR = 0.073, respectively); however, given the high odds ratios for bQTL in the top 5% of F_{ST} (Fig. 3e, Additional file 1: Fig. S17), differential CRE activity would likely persist even under constant *trans* conditions. Moreover, the lack of enrichments among diff-ATAC and diff-ChIP promoters for bQTL over non-diff (Fig. 3c), matching bQTL directionality irrespective of F_{ST} (Fig. S17, bottom), and high bQTL F_{ST} over non-diff (Fig. 4a), all relative to the positive enrichments found for tests on enhancers (including Fig. 4b) are consistent with greater evolutionary constraint on promoters and the distinct roles of enhancers in cell types that may be subject to different selection pressures [47]. Notably, while diff-ChIP enhancers and promoters both identified DE direction (Additional file 1: Fig. S14c), these results suggest that if JunD and/or NF-κ B are responsible for any of these expression differences, it is due to differences in their binding at enhancers, rather than at promoters. Moreover, we find similar proportions of diff-ATAC and diff-ChIP enhancers versus promoters (20% versus 18%, and 10% versus 9%, respectively) indicating similar levels of differential signal present in each across both methods. This genotype-level evidence restricted to differential enhancers indicates that our method of using chromatin as a spotlight on genetic variation effectively reveals otherwise hidden patterns consistent with selection (Fig 1a, left).

While our tests for greater transcription factor binding in one ancestry over the other did not show evidence of lineage-specific selection, the most enriched pathways among genes linked to higher activity CREs in AFR suggest more subtle effects of directional selection. For example, if the IFNG response pathway was under selection in one ancestry and this selection acted on a fraction of differential activity CREs regulated by transcription factor complexes more tissue- and/or response-specific than JunD or NF-κ B, this could remain undetected when aggregating over many more CREs. Importantly, any ancestry-associated differences that may exist in the regulation of these pathways as a result of selection or drift do not imply differences in underlying cellular and physiological mechanisms. Independent of these considerations, our study is limited by any changes to genome architecture introduced by Epstein-Barr virus in transforming B cells into LCLs that further mask the effects of any selection that has acted on B cells or even more relevant cell types and the noise introduced by combined analysis of multiple datasets generated by different people and/or labs. Future studies generating ABC score component data from diverse donors in cellular contexts more like those in which lineage-specific selection could have acted may find stronger evidence of it, especially if bQTL are mapped for more context-specific transcription factors.

The demographic processes that shape human genetic variation (e.g., population history, migration, and drift) can obscure the influence of selection on variants that underlie adaptive phenotypes [48]. Moreover, false signals of selection can result from under-controlled population stratification [49, 50]. These confounders along with the prevalence of adaptive variants in non-coding regions with subtle effects [16] demonstrate the need for complementary methods to identify CREs that have been subjects of selection. We anticipate that extending the application of the method presented here to more populations and cell types will elucidate the molecular underpinnings of recent human evolution with implications for understanding modern disease prevalence.

Conclusions

In generating the first population-level maps of candidate enhancer-target gene pairs in humans, we suggest *cis*-regulatory elements are poised for ancestry-dependent differential expression regulation upon stimulation or differentiation to another cell type. Mechanistically, this poising could be maintained by variants affecting the binding of transcription factors NF-kB, JunD, and PU.1 that show signs of lineage-specific selection in enhancers but not promoters. The potential effects of directional selection on immune-related pathways identified here suggest the promise of applying our chromatin-level selection test in additional cell types with roles in these pathways.

Methods

Cell culture and ATAC-seq

For detailed methods on cell culture conditions and processing, see our previous study [19]. Briefly, 2×10^3 cells from each LCL were collected and pooled by population after growth to $6–8\times10^5$ cells/mL. To prevent disproportionate cell line growth within pools throughout the collection and pooling process, sub-pools were frozen in liquid nitrogen at −180 °C. After collection of all LCLs, sub-pools were combined by population, and cells from each of the 10 pools were purified, isolated, and split into two replicates of 10^5 cells each and pelleted according to [19] for a total of 20 samples. ATAC-seq was performed using the protocol from [23] in which each sample was resuspended in 100 μl of transposition mix containing 5 μl of Tn5 Transposase and incubated in a ThermoMixer for 30 min at 37 °C and 750 rpm. Transposed DNA fragments were then eluted and PCR-amplified with total cycles determined according to [23]. Following two PCR cleanup steps, purified ATAC-seq libraries were sequenced on an Illumina HiSeq 4000 to generate 2×150 bp, paired-end reads.

