A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker

Yavor K. Bozhilov, Damien J. Downes, Jelena Telenius, A. Marieke Oudelaar, Emmanuel N. Olivier, Joanne C. Mountford, Jim R. Hughes, Richard J. Gibbons & Douglas R. Higgs

Many single nucleotide variants (SNVs) associated with human traits and genetic diseases are thought to alter the activity of existing regulatory elements. Some SNVs may also create entirely new regulatory elements which change gene expression, but the mechanism by which they do so is largely unknown. Here we show that a single base change in an otherwise unremarkable region of the human α-globin cluster creates an entirely new promoter and an associated unidirectional transcript. This SNV downregulates α-globin expression causing α-thalassaemia. Of note, the new promoter lying between the α-globin genes and their associated super-enhancer disrupts their interaction in an orientation-dependent manner. Together these observations show how both the order and orientation of the fundamental elements of the genome determine patterns of gene expression and support the concept that active genes may act to disrupt enhancer-promoter interactions in mammals as in Drosophila. Finally, these findings should prompt others to fully evaluate SNVs lying outside of known regulatory elements as causing changes in gene expression by creating new regulatory elements.
A large proportion of the single nucleotide variants (SNVs) associated with human traits and predisposition to human disease lie within non-coding regions of the genome. Therefore, the causative variants are presumed to affect the fundamental regulatory elements of the genome including enhancers, promoters and boundary elements. We previously described the emergence of a new transcriptional unit in a non-regulatory region of the ‐globin locus which causes down-regulation of ‐globin expression (−thalassaemia) in individuals from Melanesia. In the same study we identified a candidate for the causal mutation: a single nucleotide (T to C) change which resulted in a new binding site for the erythroid master regulator GATA-1 and the production of a new RNA transcript. This provides an example of a trait associated SNV that appears to create a new regulatory element rather than disrupting an existing element. Few “gain-of-function” SNVs have been identified and as yet the prevalence and full repertoire of mechanisms by which such SNVs might alter gene expression have not yet been determined.

The candidate SNV (a T to C transition at coordinate hg19 chr16:209,709) which is thought to down regulate ‐globin expression is found within an unremarkable non-coding region of the ‐globin locus. This multi gene cluster lies within a relatively small, well characterised topologically associated sub-domain (a sub-TAD of ~80 kb) flanked by CTCF-boundary elements within the telomeric region of chromosome 16. The sub-TAD includes embryonic (−HBZ), fetal/adult (−HBA1/2) and theta (−HBO) ‐globin genes which are regulated by four enhancers (R1, R2, R3 and R4) arranged in the order 5′-R1-R2-R3-R4-3′. The locus is only active in erythroid cells where active regulatory regions are found in open chromatin associated with acetylated lysine at position 27 of histone H3 (H3K27ac). Enhancers are marked by monomethyl lysine 4 of histone H3 (H3K4me3) while the difference site is bound by RNA polymerase II (RNAP II) and is marked by a wider half E-box-GATA motif, commonly co-occupied by TFs. The T-C variant generates a recognition motif for the erythroid transcription factor (TF) GATA1 and this particular GATA1 site is part of a wider half E-box-GATA motif, commonly co-occupied by TFs. The T-C variant generates a recognition motif for the erythroid transcription factor (TF) GATA1 and this particular GATA1 site is part of a wider half E-box-GATA motif, commonly co-occupied by TFs.

To understand how the local DNA sequence might facilitate the establishment of the novel C-SNV promoter we studied the site at the 5′ end of the transcribed region. We mapped the transcription start site (TSS) using RLM-RACE and identified the most likely transcription initiation element to be the core promoter element XCP(E17) (Supplementary Fig. 2). The T-C variant generates a recognition motif for the erythroid transcription factor (TF) GATA1 and this particular GATA1 site is part of a wider half E-box-GATA motif, commonly co-occupied by TFs. Correspondingly the site of the mutation is bound both by GATA1 and KLF1 in primary erythroblasts homogenous for the mutation. In addition, we identified a strong binding motif for another key erythroid TF KLF1, in the immediate vicinity of the GATA1 motif (Fig. 2A). Using ChIP-seq we found a near two-fold increase in GATA1 occupancy and over four-fold increase in KLF1 binding in the vicinity of the C-SNV (Fig. 2B). This is of interest since both GATA1 and KLF1 are master regulators of erythroid genes and have a role in establishing and/or maintaining chromatin conformation and are often found to co-occupy the same sites in erythroid cells. Of interest, we analysed the predicted chromatin accessibility of the human ‐globin gene cluster with and without the candidate SNV using a recently established, convoluted neural network (DeepHaem) [https://github.com/rschwess/deepHaem]. DeepHaem has previously been used to predict chromatin accessibility and its effect on higher order chromatin structures using DeepC. This network predicts that the C-SNV mutation alone is sufficient to create an open chromatin site whose accessibility increases as cells differentiate along the erythroid pathway (Supplementary Fig. 2b).

