Resolving the chromatin impact of mosaic variants with targeted Fiber-seq

Stephanie C. Bohaczuk, Zachary J. Amador, Chang Li, et al.

Genome Res. 2024 34: 2269-2278 originally published online December 9, 2024
Access the most recent version at doi:10.1101/gr.279747.124

Method

Resolving the chromatin impact of mosaic variants with targeted Fiber-seq

Stephanie C. Bohaczuk,[1] Zachary J. Amador,[2] Chang Li,[1] Benjamin J. Mallory,[2] Elliott G. Swanson,[2] Jane Ranchalis,[1] Mitchell R. Vollger,[1] Katherine M. Munson,[2] Tom Walsh,[1] Morgan O. Hamm,[2] Yizi Mao,[1] Andre Lieber,[1] and Andrew B. Stergachis[1,2,3]

[1]*Division of Medical Genetics, University of Washington School of Medicine, Seattle, Washington 98195, USA;* [2]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [3]*Brotman Baty Institute for Precision Medicine, Seattle, Washington 98195, USA*

Accurately quantifying the functional consequences of noncoding mosaic variants requires the pairing of DNA sequences with both accessible and closed chromatin architectures along individual DNA molecules—a pairing that cannot be achieved using traditional fragmentation-based chromatin assays. We demonstrate that targeted single-molecule chromatin fiber sequencing (Fiber-seq) achieves this, permitting single-molecule, long-read genomic, and epigenomic profiling across targeted >100 kb loci with ~10-fold enrichment over untargeted sequencing. Targeted Fiber-seq reveals that pathogenic expansions of the *DMPK* CTG repeat that underlie Myotonic Dystrophy 1 are characterized by somatic instability and disruption of multiple nearby regulatory elements, both of which are repeat length-dependent. Furthermore, we reveal that therapeutic adenine base editing of the segmentally duplicated γ-globin (*HBG1*/*HBG2*) promoters in primary human hematopoietic cells induced toward an erythroblast lineage increases the accessibility of the *HBG1* promoter as well as neighboring regulatory elements. Overall, we find that these non–protein coding mosaic variants can have complex impacts on chromatin architectures, including extending beyond the regulatory element harboring the variant.

[Supplemental material is available for this article.]

Mosaic variants play a central role in the pathogenesis of Mendelian conditions, cancer, autoinflammatory diseases, and aging via altering the amino acid sequence of protein-coding genes or the gene regulatory elements critical for the appropriate expression of these genes (Bamford et al. 2004; Holzelova et al. 2004; Poduri et al. 2013; Luks et al. 2015; Alriyami and Polychronakos 2021). Although robust tools exist for detecting mosaic variants (Kim et al. 2018; Benjamin et al. 2019; Krishnamachari et al. 2022), quantifying the functional impact of noncoding mosaic variants on overlying chromatin architectures has proven to be more challenging. For example, germline variants that alter chromatin architectures are often detected by identifying ATAC-seq or ChIP-seq peaks with allelically imbalanced read counts. However, by definition, mosaic variants have allelically imbalanced read counts, confounding the ability to accurately disentangle the relative contribution of the variant on chromatin accessibility, especially since the variant allele fraction of a mosaic variant can shift depending on which section of a tissue is profiled. Furthermore, these short-read methods are ill-suited for identifying neighboring regulatory elements that may be altered by a variant as the reads are often <150 bp in length, resulting in a myopic view of how mosaic variants impact gene regulatory architectures.

Long-read sequencing allows for the precise identification of single-nucleotide and structural variants across individual multikilobase sequencing reads. Moreover, recent advances in single-molecule chromatin fiber sequencing (Fiber-seq) enable the co-identification of genetic variants and chromatin architectures along individual chromatin fibers (Abdulhay et al. 2020; Lee et al. 2020; Shipony et al. 2020; Stergachis et al. 2020; Altemose et al. 2022; Cheetham et al. 2022). For example, Fiber-seq uses nonspecific N^6-adenine methyltransferases (m6A-MTases) to selectively mark regions of chromatin accessibility and protein occupancy along individual DNA molecules via m6A-modified bases (Stergachis et al. 2020). m6A-modified bases along with endogenous CpG methylation are then detected using PCR-free long-read sequencing (Clark et al. 2012; Marks et al. 2012; Murray et al. 2012; Loman et al. 2015; Töpfer and Wenger 2023), enabling the single-molecule detection of genetic and chromatin information with the ability to evaluate these features within challenging genomic regions that are known to play a pivotal role in many human diseases (Cooper et al. 2011). We hypothesized that the single-molecule long-read nature of Fiber-seq would be well-suited for investigating the chromatin impact of mosaic variants and sought to create a targeted version of Fiber-seq that would combine the single-molecule and nucleotide precision aspects of Fiber-seq with targeted high-molecular-weight DNA enrichment (i.e., targeted Fiber-seq) (Fig. 1A). With a known locus of interest, high-molecular-weight DNA enrichment can drastically reduce sequencing costs compared to a whole-genome approach. This is especially useful to analyze mosaic variants, where deep sequencing may be required for sufficient coverage of a variant. It also enables the

Corresponding author: absterga@uw.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279747.124.

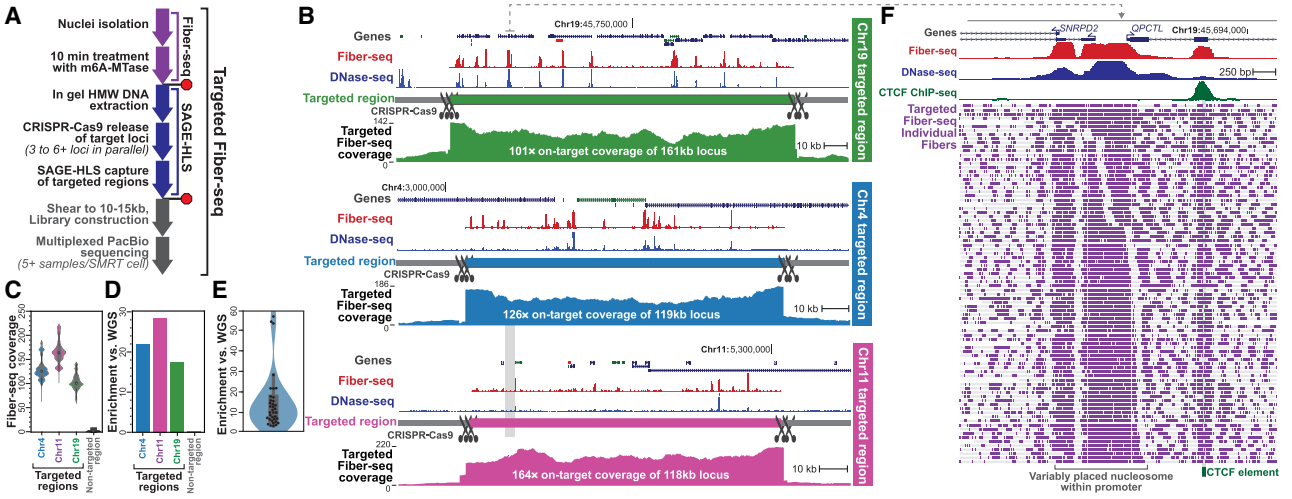

Figure 1. Targeted Fiber-seq methodology. (*A*) Schematic of targeted Fiber-seq protocol. (*B*) Fiber-seq percent actuation (red), ENCODE DNase-seq (blue), and per-base targeted Fiber-seq coverage are shown across targeted loci. (*C*) Violin plot of per-base coverage over the targeted regions, compared to a nontargeted region (Chr14:20283833–20402650, hg38). (*D*) Relative enrichment of targeted loci relative to WGS (see Methods). (*E*) Relative enrichment across all samples and targets. (*F*) Zoom-in of an illustrative locus showing m6A events (purple ticks) across individual fibers.

inclusion of more individuals and/or tissue types to more broadly capture the diversity of chromatin states contributing to a phenotype of interest.

Results

To assess the chromatin impact of variants across extended genomic loci, we designed targeted Fiber-seq to simultaneously capture multiple 100–250 kb genomic loci using a CRISPR–Cas9 targeting approach. Specifically, four wells of an HLS-SAGE cassette are each loaded with 0.5–1.5 million nuclei treated with m6A-MTase. These samples are then subjected to in-gel DNA extraction and CRISPR–Cas9 release of the targeted loci. The released DNA fragments are then separated and eluted from the gel using pulse-field electrophoresis, sheared to ∼13 kb, barcoded, multiplexed, and then sequenced using single-molecule long-read sequencing (Fig. 1A).