HiChIP

We thawed each −180°C-stored sub-pool described above and in [19] on ice, combined them by population and removed dead cells. As for ATAC-seq, to avoid disproportionate cell line growth we did not passage the cells before or after combining sub-pools. We then split each population pool into 2 replicates for crosslinking and HiChIP. For more detailed HiChIP methods, see [30]. Briefly, cells from each pool were pelleted and resuspended in 1% formaldehyde (Thermo Fisher) for crosslinking at a volume of 1 ml per million cells with incubation at room temperature for 10 min with rotation. Formaldehyde was then quenched with glycine at a 125-mM final concentration with 5 min room temperature incubation with rotation. Cells were then pelleted, PBS-washed, re-pelleted, and either used immediately in the HiChIP protocol or stored at −80 °C for HiChIP later.

HiChIP was performed as described in [25] with H3K27ac antibody (Abcam, ab4729) and the following modifications. We used a 2 min sonication time, 2 µg of antibody, 34 µl of Protein A beads (Thermo Fisher) for chromatin-antibody complex capture. Post-ChIP Qubit quantification was performed to determine the amount of Tn5 used and a number of PCR cycles performed for library generation, accounting for varying amounts of starting material. We performed size selection by PAGE purification (300–700 bp) to remove primer contamination and sequenced all libraries on an Illumina HiSeq 4000.

ATAC-seq read mapping

For the complete mapping pipeline see https://github.com/kadepettie/popABC/tree/master/mapping, which contains a nextflow implementation of the steps described in [19]. Cutadapt was used to remove sequencing adapters (arguments: -e 0.20 -a CTGTCTCTTATACACATCT -A CTGTCTCTTATACACATCT -m 5). PCR duplicate reads generated during library preparation were removed using picard MarkDuplicates (v2.18.20) (http://broadinstitute.github.io/picard/) (arguments: SORTING_COLLECTION_SIZE_RATIO=.05 MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000 MAX_RECORDS_IN_RAM=2500000 OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 REMOVE_DUPLICATES=true DUPLICATE_SCORING_STRATEGY=RANDOM). To minimize allelic mapping bias, a modified version (https://github.com/TheFraserLab/WASP/tree/atac-seq-analysis/mapping) of the WASP pipeline [38] was used for read mapping. Reads were aligned to the hg19 genome using bowtie2 [51] (arguments: -N 1 -L 20 -X 2000 --end-to-end --np 0 --n-ceil L,0,0.15) and filtered to a minimum mapping quality of 5 using samtools (v1.8) [52].

HiChIP read mapping

HiChIP reads were mapped using the nf-core [53] HiC-Pro [54] mapping pipeline (https://github.com/nf-core/hic) modified to include the same version of the WASP pipeline as was used for ATAC-seq to minimize allelic mapping bias (https://github.com/kadepettie/popABC/tree/master/hicpro). In this version, however, allele-swapped remapping was performed separately on each read end, after which reads were re-paired, to accommodate the long-range nature of the paired-end reads as in the original HiC-Pro pipeline. Filtering reads down to valid *cis* interaction pairs, we took the raw 5-Kb resolution contact maps (the ".matrix" and corresponding ".bed" file output from process "build_contact_maps") as input to our differential activity-by-contact pipeline.