To determine experimentally whether the T to C transition is the sole causative variant we introduced this point mutation homozygously into wild type cells and reverted the mutation in mutant cells (for genetic screening and quality control see Supplementary Fig. 2). Editing the mutant C allele to a T (C-T lines) abolished the accessible chromatin site surrounding the mutation and transcription at the site of the SNV became undetectable (Fig. 2C-E). Conversely, introducing the mutant C allele in wild type cells (T-C lines) produces a region of accessible chromatin at the site of the SNV and transcription of this region (Fig. 2C-E). We performed qPCR to determine if the T to C transition causes downregulation of ‐globin. The C-T lines showed an over two-fold increase in ‐globin mRNA when compared to its isogenic control C-SNV line. Conversely, the T-C lines showed over 55% reduction in ‐globin transcript (Fig. 2D). These results prove that the T to C transition is both necessary and sufficient to cause the observed phenotype and that the accumulation of erythroid TFs around the TSS is causative rather than coregulatory in creating a new promoter element.
The new promoter sequesters enhancers away from native promoters. The emergence of a new promoter, induced by key erythroid transcription factors like GATA1, in an erythroid-specific locus suggests that its expression may be regulated by the local enhancers. Indeed, GATA1 is thought to promote chromatin looping through its binding partners, among those that have shown to play a role in chromatin conformation are the cofactor FOG1 and the multimeric LDB1-complex. The LDB1-complex consists of TAL1/E-protein heterodimer bound to GATA1 via LMO2 and LDB1 and is thought to achieve transcriptional regulation through facilitating interactions between distal regulatory elements and promoters. Genome wide studies have identified frequently occurring composite TAL1 and GATA1 binding sites known as Ebox-GATA1 motifs at erythroid regulatory elements confirming their co-association. In addition, GATA1 has been shown to cooperate with the TF NF-E2 to contribute to the recruitment of the ATPase component of SWI/SNF chromatin remodelling complex in order to facilitate the formation of accessible chromatin and to maintain higher order chromatin structure. In particular, BRG1 is required for long range interactions in the mouse α-globin locus where GATA1 binds both the enhancers and the globin promoters early.
Fig. 1 New transcriptional unit bears the marks of a unidirectional promoter and causes α-globin downregulation. A Chromatin accessibility in the α-globin locus as measured by ATAC-seq. The enhancer elements (RT-R4) are highlighted in orange, the site of the T to C mutation is highlighted in green (labelled SNV for single nucleotide variant), gene annotation by Refseq is in blue. Read-densities represent an average of 3 independent differentiation experiments, 3 independent wild type iPSC lines (labelled WT) or 3 iPSC clones obtained from the same patient material homozygous for the C allele of the SNV located at coordinate (hg19) chr16:209,709 (labelled C-SNV), differentiated to erythroblasts. Coordinates (hg19) chr16:108,000-238,000. B ChIP-seq, highlighted regions are as in A. Read-densities represent an average of 3 independent experiments, 3 replicates for wild type iPSC line AH017-13 (in blue) or 3 C-SNV iPSC clones obtained from the same patient material (in red) differentiated to erythroblasts. The tracks are overlaid on top of each other, black indicates shared signal while red and blue indicate signal unique for the mutant and wild type lines, respectively. Coordinates (hg19) chr16:108,000-238,000. The level of the signal in the middle of the C-SNV transcriptional unit is most likely affected by the presence of a variable number tandem repeat (inter-ζ VNTR), a 1 kb sequence in the reference genome which in reality can be much larger (over 2 kb). The artificial reduction of the reference genome means that the same signal is collapsed to a smaller length resulting in at least 2-fold signal increase over the middle of the region of enrichment at the C-SNV locus. Since the exact sequence or size of the repeat is not known in the genomes analysed, this has not been corrected for. C Strand-specific RNA-seq of polyadenylated selected (p+−) and non-polyadenylated (p−−) RNA, read density (in RPMK) represents an average of 3 independent experiments, 3 replicates for wild type line AH017-13 or 3 clones of C-SNV iPSCs differentiated to erythroblasts. The region of the T to C mutation is highlighted in green (SNV), gene annotation by Refseq is in blue, pseudo genes are in pink. Coordinates (hg19) chr16:209,000-217,000. D qPCR quantification of HBA1/HBA2 in reference to RPS18 in mRNA obtained from 3 independent wild type iPSC lines (WT) or 3 iPSC clones obtained from the same patient material (C-SNV) differentiated to erythroblasts. All lines were differentiated twice (one replicate was removed as an outlier due to low levels): WT (n = 6) in blue, C-SNV (n = 5) in red. Violin plots display median (dotted black line) quartile lines (coloured dotted line) and individual data points (black dots). P-values are obtained using unpaired, two-tailed student t-test.

on during erythroid differentiation40. To test if the C-SNV element comes into close proximity with the α-globin enhancers, we assessed the interaction between key sites in the α-globin locus using Capture-C41. Analysis of the interaction profiles of C-SNV erythroid cells and wild type cells, from the viewpoint of the new promoter, shows an increase in mean interaction frequency between the α-globin enhancers and the new promoter (Fig. 3A). To determine if this interaction occurs at the expense of contacts between the α-globin enhancers and their cognate promoters we performed Capture-C from the viewpoint of the promoters of the α-globin genes (HBA1 and HBA2). This showed a marked decrease in mean interaction frequency between the α-globin promoters and enhancers in the presence of the new C-SNV promoter (Fig. 3B). To confirm that this observation holds true from the point of view of the α-globin enhancers we also performed Capture-C from the R2 enhancer (Fig. 3C). This demonstrates that in erythroid cells homozygous for the C allele, the α-globin enhancers display a higher mean interaction frequency over the region of the active C-SNV promoter and a reduction of interactions with the α-globin promoters, and that this is likely to be an important component underlying the reduction in α-globin expression.