As a proof of concept, we first applied targeted Fiber-seq to cultured lymphoblastoid cells (GM04820), which represented 58% of the total mass loaded onto a Pacific Biosciences (PacBio) Sequel II SMRT cell. A median of 101–164× coverage was achieved across all three targeted loci (Fig. 1B,C), corresponding to ∼20-fold enrichment compared to whole-genome approaches (Fig. 1D; Methods). A median of 10-fold enrichment was obtained across all samples (Fig. 1E).

Comprehensive, deep-coverage epigenetic profiles revealed marked heterogeneity within the chromatin architecture of individual chromatin fibers. For example, although the *SNRPD2/QPCTL* bidirectional promoter was marked by accessible chromatin on each fiber, the boundaries of this accessible promoter element, as well as the positioning of an internal nucleosome varied across the individual reads (Fig. 1F), indicating that the precise pattern of accessibility within a promoter can vary quite substantially from fiber to fiber.

Given the potential for targeted Fiber-seq to synchronously resolve alterations in both the genome and chromatin epigenome at single-molecule resolution, we next applied targeted Fiber-seq to evaluate the chromatin impact of the unstable CTG expansion in the 3′ UTR of the gene *DMPK*, which causes Myotonic Dystrophy 1 (DM1) (Brook et al. 1992; Buxton et al. 1992; Fu et al. 1992; Harley et al. 1992; Mahadevan et al. 1992). DM1 is a dominantly inherited trinucleotide repeat disorder characterized by muscle weakness, myotonia, early cataracts, and other symptoms, which become progressively more severe with age. *DMPK* CTG repeat expansions, which are somatically unstable (Anvret et al. 1993; Ashizawa et al. 1993; Wong et al. 1995), disrupt RNA pathways (Hamshere et al. 1997; Klesert et al. 1997; Miller et al. 2000; Frisch et al. 2001; Ranum and Day 2004; Ravel-Chapuis et al. 2012; Meola and Cardani 2015), and locally alter chromatin compaction and CpG methylation (Otten and Tapscott 1995; Filippova et al. 2001; López Castel et al. 2011). However, the extent to which repeat expansions impact chromatin compaction and CpG methylation is not well resolved owing to the inherent limitations of traditional methods. Specifically, the precise boundaries of chromatin accessibility disruptions, the prevalence, and the co-occurrence of these changes on individual chromatin fibers remain unknown.

To address this, we applied targeted Fiber-seq to enrich for the *DMPK* locus in fibroblasts from four symptomatic individuals in a family with DM1 (Fig. 2A). Using reads overlapping the *DMPK* CTG repeats, we quantified repeat instability at single-molecule resolution (Fig. 2B; Supplemental Fig. S1). We found that, whereas the affected individual in the first generation of this pedigree had a median pathogenic *DMPK* CTG repeat expansion of 59.5 CTGs and a range of 20 CTG units, *DPMK* CTG repeats in individuals in the second and third generations were consistently >1000 CTGs, with individual chromatin fibers differing by almost 2000 CTG repeats within an individual donor (Supplemental Fig. S1).

We next leveraged the underlying highly accurate sequencing information to haplotype-phase reads from each donor within the 160 kb targeted region, thereby identifying whether it arose from chromatin fibers containing the normal or pathogenic repeat expansion (Fig. 2C; Supplemental Fig. S2; Methods). Pathogenic *DMPK* CTG repeat expansions have previously been associated with increased CpG methylation upstream of the CTG repeat, which has been proposed to directly inhibit CTCF occupancy at an upstream element (Filippova et al. 2001). Additionally, decreased DNase I hypersensitivity of the downstream *SIX5* promoter

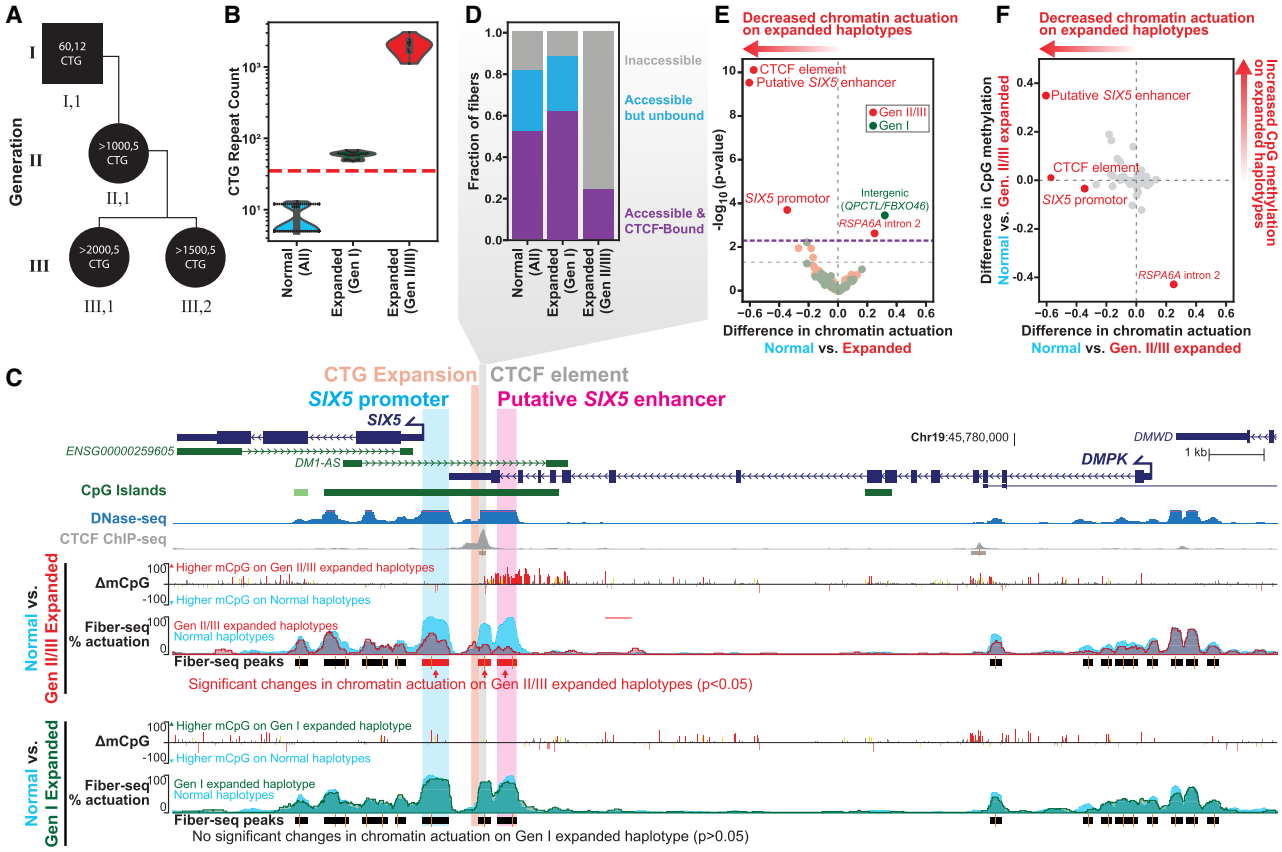

Figure 2. Haplotype-resolved chromatin accessibility and CpG methylation at the *DMPK* 3′-UTR CTG expansion in DM1 fibroblasts. (*A*) Family pedigree representing fibroblast donors. Labels *inside* each shape represent the CTG copy number within the pathogenic and normal allele, respectively. (*B*) CTG count across sequencing reads that fully span the CTG repeats, grouped by haplotypes. The red dashed line at 35 represents the threshold above which CTG expansions are unstable. (*C*) Browser tracks comparing the chromatin architecture of the normal haplotypes to the expanded haplotypes from generations II/III and generation I. The difference in CpG methylation is compared (yellow—$P<0.01$, orange—$P<0.001$, red—$P<0.0001$, Fisher's exact test), as well as percent actuation of the normal (light blue), generations II/III expanded (red), and generation I expanded (green) haplotypes. Fiber-seq peaks from the normal haplotypes are shown *below* and are colored to represent a statistically significant decrease in chromatin accessibility on the generation II/III haplotypes (red) compared to normal. There were no significant changes between the normal and generation I (green) expanded haplotype. ENCODE DNase-seq and CTCF ChIP-seq are *above* in blue and gray, respectively. (*D*) CTCF footprinting at the CTG-adjacent CTCF-binding site. Footprints were classified as accessible if they were fully overlapped by a methylation-sensitive patch, and accessible footprints were classified as CTCF bound if they did not contain any m6A ($P=0.010$, Fisher's exact test comparing CTCF bound to unbound [inaccessible plus accessible but unbound]). (*E*) Volcano plot of chromatin actuation difference at each accessible peak from the normal haplotype peak set within the targeted locus, compared between normal and expanded fibers from generation I (green) and generations II/III (red). The two peaks with increased accessibility in expanded fibers (upper *right* quadrant) are likely explained by SNPs local to each region (Supplemental Note). *P*-values were calculated by Fisher's exact test. The gray dashed horizontal line indicates the nominal significance threshold ($P<0.05$), and the purple dashed line indicates the Benjamini–Hochberg FDR-corrected significance threshold. (*F*) Plot of CpG methylation difference versus chromatin actuation difference at the peaks described in *E*. Significant and nonsignificant points from *E* are shown in red and gray, respectively.

