The chromatin remodeller ATRX facilitates diverse nuclear processes, in a stochastic manner, in both heterochromatin and euchromatin

Julia Truch1, Damien J. Downes1, Caroline Scott1, E. Ravza Gür1,2, Jelena M. Telenius1,2, Emmanouela Repapi2, Ron Schwessinger1,2, Matthew Gosden1, Jill M. Brown1, Stephen Taylor2, Pak Leng Cheong1, Jim R. Hughes1,2, Douglas R. Higgs1✉ & Richard J. Gibbons1✉

The chromatin remodeller ATRX interacts with the histone chaperone DAXX to deposit the histone variant H3.3 at sites of nucleosome turnover. ATRX is known to bind repetitive, heterochromatic regions of the genome including telomeres, ribosomal DNA and pericentric repeats, many of which are putative G-quadruplex forming sequences (PQS). At these sites ATRX plays an ancillary role in a wide range of nuclear processes facilitating replication, chromatin modification and transcription. Here, using an improved protocol for chromatin immunoprecipitation, we show that ATRX also binds active regulatory elements in euchromatin. Mutations in ATRX lead to perturbation of gene expression associated with a reduction in chromatin accessibility, histone modification, transcription factor binding and deposition of H3.3 at the sequences to which it normally binds. In erythroid cells where downregulation of α-globin expression is a hallmark of ATR-X syndrome, perturbation of chromatin accessibility and gene expression occurs in only a subset of cells. The stochastic nature of this process suggests that ATRX acts as a general facilitator of cell specific transcriptional and epigenetic programmes, both in heterochromatin and euchromatin.
The X-linked thalassaemia intellectual disability syndrome (ATRX) protein is a member of the SWI/SNF family of chromatin remodelling factors which acts as an ATP-dependent molecular motor. Together with the histone chaperone DAXX, ATRX is involved in inserting the histone variant H3.3 at sites of nucleosome turnover in a DNA replication-independent manner. The importance of the ATRX/DAXX/H3.3 complex in processing chromatin throughout the genome is highlighted by the diverse nuclear activities that are perturbed when its components are mutated or knocked out. In particular, loss of ATRX affects DNA replication, DNA repair, homologous recombination, genome stability, DNA methylation, chromatin modification and gene expression. Germline mutations in ATRX give rise to a severe form of syndromal intellectual disability (ATR-X syndrome) characteristically associated with downregulation of a-globin expression. In addition, over the past 10 years, acquired mutations in ATRX have been very frequently found in tumours which maintain their telomeres via the alternative lengthening of telomere (ALT) pathway activated in 10–15% of all cancers. Understanding the normal role of ATRX and the mechanisms by which mutations lead to inherited and acquired human disease is consequently of great current interest.

Most previous studies have shown that ATRX is located at a wide range of heterochromatic tandem repeats throughout the genome. These include rDNA, telomeric, pericentric and minisatellite repeats as well as endogenous retropositional sequences and the 3′ exons of the highly duplicated genes of the zinc finger family. Common features of these repeats are that they are frequently transcribed, may form abnormal DNA structures, including G quadruplex (G4) DNA and most are noted as regions that are difficult to replicate. At these sites, ATRX appears to maintain the repressive chromatin marks that characterized such heterochromatic loci.

Previous efforts to determine any role of ATRX in euchromatin have been severely hampered by the quality of chromatin immunoprecipitation (ChIP) experiments. Here, using a much-improved protocol for ChIP-seq we have now comprehensively and definitively analysed the binding of ATRX within euchromatin in different cell types. We confirm that ATRX binds a large proportion of zinc finger genes and also a large subset of enhancers, promoters and gene bodies. At these transcriptionally active elements, there is known to be a high turnover of H3.3 during interphase and we show here that in addition to its role at heterochromatin, the presence or absence of ATRX correlates with the levels of H3.3 at these regions of euchromatin which are normally bound by ATRX. Importantly, by comparing different cell types, we show that ATRX is recruited to regions defined by the process of transcription rather than by any specific subset of transcription factors. We also show that when ATRX is mutated, and where gene expression is perturbed, this is associated with changes in chromatin accessibility, histone modifications, binding of transcription factors and deposition of the histone variant H3.3 at the associated enhancers and/or promoters. Of interest, when analysing the role of ATRX in single cells, we show that in the absence of ATRX changes in chromatin structure at regulatory elements and consequent gene expression occur in a stochastic manner.