Reduced contacts between the enhancers and the α-globin promoters in the presence of the newly formed promoter could be explained by a model of mutually exclusive promoter competition, which has previously been proposed to explain several forms of complex gene regulation42,43. Alternatively, this could be explained by the formation of an insulator element by the new promoter, reminiscent of that observed at a subset of Drosophila promoters which act as enhancer blockers; a type of insulator element that restricts enhancer–promoter interactions44. Structural boundary elements that regulate chromatin architecture are commonly associated with the presence of the CCCTC-binding factor (CTCF)45,46. A CTCF ChIP-seq analysis showed that there are no new CTCF peaks that appear in the C-SNV lines, demonstrating that the activity of the C-SNV promoter alone is the cause of the shift in chromatin interactions (Fig. 3D).

The competition model cannot explain the full effect of the new promoter. We next wanted to understand whether the effect of the C-SNV promoter was due to mutually exclusive promoter competition which should result in the promoter exerting its effect on chromatin interactions and gene expression irrespective of its position in the sub-TAD. Alternatively, if the C-SNV promoter is acting as an enhancer blocker its location (and perhaps orientation) relative to the native promoters and enhancers in the locus would determine its effect on gene regulation. To test the competition model, we placed the sequence of the active C-SNV promoter upstream of the α-globin enhancers in a wild type line, in the antisense orientation, directed away from the enhancers (Fig. 4A) (for genome editing design and genetic screen see Supplementary Fig. 3). Importantly, this insertion still lies within the α-globin sub-TAD and therefore should be equally accessible to the enhancers. In addition, the ectopically placed C-SNV promoter lies closer to the major R2 enhancer (14 kb) than the α-globin genes (59 kb) or the original position of the C-SNV (46 kb). ATAC-seq showed that the transposed C-SNV promoter sequence opens chromatin in this new position (Fig. 4B). The associated chromatin showed a large increase in H3K27ac and H3K4me3, and a less pronounced increase in H3K4me1 (Fig. 4C) together with unidirectional transcription originating at the TSS of the transposed C-SNV promoter (Fig. 4D). This transcript extends away from the enhancers on what is now the antisense strand of DNA. Thus, when placed within the sub-TAD upstream of the enhancers, the C-SNV sequence can still act as a bona fide promoter. Interestingly, when the promoter was placed in the sense orientation, it again created a region of open chromatin marked by H3K27ac and H3K4me1 but not by H3K4me3. In this case there were no detectable RNA transcripts (Supplementary Fig. 4). Together, these observations suggest that the orientation of this promoter relative to other elements may determine whether or not it is recognised as a promoter.

To test if the transposed transcriptionally active C-SNV promoter affected α-globin expression, we performed qPCR analysis. There was a reduction (25.6%, p = 0.0322) in mean α-globin transcript levels seen between cells where the transcriptionally active C-SNV promoter is placed upstream of the α-globin enhancers and wild type cells. However, we observed a larger reduction (55.7%, p = 0.0005) in cells of the same genomic background in which the C-SNV promoter is in its original position between the enhancers and α-globin promoters (T-C) (Fig. 4E). This is not sufficient to explain the severity of the phenotype that the C-SNV promoter produces in its native location. To determine if the transposed C-SNV promoter exerts any effect on chromatin interactions in the α-globin locus when placed upstream of the enhancers we performed Capture-C from enhancer R2 and the promoters of HBA1 and HBA2. This showed no major change in chromatin interactions between the α-globin promoters and their enhancers in line with the minor change in
**Fig. 2 Point mutation causes the emergence of a de novo promoter.** A 138 bp unique sequence around the causative point mutation. The SNV is labelled using the degenerate DNA symbol Y (T/C), key features are highlighted: KLF1 binding site (magenta), GATA1 (red), composite half E-box-GATA motif (orange), core promoter element XCPE1 (blue) and TSS (black). B ChIP-qPCR for GATA1 and KLF1, sequences for TaqMan assays can be found in Supplementary Table 1. Values represent an average of 3 independent experiments, 3 replicates for wild type iPSC line AH017-13 differentiated to erythroblasts (in blue) or 3 C-SNV iPSC clones (in red). Error bars represent one standard deviation. P values are obtained using unpaired, two-tailed student t-test. C Chromatin accessibility by ATAC-seq. The enhancer elements (R1 – R4) are highlighted in orange, the site of the T to C mutation is highlighted in green (SNV), gene annotation by Refseq is in blue. Read-densities represent an average of 3 independent differentiation experiments: wild type iPSC line SB-AD2-01 (labelled WT), 3 C-SNV iPSC clones (labelled C-SNV), 4 clones of edited SB-AD2-01 cells where the T base at position 209,709 (hg19) of chr16 was changed to a C (labelled T–C), 4 clones of edited C-SNV iPSC cells (line LA01) where the C base at position 209,709 (hg19) of chr16 was changed to a T (labelled C–T). Coordinates (hg19) chr16:108,000-238,000. D qPCR quantification of HBA1/HBA2 in reference to RPS18 in mRNA obtained from independent differentiation experiments: 5 from wild type iPSC line SB-AD2-01 (WT), 3 C-SNV iPSC clones (C–SNV) differentiated to erythroblasts twice, 4 T–C clones (labelled T–C), 4 C–T clones (labelled C–T). WT (n = 5) in blue, C–SNV (n = 5) in red, T–C (n = 4) in cyan, C–T (n = 4) in magenta. Violin plots display median (dashed black line) quartile lines (coloured dotted line) and individual data points (black dots). P values are obtained using unpaired, two-tailed student t-test. E qPCR quantification of novel transcript (TaqMan assay in Supplementary Table 1) in reference to RPS18 in mRNA obtained from independent differentiation experiments as in D). P values are obtained using unpaired, two-tailed student t-test.
These observations indicate that placing the element within the sub-TAD but outside of the region between the enhancers and promoters does not significantly interfere with enhancer-promoter interactions as judged by 3C experiments. This is in line with our recent studies showing that the extension of the α-globin domain to encompass additional genes does not result in promoter competition but results in the formation of a larger chromatin hub where the extra elements together with the native elements take part in multiway interactions without causing major changes in α-globin expression