has also been observed, along with reduced *SIX5* mRNA levels (Klesert et al. 1997). The CpG methylation data obtained via targeted Fiber-seq demonstrates that pathogenic *DMPK* CTG repeat expansions from individuals in the second and third generation are associated with a focal 1 kb hyper-CpG methylated domain upstream of the CTG repeat that extends to the end of the overlapping CpG island. Notably, this hyper-CpG methylated domain does not include the upstream CTCF element, but rather initiates immediately after it (Fig. 2C). In contrast, CpG methylation downstream from the *DMPK* CTG repeat is unchanged—indicative of a focal, directional impact of the pathogenic repeat expansion on CpG methylation.

Along the nonpathogenic haplotype, the *DMPK* CTG repeat is bookended downstream by the accessible *SIX5* promoter, and upstream by both a CTCF bound accessible element and an adjacent accessible element positive for H3K4me1/H3K27ac in ENCODE fibroblast data sets (Fig. 2C; Supplemental Fig. S2). Of note, this upstream accessible element maps within a region that was previously reported to harbor a putative *SIX5* enhancer, but the exact position of the enhancer within this region had not been determined (Yanovsky-Dagan et al. 2015). Leveraging the paired Fiber-seq chromatin architectures along each fiber, we observed that the putative *SIX5* enhancer is encompassed by the 1 kb hyper-CpG-methylated domain, and chromatin accessibility is largely ablated across the repeat expansion haplotypes from the second and third generation (Fig. 2C,E,F; Supplemental Note) (P-value $= 7.3 \times 10^{-9}$, Fisher's exact test with Benjamini–Hochberg false discovery rate [FDR] correction). This result demonstrates a direct link between *DMPK* repeat expansion, focal CpG hypermethylation, and reduced chromatin accessibility at this upstream element.

Although the hyper-CpG methylated domain did not overlap the CTCF-binding motif, there was a substantial reduction in chromatin accessibility of the underlying regulatory element in expanded fibers from the second and third generations (*P*-value $= 3.8 \times 10^{-9}$, Fisher's exact test with Benjamini–Hochberg FDR correction). Utilizing the near single-base pair resolution of targeted Fiber-seq, we classified fibers at the CTCF motif as accessible and CTCF bound, accessible and unbound, or inaccessible (Fig. 2D). Compared to fibers from normal haplotypes which were 52% bound by CTCF, 24% expanded fibers from the second and third generation were CTCF bound (*P*-value = 0.010, Fisher's exact test). This reduction in binding was correlated with an increase in nucleosome occupancy overlapping the CTCF-binding motif. Together, these results indicate that hyper-CpG methylation directly within the upstream CTCF-binding element is not required to inhibit CTCF binding on CTG-expanded fibers, as has previously been suggested by in vitro studies which demonstrate that CpG methylation of a DNA oligonucleotide containing the upstream CTCF motif can inhibit CTCF binding (Filippova et al. 2001).

We next sought to detect any long-range chromatin impacts associated with these altered elements across the 160 kb targeted domain. We observed that of the 96 annotated TSSs within this targeted region, only actuation of the *SIX5* promoter was significantly reduced along expanded fibers from the second and third generations (Supplemental Figs. S3, S4) (*P*-value = 0.028, Fisher's exact test with Benjamini–Hochberg FDR correction). The canonical *DPMK* TSS had unchanged chromatin accessibility, consistent with findings that suggest posttranscriptional mechanisms drive reduced *DMPK* mRNA levels in DM1 (Hamshere et al. 1997; Frisch et al. 2001). Notably, CpG methylation was unchanged at the *SIX5* promoter, consistent with its altered promoter accessibility being predominantly driven by a loss of an enhancer, as opposed to CpG methylation mediated silencing via the repeat itself (Fig. 2F). Together, these findings indicate that pathogenic *DMPK* repeat expansions disrupt chromatin via a focal alteration in upstream CpG methylation and chromatin accessibility and implicate this upstream enhancer-like element as an enhancer for *SIX5*.

In contrast, we observed that the aforementioned differences in CpG methylation, chromatin accessibility, and CTCF occupancy were not observed along chromatin fibers from the pathogenic allele in the first generation, which contains only ~60 CTG repeats. This suggests that the observed chromatin phenotype of DM1 requires a critical number of CTG repeats, which may be reached either by age-related somatic expansion in adult-onset DM1 or congenitally present in individuals born with large germline CTG repeats.

Next, we sought to apply targeted Fiber-seq to resolve how therapeutic base editing of the fetal γ-globin promoters (encoded by *HBG1/HBG2*) in human $CD34^+$-derived erythroid cells affects chromatin structure. Reactivation of fetal γ-globin expression is a promising therapy for β-hemoglobinopathies, as γ-globin can disrupt the formation of HbS polymerization in sickle cell disease and supplement β-globin insufficiency in β-thalassemia (Eaton and Hofrichter 1987; Bollekens and Forget 1991; Akinsheye et al. 2011; Steinberg et al. 2014). The expression of *HBG1* and *HBG2* are repressed in adulthood by the transcriptional regulator BCL11A (Sankaran et al. 2008), and base editing therapies that ablate BCL11A occupancy of the *HBG1* and *HBG2* promoters are in clinical investigation (Beam Therapeutics Inc. 2022). However, our mechanistic understanding of how base editing the *HBG1* and *HBG2* promoters impacts fetal γ-globin expression is dually challenged as base editing is often incomplete, which results in a mixture of edited and unedited haplotypes within a population of cells, and the *HBG1* and *HBG2* genes are situated within a highly similar segmental duplication with 100% sequence identity over the proximal promoter, limiting short-read chromatin methods from uniquely profiling this region (Supplemental Fig. S5). To address this, we treated $CD34^+$ human hematopoietic stem cells (HSCs) isolated from two healthy donors with a targeted adenine base editor (ABE) to convert A > G within a binding site for BCL11A located −113 bp upstream of both the *HBG1* and *HBG2* TSSs. This same variant is known to cause hereditary persistence of fetal hemoglobin (HPFH) when present in the germline (Amato et al. 2014). We induced differentiation toward an erythroblast lineage and subjected these cells to targeted Fiber-seq, targeting a 118 kb region spanning from the β-globin locus control region (LCR) to the 3′HS1 (Fig. 3A). Notably, de novo genome assembly of the Fiber-seq reads from each individual revealed that one of the donors harbored a rare γ-globin gene triplication along one of their haplotypes (Supplemental Figs. S6, S7). Given the possible confounding features of this triplication event, we removed reads arising from this triplicated haplotype from our downstream analyses.

As a first step, we evaluated single-molecule editing efficiency across the β-globin locus, as well as for altered chromatin architectures induced along the base edited chromatin fibers (Fig. 3A). We observed that in addition to the targeted −113 A > G edit, edited fibers almost exclusively contained additional A > G edits in the immediate vicinity (Fig. 3B). Although only 34% and 32% of fibers demonstrated base editing at the *HBG1* and *HBG2* promoters, respectively (see Methods), we found that if one of the promoters was edited, the other promoter on that same fiber was also significantly likely to be edited ($P < 0.00001$, Fisher's exact test), suggesting that editing efficiency is largely determined by transduction and expression of the ABE (Fig. 3C). As such, we grouped fibers edited at *HBG1* and/or *HBG2* fibers for further analyses.