Differential activity-by-contact

Candidate element definition

We used the ABC model (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) to predict enhancer-gene connections in each pooled LCL population replicate (sample), with modifications to facilitate comparison of AFR and EUR population samples. For the complete differential activity-by-contact (diff-ABC) pipeline see https://github.com/kadepettie/popABC/tree/master/selection_1000G ("ABC_pipeline.nf"). We used Genrich (v0.5_dev, available at https://github.com/jsh58/Genrich) to call AFR and EUR ATAC-seq peaks jointly on the 8 samples from each with default parameters except for the following: -y -j -d 151. We then summed the reads in each peak across the corresponding 8 samples, kept the top 150,000 by read count, and resized them to 500 bp centered on the peak summit. To ensure equal contribution from peaks called separately in AFR and EUR to our candidate element input to the ABC model, we again made separate rankings by read count for each, then interleaved the two lists evenly by ranking, merging any overlaps, and taking the top 150,000 elements. We next added 500 bp gene TSS-centered regions and removed any from the resulting list that overlapped regions of the genome with known signal artifacts (https://sites.google.com/site/anshulkundaje/projects/blacklists) [55, 56]. Overlapping regions resulting from summit extensions and/or TSS additions were merged immediately following each of these steps. We defined promoter elements as those within 500 bp of an annotated TSS and the rest as enhancer elements.

Score component normalization

To ensure comparability of ABC scores between populations and replicates, and particularly samples with differing signal-to-noise ratios, we quantile normalized ATAC-seq reads per million sequenced reads (RPM), HiChIP valid *cis* interaction pair counts from 5 kb bins overlapping CREs, and each of these bins' total count (for ChIP score computation) to the mean of their respective distributions across samples separately for enhancers and promoters. Since there is a lack of consensus on an appropriate library size normalization method for HiChIP data, due to the violation of the assumption of equal visibility between interacting regions often used in HiC normalization [54, 57, 58], we relied on the combination of quantile normalization and subsequent score normalization steps to control for library size and other technical artifacts.

Quantitative HiChIP signals were computed using the quantile normalized HiChIP contact counts according to [29]. Briefly, for each gene TSS, all contact counts in CREs within 5 Mb were normalized to sum to one, then divided by the maximum of these values to normalize for comparison across genes.

Score computation

As in [29], we computed ABC scores using H3K27ac HiChIP with the fraction of regulatory input to gene *G* contributed by element *E* given by:

$$\text{ABC score}_{E,G} = \frac{A_E \times Q_{E,G}}{\sum_{\text{all elements e within 5 Mb of } G} \left(A_e \times Q_{e,G}\right)}$$

Here, the activity component (A) is quantile normalized ATAC-seq RPM, as in the original ABC score formula, but we have replaced the HiC contact component (C) with the quantitative HiChIP signal (Q) described above. We computed ATAC scores as follows:

$$\text{ATAC score}_{E,G} = \frac{A_E}{\sum_{\text{all elements } e \text{ within 5 Mb of } G} A_e}$$

We computed ChIP scores as follows, using the geometric mean of the quantile normalized HiChIP bin totals overlapping each element (vanilla coverage square root (VC-sqrt)) to estimate the aggregate H3K27ac signal at both elements:

$$\text{ChIP score}_{E,G} = \frac{\sqrt{H_E \times H_G}}{\sum_{\text{all elements } e \text{ within 5 Mb of } G} \sqrt{H_e \times H_G}}$$

where H is the total quantile normalized valid *cis* interaction pair count from the HiChIP bin overlapping element E or the promoter of gene G. VC-sqrt normalization is commonly applied to HiC data for comparison of contact frequencies across samples since the assumption of equal visibility is reasonable when considering data generated from proximity-based ligation alone (i.e., without ChIP). When applied to HiChIP, the VC-sqrt measures the difference in visibility between interacting regions relative to one another within a sample that is due to the levels of H3K27ac present at each region. Thus, when normalized by the sum of this signal across all elements within 5 Mb of the target gene, the resulting ChIP score reflects the contribution of H3K27ac levels to an ABC score. We can then use VC-sqrt normalization to estimate the contact frequency between each element independent of H3K27ac levels and extend this to compute the HiC component of an ABC score as follows:

$$\text{HiC score}_{E,G} = \frac{\frac{C_{E,G}}{\sqrt{H_E \times H_G}}}{\sum_{\text{all elements } e \text{ within 5 Mb of } G} \frac{C_{e,G}}{\sqrt{H_e \times H_G}}}$$

where C is the quantile normalized number of valid *cis* interaction pairs connecting the HiChIP bin overlapping element E and the promoter of gene G.