expression (Fig. 4F). Together these findings suggest that although there may be some competition between the newly formed promoter and the enhancers, the C-SNV promoter may also act by blocking enhancer-promoter interactions when placed between the α-globin enhancers and their promoters. One caveat of this interpretation lies in the uncertainty of whether promoter strength might determine its ability to successfully compete with other promoters. Indeed, the signal levels of the promoter mark H3K4me3 are much lower in the promoter insertion lines, for this reason we returned to examining the promoter in its native position.

The new promoter acts as orientation-dependent enhancer-blocker. Elements which limit the interactions between promoter

![Fig. 3 New promoter sequesters interactions with the α-globin enhancers away from their cognate promoters. A Top panel shows overlaid, normalised Capture-C data for the novel promoter (SNV Viewpoint) in erythroid cells derived from 3 independent differentiation experiments: 3 biologically independent samples of wild type iPSC line AH017-13 (WT in blue) or 3 clones of C-SNV iPSCs obtained from the same patient material (C-SNV in red). The mean, plus and minus one standard deviation (S.D.), of sliding 5 kb windows are visualised as a dark coloured line and a lighter coloured shadow, respectively. Differential tracks (in dashed rectangle) show a subtraction (C-SNV minus WT) of the mean value of meaningful interactions per restriction fragment: positive values (in black) display an increase in interaction over a region in the mutant cells (C-SNV) compared to wild type, while negative values (in grey) indicate a decrease. The enhancer elements (R1 to R4) are highlighted in orange, the site of the T to C mutation is highlighted in green (SNV), gene annotation by Refseq is in black (except α-globin genes in purple). Coordinates (hg19) chr16:69,200-327,999. B Data presented as in A, viewpoint is promoters of HBA1 and HBA2. C Data presented as in A, viewpoint is enhancer R2. D ChIP-seq for CTCF, enhancer elements (R1 to R4) are highlighted in orange, the site of the T to C mutation is highlighted in green (SNV), gene annotation by Refseq is in blue. Read-densities represent an average of 3 independent differentiation experiments, 3 replicates for wild type iPSC line AH017-13 (WT in blue) or 3 C-SNV iPSC clones (C-SNV in red) differentiated to erythroblasts. The third track (CTCF) displays WT and C-SNV overplayed, black indicates shared signal while red and blue indicate signal unique for the mutant and wild type lines, respectively. Coordinates (hg19) chr16:108,000-238,000.]

6
and enhancer in mammalian cells such as boundaries have been shown to commonly act in an orientation dependent manner. It is possible that this novel element may act in a similar manner, although we could find no evidence for orientation-specificity in promoters that act as enhancer-blockers in Drosophila or yeast.

To test if the orientation of the C-SNV element contributes to its ability to block enhancer-promoter interactions, we inverted the active C-SNV promoter sequence in its natural position (Fig. 5A) (for genome editing design and genetic screen see Supplementary Fig. 5). Asking if the element still acted as a unidirectional promoter after the inversion we showed that the inverted element is still capable of opening chromatin, is marked by H3K4me3 and is capable of recruiting RNAP II (Fig. 5B, C). Importantly, the ATACseq and H3K4me3 signal levels over the C-SNV promoter
were similar to those in the C-SNV lines, suggesting that even when inverted promoter strength is maintained. We also proved that the unidirectional nature of the transcriptional activity is preserved, with transcripts with similar levels of expression now extending towards the enhancers (Fig. 5D). To determine whether the C-SNV promoter still blocked enhancer-promoter interactions when inverted, we performed Capture-C from the α-globin promoters and enhancer R2, and qPCR analysis for α-globin expression. This showed that after inversion of the C-SNV promoter the α-globin promoters interact more frequently with their enhancers while the contacts between the α-globin enhancers and the inverted C-SNV promoter decrease (Fig. 5F). Consistent with this, the levels of α-globin mRNA return to wild type levels following the inversion of the C-SNV promoter (Fig. 5E). Since the C-SNV sequence appears to act as an equally strong promoter in either orientation, it should “compete” equally for the enhancer in either orientation. By contrast, we have shown that the inverted C-SNV promoter no longer causes local gene mis-regulation. These observations support the hypothesis that the new (C-SNV) promoter acts as an enhancer-blocker and that this effect is constrained by its orientation in the active locus.