Importantly, the long-reads obtained by Fiber-seq enable the unique mapping of CpG methylation, chromatin architectures, and ABE-induced base changes to each of the segmentally duplicated *HBG1* and *HBG2* genes (Fig. 3D; Supplemental Fig. S8). Comparison of the CpG methylation patterns between the edited and unedited reads demonstrated that base editing of the *HBG1* and *HBG2* promoters resulted in a significant reduction in CpG methylation over the entire span of both the *HBG1* and *HBG2* genes.

In addition to this broad change in CpG methylation, we also saw focal changes in chromatin accessibility. Specifically, base editing of the *HBG1* and *HBG2* promoters resulted in a >100% increase in chromatin accessibility of the *HBG1* promoter (Fisher's exact test *P*-value = 0.00030, *P*-value = 0.0033 after Benjamini–Hochberg FDR correction), as well as another element located 2.5 kb upstream of the *HBG1* promoter (Fisher's exact test *P*-value = 0.024, *P*-value = 0.13 after Benjamini–Hochberg FDR correction) (Fig. 3D; Supplemental Fig. S9). In contrast, chromatin accessibility at the four other accessible elements within the *HBG1/HBG2* segmental duplication exhibited nonsignificant changes along the edited fibers, including the *HBG2* promoter. The relatively modest 36% increase in *HBG2* promoter accessibility reflects higher *HBG2* promoter accessibility in unedited fibers. Accessibility of the edited *HBG1* and *HBG2* promoters was comparable, consistent with previous reports that noted equivalent expression of the *HBG1* and *HBG2* gene products, A gamma and

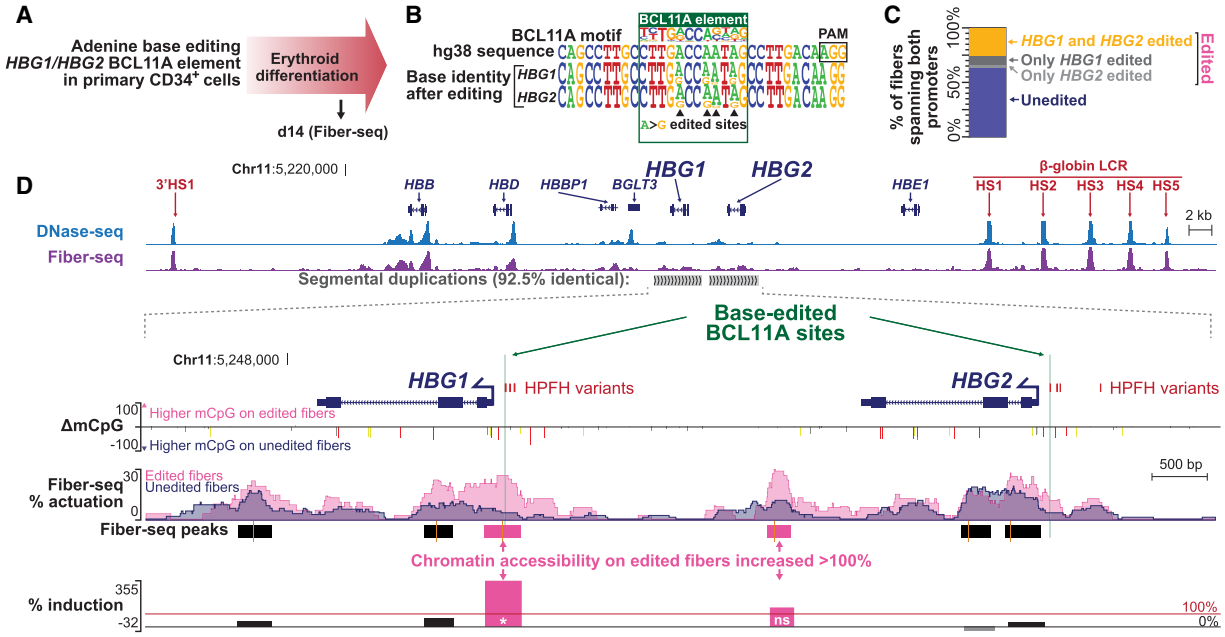

Figure 3. Therapeutic base editing of a BCL11A element within the *HBG1*/*HBG2* promoters. (*A*) Schematic of the experimental paradigm. (*B*) Sequence logo showing base editing within the BCL11A binding site in the *HBG1* and *HBG2* promoters. Letter height corresponds to the relative base frequency. (*C*) Percent of fibers with edited BCL11A sites in both *HBG1* and *HBG2* (yellow), *HBG1* only (dark gray), *HBG2* only (light gray), and neither (blue). The top three categories are grouped as "edited" reads in *D*. (*D*) UCSC browser tracks of ABE-edited CD34$^+$-derived erythroid cells. (*Top*) Comparison of ENCODE DNase-seq of treated multipotent progenitor cells (blue) and chromatin accessibility (FIRE) of CD34$^+$ derived erythroids (purple) across all fibers mapping within the targeted region. (*Bottom*) Zoom-in with comparison of edited and unedited fibers across *HBG1* and *HBG2*. The difference in CpG methylation is compared (unedited minus edited, yellow—$P < 0.01$, orange—$P < 0.001$, red—$P < 0.0001$, Fisher's exact test), as well as percent actuation of unedited fibers (blue), and *HBG1* and/or *HBG2* edited fibers (pink). Peak calls for edited reads and percent induction at each peak ([edited percent actuated – unedited percent actuated]/unedited percent actuated) are displayed at the *bottom*. Pink peaks and bars represent >100% induction. (*) indicates statistical significance (Fisher's exact test with Benjamini–Hochberg corrected FDR < 5%) (Supplemental Fig. S9).

G gamma, in a HUDEP-2 erythroid model ABE-edited using the same gRNA as in our study (Ravi et al. 2022). Notably, the four nonpromoter regulatory elements that we identified within the *HBG1/HBG2* segmental duplication were enriched for histone modifications associated with active regulatory elements in hematopoietic cells, and chromatin accessibility of these elements was coupled with the accessibility of the *HBG1/HBG2* promoters along single molecules (Supplemental Fig. S8), suggesting a putative role as *HBG1/HBG2* enhancer elements. Together, these findings demonstrate that base editing of the *HBG1* and *HBG2* −113 A > G sites preferentially occurs along the same chromatin fiber and induces chromatin accessibility at *HBG1*, likely in cooperation with adjacent enhancer-like elements.

Discussion

Overall, we demonstrate that targeted Fiber-seq enables the production of targeted long-read sequencing chromatin maps to resolve the heterogeneity of genetic and chromatin architectures with single-molecule precision. Importantly, as targeted Fiber-seq enables the synchronous measurement of DNA sequence, CpG methylation, and chromatin accessibility along the same +10 kb molecule of DNA, we can directly disentangle the functional impact of heterogeneously present genetic variants on neighboring gene regulatory programs. Application of this to the *DMPK* and *HBG1/HBG2* loci demonstrated that precise genetic changes can result in complex alterations in chromatin accessibility that extend beyond the genetically modified element or the region with disrupted CpG methylation—alterations that would be largely hidden from short-read allelic imbalance or long-read CpG methylation methods.

With combined and phased chromatin accessibility and CpG methylation on the same chromatin fibers, our results provide new insight into the chromatin phenotype of DM1. First, our results substantiate the seemingly contradictory findings that chromatin accessibility of the *SIX5* promoter is decreased (Otten and Tapscott 1995; Klesert et al. 1997) while CpG methylation is unchanged (López Castel et al. 2011). More broadly, this result highlights that although CpG methylation and chromatin accessibility are correlated (The ENCODE Project Consortium 2012), they are not interchangeable. Second, we show for the first time that reduced CTCF binding to the element located directly upstream of the CTG repeat can occur without a corresponding increase in overlapping CpG methylation. In vitro binding studies previously demonstrated that CpG methylation of the CTCF-binding motif inhibited CTCF binding (Filippova et al. 2001), and so it was proposed that CpG methylation is the mechanism by which CTCF binding is reduced in vivo. In support of our findings, more recent studies using Oxford Nanopore Technology (ONT) sequencing also show variable CpG methylation of the CTCF-binding element in individuals with DM1 (Rasmussen et al. 2022). However, these ONT studies are incapable of comeasuring CTCF occupancy. Comeasurements of CpG methylation, chromatin accessibility, and single-molecule CTCF footprinting enabled us to eliminate any confounding factors and directly disentangle this result.