E-G pair definition

To perform differential ABC score analysis across ancestries, we took predictions from the ABC model for each sample (population and replicate) and processed them according to the following steps. First, we excluded pairs with ABC score < 0.015 in all samples to avoid testing pairs unlikely to be true regulatory connections in any population [31]. Second, we excluded promoter-gene pairs with ABC scores below a stringent threshold of 0.1 because experimental data has shown the ABC model has poorer performance for this class of interactions, likely due to transcriptional interference, *trans* effects, and/or promoter competition [29]. Third, we required each enhancer-gene pair to be supported by non-zero quantile-normalized HiChIP contacts and ATAC values at the CRE in all samples, to avoid testing pairs where low ABC scores could be driven by mapping biases or low sequencing depth. Due to the difference in sequencing depths between samples, this final filtering step reduced the number of enhancer-gene pairs under consideration from 580,474 to our final set of 52,454 after removing CEU from the enhancer-gene pair-calling pipeline.

Clustering analysis

For each score type and enhancer-gene pair, values were z-score normalized across samples for comparison and visualization of enhancer-gene pairs with large differences in mean score. PCA was performed with "prcomp" and heatmaps were generated using the *pheatmap* package (v1.0.12) in R (v4.1.0).

Differential analysis

We called diff-CREs using unpaired, two-sample t-tests on each score type in AFR versus EUR samples. Log_2 fold change effect sizes were estimated as the $\log_2$-ratio of the mean EUR score over the mean AFR score. We estimated a false discovery rate (FDR) for each score type at t-test $P < 0.05$ as the ratio of expected over observed enhancer-gene pairs with $P < 0.05$, where the null P-value distribution was derived from unpaired, two-sample t-tests on one set of replicates from each population versus the other. Replicate number was randomized for each enhancer-gene pair. To estimate the null P-value distribution for tests with eight AFR and six EUR samples after CEU removal while maintaining the eight versus six sample structure of each test, one AFR population was held out at random from replicate shuffling for each enhancer-gene pair and both replicates from this population were used in the group of eight (as opposed to the seven versus seven structure that would result from splitting by replicate across all populations). Since ChIP score signal is derived from HiChIP contact count bins at 5 Kb resolution, we counted diff-ChIP and HiC for CREs from the same HiChIP bin as one in each diff-score enrichment test described below.

DE enrichments

We used hypergeometric tests (i.e., one-sided Fisher's exact tests) to determine enrichments for DE target genes among diff-CREs and matching ancestry directionality among DE genes with a diff-CRE. For the former across score types, we took the most differential CRE (top diff-CRE) by the corresponding metric (i.e., ABC, ATAC, ChIP, or HiC score) per gene, defining diff-CREs at nominal t-test P-values < 0.05, non-diff-CREs at t-test P-values ≥ 0.5, and DE genes at LFSRs < 0.05 [40, 59, 60]. Then, counting each CRE only once, we classified diff-CRE hits as any with at least one DE target gene, diff-CRE non-hits as any with no DE target genes, non-diff-CRE hits as any with no DE target genes, and non-diff-CRE non-hits as any with at least one DE target gene. For the promoter test, we took the subset of promoter CREs and additionally required diff-CRE hits and non-diff-CRE non-hits to be promoters for at least one of their DE target genes. For the enhancer test, we allowed promoter CREs to be classified as enhancers if they were not promoters for the relevant gene(s) (e.g., a distal promoter for another gene contacting the promoter of the DE gene under consideration). That is, we required diff-CRE hits and non-diff-CRE non-hits not to be promoters for any of their DE target genes.

For the matching direction tests, we took the subset of top diff-CREs with DE target genes where all DE target genes were in the same direction (AFR- or EUR-biased) and, again counting each CRE only once, classified hits as diff-CREs with higher scores in the same ancestry as that with higher expression in their DE target gene(s). For the promoter and enhancer tests, we required diff-CREs to be promoters for at least one of their DE target genes and none of their DE target genes, respectively. For each set of tests, we only report P-values in the main text that pass Bonferroni-corrected thresholds.