Discussion

In this work, we have found that a gain of function SNV causes the emergence of an active promoter in a tissue-specific gene locus, disrupting normal chromatin interactions and gene regulation. By changing the position and orientation of the novel promoter, we show that rather than simply competing for the activity of the enhancer within a defined sub-TAD, the new promoter acts in an orientation-dependent manner. It predominantly disrupts chromatin interactions only when placed between the enhancers and the promoters of the α-globin genes, and only in one orientation (see Fig. 6). This raises two possibilities to explain the effects of the C-SNV promoter at its natural location between the enhancers and the α-globin promoters. One is that the resulting reduction in α-globin expression could reflect a role for the C-SNV element as an orientation-dependent enhancer-blocker. Alternatively, it could be that the enhancer-promoter interaction is weaker when the C-SNV promoter is inverted with respect to the enhancer, with the promoter inversion altering the enhancer’s ability to recognise its target. This, in turn, might reduce the C-SNV promoter’s ability to compete for the enhancer activity. However, there is no current evidence, to our knowledge, to suggest that enhancers differ in their interactions with respect to the orientation of their cognate promoters. If so, this would contest the commonly held hypothesis that enhancers and promoters interact equally regardless of their orientation. Of interest, we have also recently shown that the α-globin promoters themselves may act to delimit enhancer contacts within the context of the sub-TAD. Together these observations support the concept that promoters may act to block enhancer-promoter interactions in mammals as in Drosophila. However, the mechanism(s) by which a promoter may act as an orientation dependent enhancer-blocker are not yet known. We can only speculate that the motif grammar, the conformation of multiprotein complexes at the promoter and/or the direction and the act of transcription may all play a role in this. Further studies varying these parameters will be required to determine such mechanisms. The observation that a tissue-specific promoter can act to block an enhancer promoter interaction in an orientation-dependent manner raises the question of whether other promoters, most notably those that are situated at the edges of self-interacting domains, might exhibit the same type of effect and whether their orientation might serve as an additional mechanism that cells use to direct specificity of interactions and shape local chromatin conformation. Finally, variants like C-SNV which generate de novo regulatory elements can only be identified by observation in the correct genotype or, as shown here, through predictive machine learning approaches. These findings should prompt others to re-evaluate SNVs lying outside of known regulatory elements when studying human traits associated with natural variation.

Methods

Induced pluripotent stem cell maintenance. Human iPSCs were cultured at 37 °C 5% CO2 in mTeSR1 medium (StemCell Technologies) on tissue culture treated plastic ware (Corning) coated with Human ESC Qualified Matrigel (Corning) with daily medium changes. Cultures were passaged 1 in 8 when 70–80% confluent using a non-enzymatic method employing a 0.02% EDTA containing solution (Versene from Lonza). Frozen stocks were stored in 30% mTeSR1 30% Knock-Out Serum Replacement (Gibco) 10% Knock-Out DMEM (Life Technologies) 10% DMSO (SIGMA) under liquid Nitrogen vapour. Frozen stocks were resurrected in mTeSR1 medium containing 10 μM ROCK inhibitor (RnD).
Wild type human induced pluripotent stem cell lines SB-AD2-01, SB-AD3-01 and AH017-13 were obtained from Oxford StemBANCC. C-SNV human induced pluripotent stem cell lines LA01, LA06 and LA13 were generated and validated in house. Ethical approval for the study was granted by North West Research Ethics Committee of NHS National Research Ethics Services, Ref: MREC 03/8/097. The experimental procedures in this study were carried out in accordance with the approved guidelines. Informed consent was obtained for the generation of human derived iPSC lines.

Reprogramming to pluripotency. Reprogramming EBV immortalised B-cells from an individual homozygous the C allele (C-SNV) was performed using the CytoTune-iPS 2.0 Sendai Reprogramming Kit (Life Technologies) according to manufacturer’s protocol using the feeder-based reprogramming method on Mitomycin C inactivated MEF plated dishes. When iPSC colonies were ready for transfer (day 18–20) they were transferred onto a Matrigel coated plate in mTeSR1 medium. Subsequently, the lines were maintained for 2 months until they reached passage 17–20 and then quality control was performed, and the lines were stored in...
Fig. 5 Changing the orientation of the promoter alleviates repression and increases α-globin enhancer-promoter contacts. A The native promoter was inverted using heterotypical loxP sites located 4011 bp upstream and 2975 bp downstream of the XCPE1 TSS (fragment chr16:205,734-212,720, hg19). B ATAC-seq: the enhancer elements (R1 to R4) are highlighted in orange, the site of the T to C mutation is highlighted in green (SNV), the inverted region is highlighted in light blue, gene annotation by Refseq is in blue. Read-densities represent an average of: 3 C-SNV iPSC clones (labelled C-SNV), 3 clones of edited C-SNV cell line LA01 where the heterotypical loxP sites are inserted but the C-SNV promoter segment is not inverted (labelled nonINV), 3 clones of edited patient line LA01 where the C-SNV promoter segment is inverted (labelled INV). Reads for C-SNV and nonINV were mapped to the normal genome, reads for INV were mapped to a custom genome where a 7 kb segment containing the C-SNV promoter is inverted. Coordinates (hg19) chr16:108,000-238,000. C ChIP-seq, highlighted regions are as in a). Read-densities represent an average of 2 or 3 independent differentiation experiments, 3 C-SNV iPSC clones (C-SNV in red), 2 clones of INV cells (INV in purple). Reads for C-SNV were mapped to the normal genome, reads for INV were mapped to a custom genome. Coordinates (hg19) chr16:108,000-238,000. D Strand-specific RNA-seq of polyadenylated selected (pA+) and non-polyadenylated (pA-) RNA, read density (in RPKM) represents an average of: 3 C-SNV iPSC clones (C-SNV in red), 2 clones of INV cells (INV in purple). The inverted segment is highlighted in blue, the location of the LoxP sites is marked by orange triangles, the region of the T to C mutation is highlighted in green (SNV), gene annotation by Refseq is in blue with pseudogenes in pink. Reads for C-SNV were mapped to the normal genome, reads for INV were mapped to a custom genome. Coordinates (hg19) chr16:202,000-217,000. E qPCR quantification of HBA1/HBA2 in mRNA obtained from independent differentiation experiments: 3 independent wild type iPSC lines differentiated twice (WT), 3 C-SNV iPSC clones (C-SNV) differentiated twice (one replicate removed as an outlier), 3 clones of nonINV cells (nonINV), 3 clones of INV cells (INV). WT (n = 6) in blue, C-SNV (n = 5) in red, nonINV (n = 3) in orange, INV (n = 3) in purple. Violin plots display median (dashed black line) quartile lines (coloured dotted line) and individual data points (black dots). P-values are obtained using unpaired, two-tailed student t-test. F Top panel shows overlaid, normalised Capture-C data for either the promoters of the α-globin genes (left graph) or enhancer R2 (right graph) in erythroid cells derived from 3 independent differentiation experiments: 3 C-SNV clones (C-SNV in red) or 3 clones of INV cells (INV in purple). The mean, plus and minus one standard deviation (S.D.), of sliding 5 kb windows are visualised as a dark coloured line and a lighter coloured shadow, respectively. Differential tracks (in dashed rectangle) show a subtraction (INV minus C-SNV) of the mean value of meaningful interactions per restriction fragment. Reads were mapped to the wild type genome. Coordinates (hg19) chr16:69,200-327,999.