In addition, we demonstrate the utility of targeted Fiber-seq for resolving both genetic and gene regulatory alterations within complex genomic regions such as segmental duplications.

Specifically, we were able to precisely disentangle the genetic and chromatin architectures of the *HBG1* and *HBG2* genes, demonstrating that the γ-globin −113 A > G variant that is known to cause HPFH results in the focal increase in chromatin accessibility at the γ-globin gene promoters, as well as at adjacent regulatory elements. Furthermore, we were able to perform de novo genome assembly on our data to show that one of the donors harbored a rare γ-globin gene triplication along one of their haplotypes. This later observation enabled us to selectively remove reads arising from this triplicated haplotype from our analyses above, reducing a possible confounding feature that would have been largely hidden from short-read-based chromatin methods.

Our results provide mechanistic insights into the pathogenesis of both DM1 and HPFH, precisely delineating enhancer elements that potentially play a critical role in both of these conditions. Specifically, we identify that pathogenic *DMPK* repeat expansions disrupt an upstream accessible chromatin element with enhancer-like activity and implicate this upstream enhancer-like element as an enhancer for *SIX5*, providing a mechanistic basis for altered *SIX5* expression, which is one of the hallmark features of DM1. We also found that the size of the CTG expansion was correlated with both somatic instability and disruptions in chromatin architecture, raising the possibility that a common mechanism might contribute to both processes. Notably, these findings were performed in patient-derived fibroblasts, and further studies using primary tissues from patients will be helpful to further evaluate this mechanism.

Together, these findings emphasize the benefit of targeted single-molecule chromatin accessibility assays, such as targeted Fiber-seq, to fully capture the genetic and functional impact of germline and mosaic variants. Here, we demonstrated that targeted Fiber-seq can be successfully applied to both cell lines and primary human cells. As whole-genome Fiber-seq has also been demonstrated on nuclei isolated from biopsied human tissue (Grasberger et al. 2024), we expect that the targeted Fiber-seq methodology would translate to primary human tissue samples as well. We anticipate that further advances in target capture enrichment will build upon the benefits delineated in this manuscript. Furthermore, by unraveling the chromatin basis of these two disease-associated variants, this study highlights additional potential therapeutic epigenetic targets for both DM1 and β-hemoglobinopathies.

Methods

Cell line culture

The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM04820 (lymphoblastoid), GM06076 (fibroblast), GM04601 (fibroblast), GM04602 (fibroblast), and GM04608 (fibroblast). Lymphoblastoid cells were maintained in suspension in Iscove's Modified Dulbecco's Medium (IMDM) media supplemented with 10% FBS (HyClone SH30396.03IH25-40) and antibiotic (100 IU/mL penicillin, 100 μg/mL streptomycin, Gibco 15140122) at 37°C and 5% CO_2 in T-75 flasks. Cells were split 1:10 every 3–4 days. Fibroblasts were maintained in DMEM media with the same supplementation and incubation as above. Cells were split 1:4 every 5–7 days with 0.25% trypsin-EDTA (Gibco 25200056).

CD34$^+$ cell culture and transduction

CD34$^+$ cells from G-CSF-mobilized healthy adult donors were provided by the Fred Hutch Cell Processing Facility. The cells were recovered from frozen stocks and incubated overnight in StemSpan H3000 medium (STEMCELL Technologies) supplemented with penicillin/streptomycin, Flt3 ligand (Flt3L, 25 ng/mL), interleukin 3 (10 ng/mL), thrombopoietin (TPO, 2 ng/mL), and stem cell factor (SCF, 25 ng/mL). Cytokines and growth factors were from Peprotech. CD34$^+$ cells were transduced in low- attachment 6-well plates with an all-in-one base editing vector HDAd-ABE8e-sgHBG2 targeting the BCL11A binding sites in the *HBG1/2* promoters for fetal hemoglobin reactivation (Li et al. 2022).

In vitro erythroid differentiation of CD34$^+$ cells with O^{6}BG/BCNU selection

Differentiation of human CD34$^+$ cells into erythroid cells was done based on the protocol developed by Douay and Giarratana (2009). In brief, in step 1, cells at a density of 10^4 cells/mL were incubated for 7 days in IMDM supplemented with 5% human plasma, 2 IU/mL heparin, 10 g/mL insulin, 330 g/mL transferrin, 1 M hydrocortisone, 100 ng/mL SCF, 5 ng/mL IL-3, 3 U/mL erythropoietin (Epo), glutamine, and penicillin/streptomycin. In step 2, cells at a density of 1×10^5 cells/mL were incubated for 3 days in IMDM supplemented with 5% human plasma, 2 IU/mL heparin, 10 g/mL insulin, 330 g/mL transferrin, 100 ng/mL SCF, 3 U/mL Epo, glutamine, and Pen/Strep. In step 3, cells at a density of 1×10^6 cells/mL cells were incubated for 4 days in IMDM supplemented with 5% human plasma, 2 IU/mL heparin, 10 g/mL insulin, 330 g/mL transferrin, 3 U/mL Epo, glutamine, and Pen/Strep. For the enrichment of transduced cells, 48 h posttransduction, CD34$^+$ cells were incubated with 50 M O^6-Benzylguanine (O^{6}BG) for 1 h. Without washing, 35 M Carmustine (BCNU) was added for 2.5 more hours incubation, after which cells were washed and resuspended in fresh medium. Both drugs were purchased from Millipore/Sigma and freshly prepared. O^{6}BG/BCNU selection was used to enrich for edited cells in samples PS00208.erythroid1, PS00209.erythroid2, and PS00316.erythroid1.

Fiber-seq

In-house Hia5 preparation and Fiber-seq were performed as described (Stergachis et al. 2020). Briefly, $2–4 \times 10^6$ cells were washed with PBS, lysed in lysis buffer (15 mM Tris-Cl, pH 8, 15 mM NaCl, 60 mM KCl, 1 mM EDTA pH 8, 0.5 mM EGTA pH 8, 0.5 mM spermadine, 0.025% IGEPAL) and treated with Hia5 at 100 U per million cells (25°C, 10 min). The reaction was split into four equal aliquots, and SDS was added to a final concentration of 2%. The volume of each aliquot was adjusted to 70 μL with suspension buffer M2 (Sage Science). Samples used in this study are described in Supplemental Table S1.

crRNA design

Three crRNAs per side per target were designed within a ~3 kb span with attention to avoid annotated repeat elements or common variants. The following target windows represent the region flanked by the innermost crRNAs: Chr4:3006837–3125654 (Chr 4 target), Chr11:5186600–5304585 (Chr 11 target), and Chr19:45664901–45825746 (Chr 19 target). crRNAs were synthesized by IDT. crRNA sequences are detailed in Supplemental Table S2.

HLS-CATCH

HLS-CATCH was performed according to the manufacturer's protocol (Sage Science), using the Non-Core Workflow "CATCH 100–300 kb extr3h enhINJ Sep3h," with modifications as described: For the extraction step, post fiber-seq nuclei aliquots ($\sim 6 \times 10^5$–1×10^6 cells/lane) were added to HLS-Sage sample wells,

and 210 µL of HLS lysis buffer A was added to the reagent wells. Cassettes were sealed and run for 3 h, 55 V with wave index 1–1. Meanwhile, crRNAs (3 per target per side) were diluted to 100 µM in nuclease-free duplex buffer (IDT 11-01-03-01) and annealed and prepared as specified, using NEB Cas9 (M0386M) or NEB EnGen Spy Cas9 HF1 (M0667M). The Cas9 reaction mix was injected for 2 min, 80 V with wave index 2–1, and incubated for 30 min at RT. The reagent well was replaced with lysis buffer A and ran for 3 h, 55 V with a wave index 3–2. The sample was eluted for 1.5 h, 50 V with wave index 3–1.