Detailed analysis of the role of ATRX at the well-characterized α-globin locus in erythroid cells illustrates how mutations in many widely expressed chromatin remodelling factors may lead to stochastic changes in gene expression and cell fate via a wide range of nuclear processes which they facilitate. In the case of ATRX, this may occur predominantly via interdependent roles in DNA replication, chromatin modification and chromatin accessibility.

Results
ATRX ChIP-seq confirms ATRX association with heterochromatin and reveals a variety of ATRX binding sites in euchromatin. We performed ChIP-seq of endogenous ATRX, with two distinct antibodies, in human lymphoblastoid cell lines (LCLs) derived from three independent unaffected individuals (TA-Ctr, FF-Ctr and CB-Ctr). Using a significantly improved protocol we produced sensitive and reproducible ATRX ChIP-seq datasets associated with low background inputs (Supplementary Fig. 1a). This identified 8,454 ATRX binding sites genome-wide and we selected a subset of ATRX-enriched regions to confirm our results by ChIP-qPCR (Supplementary Fig. 1b).

We confirmed the affinity of ATRX for repetitive regions including pericentromeric DNA, rDNA and GC-rich, low complexity repeat sequences by direct mapping and sequence analysis at peak-called regions. In addition, we identified ATRX enrichment at several genes encoding RNA and the small nuclear RNA U2 (Supplementary Fig. 2a and b). This analysis also showed ATRX associated with the KRAB domain-containing Cys2-His2 zinc fingers (C2-H2 ZNF) (Supplementary Fig. 2c) consistent with observations in other human cell types.

ATRX has been previously shown to bind directly to G4 in vitro and putative quadruplex sequence (POQ) containing targets in vivo and here we observed that almost half of the ATRX binding sites in LCLs overlap POQ (Supplementary Fig. 2d). Peak calling revealed two main patterns of ATRX binding sites including broad shape peaks, similar to those observed with histone marks such as H3K9me3, contrasting with sharper peaks, usually observed with transcription factors (Fig. 1a and ii) possibly reflecting different modes of ATRX recruitment.

With the reduced background afforded by the modified ChIP-seq protocol, it is now evident that, using short-read sequencing, most mappable ATRX targets are present at a subset of genes and many regions of increased chromatin accessibility. Almost two thirds of ATRX binding sites are intragenic with 25% spanning gene promoters, and 39% within gene bodies (Fig. 1b). ATRX binding sites were significantly enriched at transcription start sites (TSS) compared to matching size random fragments (p-value < 2.2e−16, odds ratio = 23.3, Fisher’s exact test) (Supplementary Fig. 3a).

We further analysed the average distribution of ATRX across the entire genome and observed a strong signal enrichment at TSS (Fig. 1c). ATRX binding sites were enriched in annotated CpG islands which were present in 29% (compared to 2% in matching size random fragments, p-value < 2.2e−16, odds ratio = 18.9, Fisher’s exact test) (Fig. 1d and Supplementary Fig. 3b). Using a BioCAP-seq approach, which detects non-methylated CpG islands, we found that most of the ATRX bound CpG islands are unmethylated and associated with promoter regions (Fig. 1d, e and Supplementary Fig. 4). At these sites, ATRX peaks were enriched in the active chromatin mark H3K4me3 and depleted of H3K9me3 signal (Fig. 1e and f). Only 113 ATRX binding sites were fully methylated on both alleles (p-value < 2.2e−16, odds ratio = 0.1, Fisher’s Exact Test, compared to matching size random fragments) (Fig. 1d and Supplementary Fig. 3c); such sites were correlated with high levels of H3K9me3 (Fig. 1f) representative of heterochromatic regions to which ATRX is known to be recruited.