Fig. 6 C-SNV promoter behaviour in different contexts within the α-globin locus. Schematic shows the effect the active C-SNV promoter has on the regulation of the α-globin locus when placed either in its native location (labelled C-SNV promoter in green) or inserted behind the α-globin enhancers (labelled Promoter insertion in green). Elements of importance are highlighted on an artificially merged ATACseq track: enhancers R1-R4 in orange and HBA2/1 genes in red. The effects of the different genomic arrangements are adjacent to the conformation of the element on the schematic.
cryogenic store. Human iPSC lines at passage 17–20 were assessed for pluripotency marker expression using the PSC 4-Marker Immunocytochemistry Kit (Invitrogen) as per manufacturer’s instructions. Nuclei were stained with DAPI and analyzed using epifluorescence microscopy on a Leica DMi8 with a high quality genomic DNA using the Infinium HD assay on Human CytoSNP 12 BeadChip v2.1 (Illumina) at the Wellcome Trust Centre for Human Genetics (Oxford). Data was analysed using the KaryoStudio package (Illumina).

Genome editing of iPSCs. DNA was introduced into human iPSCs by a lipofection based method using Fugene 6 (Promega). Up to 8 μg of DNA was used per 105 cells, equal amounts of single-stranded repair template DNA and p5pCas9 (BB)-2A-Puro (PX459) V2.0 plasmid DNA. For point mutation and box site insertion single stranded oligonucleotides were ordered from IDT. For promoter insertion single stranded DNA was generated in house using λ exonuclease and a 5′ phosphorylated primer (for sequences see Supplementary Table 1). Exogenous DNA containing cells were selected using puromycin for 24 h, cells were then maintained as normal for one passage. The puromycin resistant cells were brought to a single cell suspension using Accutase (Millipore) and seeded at low density (~800 cells in suspension per 10cm2). After 7–9 days iPSC colonies were picked into a 96-well plate coated with Matrigel in mTeSR1 medium with 10 μM ROCK inhibitor. The 96 well plate was split to create 2 replica plates, one of the replica plates was used for genotyping the other one was used to expand the clones with correct edits.

Erythroid differentiation of iPSCs. Prior to the initiation of differentiation protocol, human iPSCs were cultured for at least 15 days on Vitronectin (Invitrogen) coated plates (Corning TC treated) in STEMPRO HESCA media (Invitrogen) supplemented with 20 ng/ml bFGF (RnD). Cultures were maintained in this way for 5–6 passages at high confluency and split 1 in 4 when over 85% confluent. Human iPSCs were differentiated as published by Olivier et al.12 and harvested on day 21 of the differentiation protocol. Flow cytometry analysis was performed on 2 × 106 cells resuspended in 1 ml of 2% bovine serum albumin (BSA) buffered saline and labelled for 20 min at 4 °C with 1:200 dilution of fluorescein isothiocyanate (FITC) conjugated anti-CD71 (BD Biosciences; 555536) and 1:200 dilution of phycoerythrin (PE) conjugated anti-CD235a (BD Biosciences; 340947). After staining cells were harvested and resuspended in PBS containing 0.02% Hoechst 33342 and 50 μM acetyl-Histone H3 (Lys27) antibody 17-683 (Merck Millipore) 2 μg per 1 ml of chromatin from 1 × 107 cells; Polyclonal anti-CTCF antibody ab195391 (Abcam) 4 μg per 1 ml of chromatin from 1 × 107 cells; Polyclonal anti-KLF1 antibody was kindly provided by the Perkins Center for Cancer Research (Cambridge, MA) 2 μg per 1 ml of chromatin from 1 × 107 cells. Chromatin was eluted off beads using 20 mM Tris-HCl, 5 mM EDTA, 50 mM NaCl. The diluted chromatin was precleared twice with 10 µl of a 1:1 mix of protein G Dynabeads and 1:200 dilution of anti-CTCF antibody. PCR duplicates and ploidy regions were removed, and biological replicates were normalised to reads per 105.