Library preparation and PacBio sequencing

Post-CATCH samples were purified with 0.5× volume of Ampure PB beads and eluted in 52 µL of elution buffer (PacBio 101-633-500). One microliter was used to quantify sample concentration using a Qubit dsDNA HS Assay according to the manufacturer's protocol. Samples were sheared to 10–15 kb with two passes (one inversion) through a Covaris gTUBE at 3200 rpm for 2–4 min per pass in an Eppendorf 5424R centrifuge. Approximately 50 µL per sample was recovered and barcoded using the SMRTbell Express Template Prep Kit 2.0 (PacBio) with 5 µL of barcoded adapter from the Barcoded overhang adapter kit 8A/B for Sequel II sequenced samples or the SMRTbell prep kit 3.0 with the SMRTbell adapter index plate 96A, according to the protocol for each kit with adjustment for larger sample volume. Samples were sequenced by UW PacBio Sequencing Services on a Sequel II SMRT Cell 8 M with a 30 h movie or on a Revio SMRT Cell 25 M with a 24 h movie.

Fiber-seq processing

Circular consensus sequence reads were generated from raw PacBio subread files using PacBio CCS (https://ccs.how/) with chunking and average kinetics information included. Chunks were combined with pbmerge (https://pbbam.readthedocs.io/en/latest/tools/pbmerge.html). CCS reads were demultiplexed with lima (https://lima.how/). CpG methylation was called on demultiplexed BAM files directly from polymerase kinetics using Primrose/Jasmine (Töpfer and Wenger 2023), keeping kinetics for subsequent m6A calling. m6A and nucleosome calls were generated with fibertools using a 55 bp nucleosome threshold, 100 bp combined nucleosome length, and 25 bp distance from the end (Jha et al. 2024). A small percentage of reads with the proportion of methylated adenine to all adenines below 0.02 or above 0.4 were excluded from further analysis. Scripts to filter m6A reads by m6A proportion are available here (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/m6a_percent_filtering) and in Supplemental Code.

Alignment

Reads were aligned to hg38 with pbmm2, which revealed that one of the two CD34$^+$ cell donors had a heterozygous duplication of *HBG2* (Supplemental Figs. S6, S7). To resolve this, a custom assembly of the β-globin locus was created with a separate sample from the same donor (PS00148) using hifiasm (Cheng et al. 2021) with the -f0 flag to create an assembly from reads aligning to Chr11:5180424–5305559 in hg38. hifiasm yielded two contigs: a 133 kb contig containing one copy each of *HBG1* and *HBG2* (2γ haplotype) and a 126 kb contig containing one *HBG1* copy and two *HBG2* copies (3γ haplotype). The donor-specific assembly is available here (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/ABE_hifiasm_assembly) and in Supplemental Data, and NucFreq plots (Vollger et al. 2019) of alignment to hg38 and the donor-specific assembly are shown in Supplemental Figure S7. Reads from this donor aligning to Chr11:5186600–5304585 in hg38 were remapped to the hifiasm custom genome and filtered for primary alignments only with MAPQ score >10. Reads mapping to the 2γ haplotype were filtered from the hg38 alignment and merged with all reads from the second donor, who was homozygous for the 2γ haplotype. Coverage of the 3γ haplotype was insufficient for follow-up analyses.

Phasing

DM1 reads were phased separately for each sample. First, DeepVariant (Poplin et al. 2018) was used to identify variants and produce a VCF file. Reads were phased with HiPhase (Holt et al. 2024) using the "*k*-mer variant phasing" pipeline with default settings (https://github.com/mrvollger/k-mer-variant-phasing).

Fiber-seq chromatin accessibility and peak actuation comparisons

To assess chromatin percent actuation, we used the Fiber-seq Inferred Regulatory Element (FIRE) pipeline version 0.0.4 (https://github.com/fiberseq/fire) (Vollger et al. 2024), a method that uses a semisupervised machine learning algorithm to predict the likelihood that a methylation-sensitive patch is a regulatory element on individual chromatin fibers. The configuration files used to run FIRE are available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/FIRE_config) and in Supplemental Code. The Fiber-seq chromatin percent actuation tracks in Figures 1, 2, and 3 represent the percent of fibers with FIRE elements at each base pair. Peaks are called with min_frac_accessible: 0.15. Reads were grouped as follows for FIRE analysis: PS00118 (Fig. 1); normal haplotypes from all DM1 samples (PS00150, PS00151, PS00152, PS00153, PS00442, PS00443, and PS00444), expanded haplotypes from generation II/III DM1 donors (PS00150, PS00151, PS00152, PS00153, PS00443, and PS00444), expanded haplotype from generation I DM1 donor (PS00442) (Fig. 2); edited and unedited 2γ-haplotype from CD34$^+$ donor 1 and both haplotypes from CD34$^+$ donor 2 (PS00196, PS00208, PS00209, and PS00316) (Fig. 3D, top track in purple), edited reads from 2γ-haplotype from CD34$^+$ donor 1 and both haplotypes from CD34$^+$ donor 2 (PS00196, PS00208, PS00209, and PS00316), unedited reads from 2γ-haplotype from CD34$^+$ donor 1 and both haplotypes from CD34$^+$ donor 2 (PS00196, PS00208, PS00209, and PS00316) (Fig. 3).

Fraction actuation over each peak was computed as the number of reads with a FIRE score ≤0.1 anywhere within the ∼100–200 bp peak divided by the total number of reads mapped within the peak. For DM1 analyses, normal and expanded peaks were compared across peaks called for normal haplotypes from all DM1 samples (Fig. 3C). For comparison of edited and unedited CD34$^+$-derived erythroids, edited and unedited peaks were compared across all peaks called for the edited haplotype. Fisher's exact test was performed in Python (scipy.stats.fisher_exact) with Benjamini–Hochberg FDR correction (scipy.stats.false_discovery_control) using FDR < 5% for statistical significance. The FDR-corrected significance threshold was calculated as 0.05 × (rank of the smallest significant FDR-corrected *P*-value + 1)/total number of comparisons, corresponding to the *P*-value that would have been required for the smallest *P*-value point that did not reach FDR-corrected significance to be significant. In Figure 2E and Supplemental Figure S4, the FDR-corrected significance threshold is plotted for generation II/III samples, but generation I points above and below this line are accurately classified as significant or nonsignificant, respectively, following FDR correction. Percent induction due to editing (Fig. 3C) is calculated as (edited percent actuated – unedited percent actuated)/unedited percent actuated. Scripts to reproduce analyses and figures are available on GitHub

(https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/compare_perc_actuation) and in Supplemental Code.

CpG methylation

To compare CpG methylation, the haplotype tag (1 or 2) was used to label reads within each BAM, and then BAM files to compare were merged. "aligned_bam_to_cpg_scores" (Pacific Biosciences 2023) was run with default settings using the "pileup_calling_model.v1.tflite" model. For CpG differences tracks (Figs. 2C, 3D), MethBat (Holt and Saunders 2023) was used to calculate the difference in CpG methylation and an associated *P*-value at each CpG (i.e., 1 bp input regions). CpG sites that were present as heterozygous variants (>10% frequency) were excluded. For CpG methylation over peaks (Fig. 2F), peaks were used as input regions. Scripts to reproduce CpG tracks are available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/CpG_tracks) and in Supplemental Code.

Coverage and enrichment

Coverage was calculated at each base pair within the targeted loci. Enrichment was calculated as median sequencing depth/(mass fraction × expected WGS coverage), where mass fraction is sample mass/total mass loaded onto the multiplexed SMRT cell, and an expected WGS coverage of 10 and 30 was used for Sequel II and Revio, respectively. Scripts to reproduce analyses are available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/coverage) and in Supplemental Code.

DM1 repeat expansion length

The CTG/CAG repeat was counted on reads that were anchored both 5′ and 3′ of the repeat, i.e., those where either the primary alignment mapped both 5′ and 3′ of the repeat, or a primary alignment mapped either 5′ or 3′ and a supplementary alignment mapped to the missing side. The length of the expansion was calculated by finding the position of the first and last CAG within the repeat region, subtracting these positions to find the total length of bases, and then dividing by three for the length in CAG units. A custom script to reproduce this analysis is available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/DM1_CAG).

CTCF footprinting

CTCF footprinting was computed using a custom script, which assigns fibers as "accessible and CTCF bound" if the core of the CTCF-binding site (Yin et al. 2017) (Chr19:45770329–45770342 in hg38) is fully overlapped by a methylation-sensitive patch (MSP) but does not contain any m6A calls within, "accessible but unbound" if the core is fully overlapped by an MSP but does contain m6A calls within, or "inaccessible" if the core is overlapped by a nucleosome. Statistics represent Fisher's exact test comparing "accessible and CTCF bound" to unbound (i.e., "accessible and unbound" + "inaccessible"). A custom script to reproduce this analysis is available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/DM1_CTCF) and in Supplemental Code.