TF bQTL and H3K4me3 QTL enrichment analysis

We used hypergeometric tests to determine enrichments for each QTL type among diff-CREs relative to non-diff-CREs and matching ancestry directionality among diff-CREs with a QTL, using the same definitions for diff- and non-diff-CREs as in our DE enrichment analyses. In each test, we considered the QTL with the lowest P-value per CRE for CREs with multiple QTL of the given type. For the directional analyses, we defined bQTL directionality as AFR if the high-affinity allele was present in AFR at a greater frequency than in EUR and vice versa. For CREs with multiple bQTL, we additionally required that they all match the direction for inclusion in each test. For binomial sign tests (see Additional file 1: Fig. S23a-b), we performed two-sided binomial tests on the number of QTL matching directionality in diff-CREs in the AFR direction out of the total number matching direction in diff-CREs with a null probability of this proportion across all CREs.

GO analysis

We used the R package fgsea (v1.20.0) [61] to perform gene set enrichment analysis on genes ranked by the value of their most differential CRE according to the following T-test statistic [62]:

$$\frac{\mu_{EUR} - \mu_{AFR}}{\sqrt{\frac{\sigma^2_{EUR}}{n_{EUR}} + \frac{\sigma^2_{AFR}}{n_{AFR}}}}$$

Where μ is the mean score, σ is the standard deviation, and n is the number of samples for each ancestry. fgsea was run on these ranked lists for each score type using the C5 GO biological processes and MSigDB Hallmark gene sets with default arguments except: minSize = 15, maxSize = 500.

F_{ST} analysis

F_{ST} for all variants was obtained using VCFtools' calculation of Weir and Cockerham F_{ST} [9] between individuals from the African (ESN, GWD, LWK, and YRI) and European (CEU, FIN, IBS, and TSI) populations in our ATAC-seq and HiChIP data on a per-site basis. Variants with NA values were removed and negative estimations were adjusted to zero. For diff- versus non-diff CRE F_{ST} Wilcoxon tests independent of their containing bQTL or H3K4me3 QTL, we took the maximum F_{ST} value per CRE.

To control for possible allele frequency differences in our diff- versus non-diff CRE bQTL F_{ST} Wilcoxon tests, we took the combined set of diff- and non-diff CRE bQTL in each test, split them by mean allele frequency across AFR and EUR populations into 10 decile bins, and performed separate tests within each of these bins.

iHS analysis

iHS for all populations were obtained from Johnson and Voight (2018) [63] and overlapped with bQTL in our CREs. For Wilcoxon tests analogous to those in our F_{ST} analysis, we used the maximum iHS observed across all populations.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03165-2.

Additional file 1.

Additional file 2. HiChIP read mapping statistics. Number of reads of each category at each mapping step (percentages are of the total read pairs entering each mapping step).

Additional file 3. CEU scores. ABC, ATAC, ChIP, and HiC scores for both CEU replicates for all E-G pairs after filtering across all 16 samples.

Additional file 4. AFR scores. ABC, ATAC, ChIP, and HiC scores for all African ancestry samples for all E-G pairs after filtering across the 14 samples not including CEU.

Additional file 5. EUR scores. ABC, ATAC, ChIP, and HiC scores for all European ancestry samples for all E-G pairs after filtering across the 14 samples not including CEU.

Additional file 6. Differential score statistics. Results of tests for differential ABC, ATAC, ChIP, and HiC scores between AFR and EUR populations for all E-G pairs after filtering across the 14 samples not including CEU.

Additional file 7. Peer review history.

Acknowledgements

We thank members of the Fraser lab for helpful conversations, advice, and feedback on the manuscript; and Joseph Nasser, Kristy Mualim, and Jesse Engreitz for help with Activity-by-Contact scores.

Review history

The review history is available as Additional file 7 .

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

HBF conceived the study. KPP and HBF conceived analysis methods. HiChIP experiments were performed by MM, MK, and KPP and funded by HYC. KPP performed all analyses and designed all graphics. AJL and JA provided unpublished data. KPP wrote the manuscript with input from all authors. HBF supervised all aspects of the work. All authors read and approved the final manuscript.

Funding

This work was funded by NIH grant R01GM134228. KPP was supported by NIH training grant T32GM007276 and the NSF Graduate Research Fellowship Program. HiChIP data generation was supported by NIH grant RM1-HG007735 to HYC.