The antibiotics used are: RNA Polymerase II N-20 antibody SC-899 (Santa Cruz Biotechnology) 6 μg per 1 ml of chromatin from 1 × 105 cells; Monoclonal anti-acetyl-Histone H3 (Lys27) antibody 17-683 (Merck Millipore) 2 μg per 1 ml of chromatin from 1 × 105 cells; Polyacrylamide gel electrophoresis (PAGE) glycine buffer 20 μg per 1 ml of chromatin from 1 × 105 cells; Polyclonal anti-trimethyl-Histone H3 (Lys4) antibody 07-473 (Merck Millipore) 1 μg per 1 ml of chromatin from 1 × 105 cells; Polyclonal anti-monomethyl-Histone H3 (Lys4) antibody ab195391 (Abcam) 4 μg per 1 ml of chromatin from 1 × 105 cells; Polyclonal anti-CTCF antibody 07-729 (Merck Millipore) 10 μl serum per 1 ml of chromatin from 1 × 105 cells; Polyclonal anti-KLF1 antibody was kindly provided by the Perkins Laboratory, Translational Research Institute, Brisbane, Australia (used 30 μl serum per 1 ml of chromatin from 1 × 105 cells).

RNA expression analysis and TSS mapping. Total RNA was isolated from 5 × 105 day 21 erythroid cells using the Direct-zol RNA kit (Zymo) as per manufacturer’s instructions. mRNAs were reverse-transcribed using the Transcriptor First Strand Synthesis Kit (Roche) according to the manufacturer’s instructions. cDNA was quantified using the IQ SYBR Green kit (Thermo Fisher Scientific). The expression of each of the genes was normalized to the expression of the housekeeping reference gene. The statistical analysis was performed using the t-test.

Next generation capture C. Next generation capture C was performed as previously published41. Briefly, 7 × 104 cells per replicate were lysed and nuclei were isolated prior to hybridization with 500 ng of Human CytoSNP v2.1 (Illumina) for 30 min at 37 °C. Tagged DNA was purified using the Infinium HD Assay MethylElute kit (Illumina). The eluted DNA was then quantified using a NanoDrop (ThermoFisher Scientific) and an Agilent 2100 Bioanalyzer (Agilent). The DNA was fragmented using an Infinium HD assay on Human CytoSNP 12 BeadChip v2.1 (Illumina) at the Wellcome Trust Centre for Human Genetics (Oxford). Data was analysed using the KaryoStudio package (Illumina).
deviation for each replicate (n = 3). Mean reporter counts were divided into 250 bp bins and smoothed using a 5 kb window.

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

**Data availability**

The datasets generated during and/or analyzed during the current study are available in FASTA format as supplementary files in GEO Subseries GSE159871. Figures that have associated raw data files: Fig. 1A–C; Fig. 2C; Fig. 3A–D; Fig. 4B–I; Supplementary Fig. 4b–d). Source data are provided with this paper.

Received: 5 November 2020; Accepted: 19 May 2021; Published online: 21 June 2021

**References**

1. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet* 7, 862–872 (2006).

2. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195 (2012).

3. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking sequence variation in regulatory DNA. *Genome Res* 22, 1748–1759 (2012).

4. Wienert, B. et al. KLFI drives the expression of fetal hemoglobin in British HFFH. *Blood* 130, 803–807 (2017).

5. Martyn, G. E. et al. A natural regulatory mutation in the proximal promoter, hg19_Vas (promoter insertion in anti-sense behind enhancers) or hg19_Vs (promoter insertion in sense behind enhancers) based on reference human genome hg19, as indicated in figure legends. Sequences for chromosome 16 from the custom genomes are available in FASTA format as supplementary files in GEO Subseries GSE159871. Figures that have associated raw data files: Fig. 1A–C; Fig. 2C; Fig. 3A–D; Fig. 4B–I; Supplementary Fig. 4b–d). Source data are provided with this paper.

6. Hughes, J. R. et al. High-resolution analysis of cis-acting regulatory networks. *Trends Genet* 30, 1–9 (2014).

7. Wienert, B. et al. Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nat. Commun.* 6, 7085 (2015).

8. Hughes, J. R. et al. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl Acad. Sci. USA* 102, 9830–9835 (2005).

9. Hughes, J. R. et al. High-resolution analysis of cis-acting regulatory networks at the alpha-globin locus. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368, 20120361 (2013).

10. Olivier, E. N. et al. High-efficiency serum-free feeder-free erythroid differentiation of human pluripotent stem cells using small molecules. *Stem Cells Transl. Med.* 5, 1394–1405 (2016).

11. Bernstein, B. E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169–181 (2005).

12. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130, 77–88 (2007).

13. Kim, T. H. et al. A high-resolution map of active promoters in the human genome. *Nature* 436, 787–880 (2005).

14. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560 (2007).

15. Tokusumi, Y., Ma, Y., Song, X., Jacobson, R. H. & Takada, S. The new core promoter element XCPF1 (X Core Promoter Element 1) directs activator-, -mediator-, and TATA-binding protein-dependent but TFID-independent RNA polymerase II transcription from TATA-less promoters. *Mol. Cell Biol.* 27, 1844–1858 (2007).

16. El Omari, K. et al. Structural basis for LMO2-driven recruitment of the SCL: E47bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell Rep.* 4, 135–147 (2013).

17. Kassouf, M. T. et al. Genome-wide identification of TALI’s functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res.* 20, 1064–1083 (2010).

18. Soler, E. et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev.* 24, 277–289 (2010).

19. De Gobbi, M. et al. A regulatory SNP causative of primary erythroid disease by creating a new transcriptional promoter. *Science* 312, 1215–1217 (2006).