Identification of ABE edits

Reads mapping to the edited BCL11A site (Li et al. 2022) within the *HBG1* and/or *HBG2* promoters were categorized as "unedited" if they contained no indels or mismatches within a window of −4/+5 bp from the intended edit site located 113 bp upstream of the *HBG1/2* TSS. Reads were categorized as "edited" if they contained the expected A > G (T > C in the reference genome), with other edits within the window also tolerated. Reads not classified as "edited" or "unedited" were excluded from further analyses. Scripts to classify reads by editing status are available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/ABE_editing) and in Supplemental Code.

ENCODE tracks

The following ENCODE (The ENCODE Project Consortium 2012) tracks were included: DNase-seq of GM12878, ENCFF960FMM (Fig. 1B,F); DNase-seq of fibroblasts, ENCFF302JEV, CTCF ChIP-seq of fibroblasts, ENCFF080HIA (Fig. 2C); DNase-seq of hematopoietic multipotent progenitor cell treated with interleukin-3 for 8 days, kit ligand for 8 days, hydrocortisone succinate for 8 days, erythropoietin for 8 days, ENCFF800PMS (Fig. 3C).

Coaccessibility and codependency

Coaccessibility between each pair of peaks was quantified by evaluating fibers that (1) entirely overlapped each of the two peaks and (2) had at least 46 bp between the read ends and peak edges. Fibers were classified as accessible at a peak if they contained a FIRE element (FIRE score ≤0.05) that overlapped the peak by ≥1 bp. Overlaps were computed using BEDTools intersect (v2.31.0) (Quinlan and Hall 2010). Coaccessibility data were filtered to only include peak pairs that were overlapped by a minimum of 20 fibers. Codependency scores were calculated as the difference between the observed and expected coaccessibility given the null hypothesis that the actuation of each peak is independent of the other. The expected coaccessibility was calculated as the product of the actuation proportion at each of the two peaks. The resulting scores were multiplied by 4 to scale the range of possible values from −1 (only one peak is actuated per fiber) to +1 (actuation at both peaks for a given fiber). Scripts to reproduce this analysis are available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq/tree/main/ABE_codependency) and in Supplemental Code.

Software availability

The code to reproduce figures and analyses in this study is available on GitHub (https://github.com/StephanieBohaczuk/Targeted-Fiber-seq) and as a Supplemental File called Supplemental Code.

Data access

Sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1125891.

Competing interest statement

A.B.S. is a coinventor on a patent relating to the Fiber-seq method (US17/995,058). A.L. is an academic cofounder of Ensoma Inc. The remaining authors declare no competing financial interests.

Acknowledgments

The authors thank Stephen Tapscott for the discussion of the Myotonic Dystrophy 1 data, Shane Neph for assisting in data submission to the NCBI BioProject database, and Chris Boles for technical guidance on the HLS-CATCH method. A.B.S. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a Pew Biomedical Scholar. This study was supported by

the National Institutes of Health (NIH) grants 1DP5OD029630 and UM1DA058220 to A.B.S., as well as a UW ADRC Developmental Project (NIH grant P30AG066509) to A.B.S. M.R.V. and S.C.B. were supported by a training grant (T32) from the NIH (2T32GM007454-46).

Author contributions: Conceptualization and design: S.C.B. and A.B.S. Experimental design and execution: S.C.B., Z.J.A., C.L., B.J.M., J.R., K.M.M., T.W., Y.M., A.L., and A.B.S. Computational experiments: S.C.B., Z.J.A., E.G.S., M.O.H., and M.R.V. Data generation: S.C.B., Z.J.A., C.L., J.R., K.M.M., and A.B.S. Manuscript writing: S.C.B. and A.B.S., with input from all authors.

References

Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, et al. 2020. Massively multiplex single-molecule oligonucleosome footprinting. *eLife* **9:** e59404. doi:10.7554/eLife.59404

Akinsheye I, Alsultan A, Solovieff N, Ngo D, Baldwin CT, Sebastiani P, Chui DHK, Steinberg MH. 2011. Fetal hemoglobin in sickle cell anemia. *Blood* **118:** 19–27. doi:10.1182/blood-2011-03-325258

Alriyami M, Polychronakos C. 2021. Somatic mutations and autoimmunity. *Cells* **10:** 2056. doi:10.3390/cells10082056

Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, Detweiler AM, Neff N, Miga KH, Straight AF, et al. 2022. DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. *Nat Methods* **19:** 711–723. doi:10.1038/s41592-022-01475-6

Amato A, Cappabianca MP, Perri M, Zaghis I, Grisanti P, Ponzini D, Di Biagio P. 2014. Interpreting elevated fetal hemoglobin in pathology and health at the basic laboratory level: new and known γ-gene mutations associated with hereditary persistence of fetal hemoglobin. *Int J Lab Hematol* **36:** 13–19. doi:10.1111/ijlh.12094

Anvret M, Ahlberg G, Grandell U, Hedberg B, Johnson K, Edström L. 1993. Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum Mol Genet* **2:** 1397–1400. doi:10.1093/hmg/2.9.1397

Ashizawa T, Dubel JR, Harati Y. 1993. Somatic instability of CTG repeat in myotonic dystrophy. *Neurology* **43:** 2674–2678. doi:10.1212/WNL.43.12.2674

Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, et al. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91:** 355–358. doi:10.1038/sj.bjc.6601894

Beam Therapeutics Inc. 2022. BEACON: a study evaluating the safety and efficacy of BEAM-101 in patients with severe sickle cell disease (BEACON). https://clinicaltrials.gov/study/NCT05456880.

Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. 2019. Calling somatic SNVs and indels with Mutect2. bioRxiv doi:10.1101/861054

Bollekens JA, Forget BG. 1991. Delta beta thalassemia and hereditary persistence of fetal hemoglobin. *Hematol Oncol Clin North Am* **5:** 399–422. doi:10.1016/S0889-8588(18)30422-2

Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T, et al. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* **68:** 799–808. doi:10.1016/0092-8674(92)90154-5

Buxton J, Shelbourne P, Davies J, Jones C, Van Tongeren T, Aslanidis C, de Jong P, Jansen G, Anvret M, Riley B, et al. 1992. Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. *Nature* **355:** 547–548. doi:10.1038/355547a0

Cheetham SW, Jafrani YMA, Andersen SB, Jansz N, Kindlova M, Ewing AD, Faulkner GJ. 2022. Single-molecule simultaneous profiling of DNA methylation and DNA-protein interactions with Nanopore-DamID. bioRxiv doi:10.1101/2021.08.09.455753v2

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18:** 170–175. doi:10.1038/s41592-020-01056-5

Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J. 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* **40:** e29. doi:10.1093/nar/gkr1146

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43:** 838–846. doi:10.1038/ng.909

Douay L, Giarratana M-C. 2009. Ex vivo generation of human red blood cells: a new advance in stem cell engineering. *Methods Mol Biol* **482:** 127–140. doi:10.1007/978-1-59745-060-7_8

Eaton WA, Hofrichter J. 1987. Hemoglobin S gelation and sickle cell disease. *Blood* **70:** 1245–1266. doi:10.1182/blood.V70.5.1245.1245

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Filippova GN, Thienes CP, Penn BH, Cho DH, Hu YJ, Moore JM, Klesert TR, Lobanenkov VV, Tapscott SJ. 2001. CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat Genet* **28:** 335–343. doi:10.1038/ng570

Frisch R, Singleton KR, Moses PA, Gonzalez IL, Carango P, Marks HG, Funanage VL. 2001. Effect of triplet repeat expansion on chromatin structure and expression of DMPK and neighboring genes, SIX5 and DMWD, in myotonic dystrophy. *Mol Genet Metab* **74:** 281–291. doi:10.1006/mgme.2001.3229

Fu YH, Pizzuti A, Fenwick RG Jr, King J, Rajnarayan S, Dunne PW, Dubel J, Nasser GA, Ashizawa T, de Jong P, et al. 1992. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255:** 1256–1258. doi:10.1126/science.1546326

Grasberger H, Dumitrescu AM, Liao X-H, Swanson EG, Weiss RE, Srichomkwun P, Pappa T, Chen J, Yoshimura T, Hoffmann P, et al. 2024. STR mutations on chromosome 15q cause thyrotropin resistance by activating a primate-specific enhancer of MIR7-2/MIR1179. *Nat Genet* **56:** 877–888. doi:10.1038/s41588-024-01717-7

Hamshere MG, Newman EE, Alwazzan M, Athwal BS, Brook JD. 1997. Transcriptional abnormality in myotonic dystrophy affects *DMPK* but not neighboring genes. *Proc Natl Acad Sci* **94:** 7394–7399. doi:10.1073/pnas.94.14.7394

Harley HG, Brook JD, Rundle SA, Crow S, Reardon W, Buckler AJ, Harper PS, Housman DE, Shaw DJ. 1992. Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature* **355:** 545–546. doi:10.1038/355545a0

Holt M, Saunders C. 2023. MethBat: a battery of methylation tools for PacBio HiFi reads: GitHub Repository. https://github.com/PacificBiosciences/MethBat.

Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, Eberle M. 2024. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40:** btae042. doi:10.1093/bioinformatics/btae042

Holzelova E, Vonarbourg C, Stolzenberg M-C, Arkwright PD, Selz F, Prieur A-M, Blanche S, Bartunkova J, Vilmer E, Fischer A, et al. 2004. Autoimmune lymphoproliferative syndrome with somatic *Fas* mutations. *N Engl J Med* **351:** 1409–1418. doi:10.1056/NEJMoa040036

Jha A, Bohaczuk SC, Mao Y, Ranchalis J, Mallory BJ, Min AT, Hamm MO, Swanson E, Dubocanin D, Finkbeiner C, et al. 2024. DNA-m6A calling and integrated long-read epigenetic and genetic analysis with *fibertools*. *Genome Res* **34:** 1976–1986. doi:10.1101/gr.279095.124

Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, et al. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15:** 591–594. doi:10.1038/s41592-018-0051-x

Klesert TR, Otten AD, Bird TD, Tapscott SJ. 1997. Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of DMAHP. *Nat Genet* **16:** 402–406. doi:10.1038/ng0897-402

Krishnamachari K, Lu D, Swift-Scott A, Yeraliyev A, Lee K, Huang W, Leng SN, Skanderup AJ. 2022. Accurate somatic variant detection using weakly supervised deep learning. *Nat Commun* **13:** 4248. doi:10.1038/s41467-022-31765-8

Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat Methods* **17:** 1191–1199. doi:10.1038/s41592-020-01000-7

Li C, Georgakopoulou A, Newby GA, Everette KA, Nizamis E, Paschoudi K, Vlachaki E, Gil S, Anderson AK, Koob T, et al. 2022. In vivo base editing by a single i.v. vector injection for treatment of hemoglobinopathies. *JCI Insight* **7:** e162939. doi:10.1172/jci.insight.162939

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12:** 733–735. doi:10.1038/nmeth.3444

López Castel A, Nakamori M, Tomé S, Chitayat D, Gourdon G, Thornton CA, Pearson CE. 2011. Expanded CTG repeat demarcates a boundary for abnormal CpG methylation in myotonic dystrophy patient tissues. *Hum Mol Genet* **20:** 1–15. doi:10.1093/hmg/ddq427

Luks VL, Kamitaki N, Vivero MP, Uller W, Rab R, Bovée JVMG, Rialon KL, Guevara CJ, Alomari AI, Greene AK, et al. 2015. Lymphatic and other vascular malformative/overgrowth disorders are caused by somatic

mutations in PIK3CA. *J Pediatr* **166:** 1048–1054.e5. doi:10.1016/j.jpeds.2014.12.069
Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, Narang M, Barceló J, O'Hoy K, et al. 1992. Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. *Science* **255:** 1253–1255. doi:10.1126/science.1546325
Marks P, Banerjee O, Alexander D. 2012. Detection and identification of base modifications with single molecule real-time sequencing data. https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/whitepaper/kinetics.pdf.
Meola G, Cardani R. 2015. Myotonic dystrophies: an update on clinical aspects, genetic, pathology, and molecular pathomechanisms. *Biochim Biophys Acta* **1852:** 594–606. doi:10.1016/j.bbadis.2014.05.019
Miller JW, Urbinati CR, Teng-Umnuay P, Stenberg MG, Byrne BJ, Thornton CA, Swanson MS. 2000. Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J* **19:** 4439–4448. doi:10.1093/emboj/19.17.4439
Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ. 2012. The methylomes of six bacteria. *Nucleic Acids Res* **40:** 11450–11462. doi:10.1093/nar/gks891
Otten AD, Tapscott SJ. 1995. Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. *Proc Natl Acad Sci* **92:** 5465–5469. doi:10.1073/pnas.92.12.5465
Pacific Biosciences. 2023. pb-CpG-tools: GitHub Repository. https://github.com/PacificBiosciences/pb-CpG-tools.
Poduri A, Evrony GD, Cai X, Walsh CA. 2013. Somatic mutation, genomic variation, and neurological disease. *Science* **341:** 1237758. doi:10.1126/science.1237758
Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36:** 983–987. doi:10.1038/nbt.4235
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033
Ranum LPW, Day JW. 2004. Myotonic dystrophy: RNA pathogenesis comes into focus. *Am J Hum Genet* **74:** 793–804. doi:10.1086/383590
Rasmussen A, Hildonen M, Vissing J, Duno M, Tümer Z, Birkedal U. 2022. High resolution analysis of DMPK hypermethylation and repeat interruptions in myotonic dystrophy type 1. *Genes (Basel)* **13:** 970. doi:10.3390/genes13060970
Ravel-Chapuis A, Bélanger G, Yadava RS, Mahadevan MS, DesGroseillers L, Côté J, Jasmin BJ. 2012. The RNA-binding protein Staufen1 is increased in DM1 skeletal muscle and promotes alternative pre-mRNA splicing. *J Cell Biol* **196:** 699–712. doi:10.1083/jcb.201108113
Ravi NS, Wienert B, Wyman SK, Bell HW, George A, Mahalingam G, Vu JT, Prasad K, Bandlamudi BP, Devaraju N, et al. 2022. Identification of novel HPFH-like mutations by CRISPR base editing that elevate the expression of fetal hemoglobin. *eLife* **11:** e65421. doi:10.7554/eLife.65421
Sankaran VG, Menne TF, Xu J, Akie TE, Lettre G, Van Handel B, Mikkola HKA, Hirschhorn JN, Cantor AB, Orkin SH. 2008. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor *BCL11A*. *Science* **322:** 1839–1842. doi:10.1126/science.1165409
Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. 2020. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17:** 319–327. doi:10.1038/s41592-019-0730-2
Steinberg MH, Chui DHK, Dover GJ, Sebastiani P, Alsultan A. 2014. Fetal hemoglobin in sickle cell anemia: a glass half full? *Blood* **123:** 481–485. doi:10.1182/blood-2013-09-528067
Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368:** 1449–1454. doi:10.1126/science.aaz1646
Töpfer A, Wenger A. 2023. Jasmine: predict 5mC in PacBio HiFi reads. https://github.com/PacificBiosciences/jasmine.
Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods* **16:** 88–94. doi:10.1038/s41592-018-0236-3
Vollger MR, Swanson EG, Neph SJ, Ranchalis J, Munson KM, Ho C-H, Sedeño-Cortés AE, Fondrie WE, Bohaczuk SC, Mao Y, et al. 2024. A haplotype-resolved view of human gene regulation. bioRxiv doi:10.1101/2024.06.14.599122v1
Wong LJ, Ashizawa T, Monckton DG, Caskey CT, Richards CS. 1995. Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *Am J Hum Genet* **56:** 114–122.
Yanovsky-Dagan S, Avitzour M, Altarescu G, Renbaum P, Eldar-Geva T, Schonberger O, Mitrani-Rosenbaum S, Levy-Lahad E, Birnbaum RY, Gepstein L, et al. 2015. Uncovering the role of hypermethylation by CTG expansion in myotonic dystrophy type 1 using mutant human embryonic stem cells. *Stem Cell Reports* **5:** 221–231. doi:10.1016/j.stemcr.2015.06.003
Yin M, Wang J, Wang M, Li X, Zhang M, Wu Q, Wang Y. 2017. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res* **27:** 1365–1377. doi:10.1038/cr.2017.131

Received July 9, 2024; accepted in revised form October 15, 2024.