Availability of data and materials

Genotype data for all individuals from populations used in our ATAC-seq and F_{ST} analyses is from 1000 Genomes Project Phase 3 release [19] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).
All HiChIP reads are available as fastq files at NCBI SRA, project ID PRJNA898623 [64].
Ancestry associated differential expression data from RNA-seq in LCLs after four-hour exposure to twelve cellular environments is from Supplemental Table S8 of Lea et al. [59] with additional files and analysis code at https://github.com/AmandaJLea/LCLs_gene_exp.
Ancestry associated differential expression data from single cell RNA-seq in PBMCs is from Randolph et al. [40], available at NCBI GEO, Accession no. GSE162632 [65].
bQTL and H3K4me3 QTL are from supplemental Table S1 of Tehranchi et al. [18].
CTCF QTL are from Ding et al [42].
All pipelines and code used for analyses in this paper are available on Zenodo at https://zenodo.org/records/10396417 and on github at https://github.com/kadepettie/popABC/tree/master [66].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 November 2022 Accepted: 4 January 2024
Published online: 15 January 2024

References

1. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature. 2004;428(6984):717–23.
2. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007;39(1):31–40.
3. Manceau M, Domingues VS, Mallarino R, Hoekstra HE. The developmental role of Agouti in color pattern evolution. Science (80-). 2011;331(6020):1062–5.
4. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science (80-). 2012;337(6099):1190–5.
5. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22(9):1748–59.
6. Cano-Gamez E, Trynka G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. Front Genet. 2020;11:424 Frontiers Media S.A.
7. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: from association to function. Am J Hum Genet. 2018;102(5):717–30 Cell Press.
8. Sohail MS, Louie RHY, McKay MR, Barton JP. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. Nat Biotechnol. 2020;2020(30):1–8.
9. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution (N Y). 1984;38(6):1358.
10. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008;180(2):977–93.
11. Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185(4):1411–23.
12. Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, et al. Adaptations to Climate-Mediated Selective Pressures in Humans. Nachman MW, editor. PLoS Genet. 2011;7(4):e1001375.
13. Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, et al. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc Natl Acad Sci U S A. 2010;107(SUPPL. 2):8924–30.
14. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, et al. Adaptation to climate across the Arabidopsis thaliana genome. Science (80-). 2011;334(6052):83–6.
15. Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, Clerici M, et al. Genome-Wide Identification of Susceptibility Alleles for Viral Infections through a Population Genetics Approach. Malik HS, editor. PLoS Genet. 2010 Feb 19;6(2):e1000849.
16. Fraser HB. Gene expression drives local adaptation in humans. Genome Res. 2013;23(7):1089–96.
17. Kaplow IM, MacIsaac JL, Mah SM, McEwen LM, Kobor MS, Fraser HB. A pooling-based approach to mapping genetic variants associated with DNA methylation. Genome Res. 2015;25(6):907–17.
18. Tehranchi AK, Myrthil M, Martin T, Hie BL, Golan D, Fraser HB. Pooled ChIP-Seq links variation in transcription factor binding to complex disease risk. Cell. 2016;165(3):730–41.
19. Tehranchi A, Hie B, Dacre M, Kaplow I, Pettie K, Combs P, et al. Fine-mapping cis-regulatory variants in diverse human populations. Elife. 2019;16:8.
20. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (80-). 2020;369(6509):1318–30.
21. Greenwald WW, Li H, Benaglio P, Jakubosky D, Matsui H, Schmitt A, et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun. 2019;10(1):1054.
22. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, et al. Genetic Control of chromatin states in humans involves local and distal chromosomal interactions. Cell. 2015;162(5):1051–65.
23. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015;2015(1):21.29.1-21.29.9.
24. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.
25. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13(11):919–22.
26. Mumbach MR, Granja JM, Flynn RA, Roake CM, Satpathy AT, Rubin AJ, et al. HiChIRP reveals RNA-associated chromosome conformation. Nat Methods. 2019;16(6):489–92.
27. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47(6):598–606.
28. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, et al. Population variation and genetic control of modular chromatin architecture in humans. Cell. 2015;162(5):1039–50.
29. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9 Nature Research.
30. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet. 2017;49(11):1602–12.
31. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;7(17):1–6.
32. Peng PC, Khoueiry P, Girardot C, Reddington JP, Garfield DA, Furlong EEM, et al. The Role of Chromatin Accessibility in cis-Regulatory Evolution. Zufall R, editor. Genome Biol Evol. 2019;11(7):1813–28.