20. Hodge, D. et al. A global role for EKLF in definitive and primitive erythropoiesis. *Blood* 107, 3359–3370 (2006).

21. Love, P. E., Warrecha, C. & Li, L. Ldb1 complexes: the new master regulators of erythroid gene expression. *Trends Genet.* 30, 1–9 (2014).

22. Tallack, M. R. et al. Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq in primary erythroid cells. *Genome Res.* 20, 1052–1063 (2010).

23. Love, P. E., Warrecha, C. & Li, L. Ldb1 complexes: the new master regulators of erythroid gene expression. *Trends Genet.* 30, 1–9 (2014).

24. Bartman, C. R., Hsu, S. C., Hsiung, C. C., Raj, A. & Blobel, G. A. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol. Cell* 28, 1278–1290 (2014).

25. Li, L. et al. Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation. *Blood* 121, 4575–4585 (2013).

26. Kim, S. I., Bultman, S. J., Kiefer, C. M., Dean, A. & Bresnick, E. H. BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc. Natl Acad. Sci. USA* 106, 2259–2264 (2009).

27. Telenius, J. & Hughes, J. R. NGseqBasic - a single-command UNIX tool for ATAC-seq, DNase-seq, Cut-and-Run, and Chip-seq data mapping.

28. Kang, Y., Kim, Y. W., Yun, J., Shin, J. & Kim, A. KLF1 stabilizes GATA-1 and TAL1 occupancy in the human beta-globin locus. *Biochim. Biophys. Acta* 1849, 282–295 (2015).

29. Kim, S. I., Bultman, S. J., Kiefer, C. M., Dean, A. & Bresnick, E. H. Dissecting molecular steps in chromatin domain activation during hematopoietic differentiation. *Mol. Cell* 27, 4551–4565 (2007).

30. Anguita, E., Johnson, C. A., Wood, W. G., Turner, B. M. & Higgs, D. R. The role of BRG1 in erythropoiesis. *Science* 312, 3135–3137 (1997).

31. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233–1244 (2012).

32. Riezman, H. & Klumperman, J. The endosome system. *Nature Rev. Mol. Cell Biol.* 12, 587–600 (2011).

33. Wang, J. et al. Generation of human induced pluripotent stem cells by overexpression of OCT3/4, SOX2, c-Myc, and KLF4. *Science* 318, 1358–1361 (2007).

34. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 363–376 (2006).

35. Poueymirou, W. T., Wang, C. Y., Saha, S. & Thomson, J. A. Generation of embryonic cell lines from adult human diploid fibroblasts. *Science* 233, 1130–1134 (1986).

36. Cooper, D. N. et al. The ENCODE project, a trans-disciplinary, trans-institutional human functional genomics project: overview of the project and its results. *Nature* 447, 499–511 (2007).

37. Robinson, J. R., Bullock, A. C., Taft, R. J. & Young, R. A. Chromatin looping and transcription regulation. *Cell* 169, 1397–1408 (2017).

38. McCarroll, S. A. et al. A distal regulatory element that controls expression of the mouse beta-globin cluster. *Nature* 431, 95–99 (2004).

39. Kellis, M. et al. Integrated genomic analysis of dynamics, evolution, and regulation at the beta-globin locus. *Science* 314, 1035–1046 (2006).

40. Soler, E., Mikkelsen, T. S. & Walter, W. J. Erythroid-specific transcription factors control disease-associated beta-thalassemia. *Nat. Genet.* 37, 209–213 (1995).

41. Dorflinger, L. et al. The role of the TAL1 transcription factor in mouse embryonic stem cell differentiation. *Eur. J. Haematol.* 55, 229–238 (1995).

42. Hsiung, C. C., Raj, A. & Blobel, G. A. Enhancer transcription factor GATA-1. *Genes Dev.* 18, 1505–1509 (2004).

43. Telenius, J. & Hughes, J. R. NGseqBasic - a single-command UNIX tool for ATAC-seq, DNase-seq, Cut-and-Run, and Chip-seq data mapping.
high-resolution visualisation, and quality control. bioRxiv, https://www.biorxiv.org/content/10.1101/393413v1 (2018).

52. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).

53. Telenius, J. M. et al. CaptureCompendium: a comprehensive toolkit for 3C analysis. bioRxiv, https://www.biorxiv.org/content/10.1101/2020.02.17.952572v1 (2020).

Acknowledgements

This work was supported by an MRC programme grant to DRH (MC_uu_12009). We would like to thank the Cadet laboratory (WIMM, Oxford) for providing wild type human iPSC lines. We thank Martin Larke (WIMM, Oxford) and Julia Truch (WIMM, Oxford) for advice and assistance with computational work. We thank Jacqueline Sloane-Stanley (WIMM, Oxford) for assistance with tissue culture. We thank the MRC WIMM Flow Cytometry Facility and MRC WIMM Sequencing Facility for providing indispensable services.

Author contributions

D.R.H., R.J.G. and Y.K.B. conceived the study and wrote the manuscript. Y.K.B. performed experiments. Y.K.B. and D.J.D. performed data analysis. D.J.D., J.T. and A.M.O. provided analysis tools and reagents. J.R.H., J.C.M. and E.N.O. provided methods and expertise. All authors read and commented on the manuscript.

Competing interests

J.R.H. is a founder and shareholder of Nucleome Therapeutics. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23980-6.

Correspondence and requests for materials should be addressed to R.J.G. or D.R.H.

Peer review information Nature Communications thanks David Bodine and the other, anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© Crown 2021