33. Edsall LE, Berrio A, Majoros WH, Swain-Lenz D, Morrow S, Shibata Y, et al. Evaluating Chromatin Accessibility Differences Across Multiple Primate Species Using a Joint Modeling Approach. O'Neill R, editor. Genome Biol Evol. 2019;11(10):3035–53.
34. Swain-Lenz D, Berrio A, Safi A, Crawford GE, Wray GA. Comparative Analyses of Chromatin Landscape in White Adipose Tissue Suggest Humans May Have Less Beigeing Potential than Other Primates. Lerat E, editor. Genome Biol Evol. 2019;11(7):1997–2008.
35. Reddy SI, Burakoff R. Inflammatory Bowel Disease in African Americans. Inflamm Bowel Dis. 2003;9(6):380–5.
36. Krishnan E, Hubert HB. Ethnicity and mortality from systemic lupus erythematosus in the US. Ann Rheum Dis. 2006;65(11):1500.
37. Ogdie A, Matthias W, Thielen RJ, Chin D, Saffore CD. Racial differences in prevalence and treatment for psoriatic arthritis and ankylosing spondylitis by insurance coverage in the USA. Rheumatol Ther. 2021;8(4):1725.
38. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12(11):1061–3.
39. Yuan Y, Tian L, Lu D, Xu S. Analysis of genome-Wide RNA-sequencing data suggests age of the CEPH/Utah (CEU) Lymphoblastoid Cell Lines Systematically Biases Gene Expression Profiles. Sci Rep. 2015;5(1):1–5.
40. Randolph HE, Fiege JK, Thielen BK, Mickelson CK, Shiratori M, Barroso-Batista J, et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science. 2021;374(6571):1127–33.
41. Karlić R, Chung HR, Lasserre J, Vlahoviček K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010;107(7):2926–31.
42. Ding Z, Ni Y, Timmer SW, Lee BK, Battenhouse A, Louzada S, et al. Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. Gibson G, editor. PLoS Genet. 2014;10(11):e1004798.
43. Fraser HB, Moses AM, Schadt EE. Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc Natl Acad Sci U S A. 2010;107(7):2977–82.
44. Fraser HB. Genome-wide approaches to the study of adaptive gene expression evolution. BioEssays. 2011;33(6):469–77.
45. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014;24(6):885–95.
46. Zhang T, Zhang Z, Dong Q, Xiong J, Zhu B. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. Genome Biol. 2020;21(1):1–7.
47. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. 2015;16(3):144–54.
48. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The Role of Geography in Human Adaptation. Schierup MH, editor. PLoS Genet. 2009;5(6):e1000500.
49. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 2019;1:8.
50. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK biobank. Elife. 2019;1:8.
51. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
52. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):1–4.
53. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276–8.
54. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16(1):259.
55. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9(1):1–5.
56. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
57. Denker A, de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. Genes Dev. 2016;30(12):1357–82.
58. Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, et al. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. PLOS Comput Biol. 2019;15(4):e1006982.
59. Lea AJ, Peng J, Ayroles JF. Diverse environmental perturbations reveal the evolution and context-dependency of genetic effects on gene expression levels. Genome Res. 2022;32(10):1826–39.
60. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet. 2018;26:1.
61. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis bioRxiv. 2021;1:060012.
62. Zyla J, Marczyk M, Weiner J, Polanska J. Ranking metrics in gene set enrichment analysis: do they matter? BMC Bioinformatics. 2017;18(1):256.
63. Johnson KE, Voight BF. Patterns of shared signatures of recent positive selection across human populations. Nat Ecol Evol. 2018;2(4):713–20.
64. HiChIP on pooled LCLs from 1000 Genomes Project. PRJNA898623. BioProject. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA898623. (2022).
65. Randolph HE, Fiege JK, Thielen BK, Mickelson CK, Shiratori M, Barroso-Batista J, et al. Single-cell RNA-sequencing reveals pervasive but highly cell type-specific genetic ancestry effects on the response to viral infection. GSE162632. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162632. 2021.
66. Pettie K.kadepettie, popABC: source code for Pettie, et al 2023 Zenodo 10.5281/zenodo.10396417

Publisher's Note