Given the strong association, we found here between ATRX and euchromatin, we investigated chromatin accessibility at ATRX binding sites. Using ATAC-seq we found that 87% of ATRX binding sites associate with regions of open chromatin, in line with ATRX enrichment at transcriptionally active chromatin regions (Fig. 1e, g). Comparison with matching size random fragments showed that ATRX was significantly enriched at
regions of open chromatin ($p$-value $< 2.2 \times 10^{-16}$, odds ratio = 108.9, Fisher’s exact test) (Supplementary Fig. 3d).

**ATRX binding sites** are associated with active regulatory elements. To determine the nature of the ATRX binding sites, the data were annotated with Genomic STate ANnotation (GenoSTAN)\textsuperscript{20}. Exploring 77,597 sites encompassing the open chromatin regions identified by ATAC-seq and ATRX binding sites, we identified seven chromatin states (active promoter (P), poised Promoter (Pp), active enhancers (E), enhancer-CTCF binding site (EC), CTCF site (C), repressed region (R) and background (B)) using this unsupervised approach (Fig. 2a and Supplementary Fig. 5). Of these, we observed a significant
enrichment of ATRX binding sites at active enhancers (41% vs 22% of all sites, \( p \)-value < 2.2e\(^{-16} \), odds ratio = 2.5, Fisher's exact test) and promoters (32% vs 16% of all sites, \( p \)-value < 2.2e\(^{-16} \), odds ratio = 2.4, Fisher's exact test) (Fig. 2a and Supplementary Fig. 5c). These annotations were supported by analysing the chromatin state segmentation by Hidden Markov Model (ChromHMM) from an ENCODE LCLs dataset\(^{21} \) (Supplementary Fig. 6a). Consistent with most ATRX binding at active regulatory elements, 95% of the ATRX binding sites contained at least one transcription factor binding site (TFBS) (Fig. 2b) (\( p \)-value < 2.2e\(^{-16} \), odds ratio = 141, Fisher's exact test compared to matching size random fragments) (Supplementary Fig. 3e) with an over representation of RUNX3 in the LCLs analysed (present in 85% of ATRX binding sites) (Fig. 2c). Motif analyses supported an enrichment of RUNX family motifs at ATRX binding sites in regions of open chromatin in LCLs (Fig. 2d and Supplementary Fig. 6b), further highlighted with the metagenomic analysis (Fig. 1e).

Chromosome conformation capture (Capture-C) which identifies interactions between enhancers and promoters, was used to confirm the identity of gene-specific enhancers. We confirmed that ATRX signals at interacting promoters and enhancers were not due to artefacts caused by the binding of proteins at one site brought into close proximity and crosslinked to another site.

enrichment of ATRX binding sites at active enhancers (41% vs 22% of all sites, \( p \)-value < 2.2e\(^{-16} \), odds ratio = 2.5, Fisher's exact test) and promoters (32% vs 16% of all sites, \( p \)-value < 2.2e\(^{-16} \), odds ratio = 2.4, Fisher's exact test) (Fig. 2a and Supplementary Fig. 5c). These annotations were supported by analysing the chromatin state segmentation by Hidden Markov Model (ChromHMM) from an ENCODE LCLs dataset\(^{21} \) (Supplementary Fig. 6a). Consistent with most ATRX binding at active regulatory elements, 95% of the ATRX binding sites contained at least one transcription factor binding site (TFBS) (Fig. 2b) (\( p \)-value < 2.2e\(^{-16} \), odds ratio = 141, Fisher's exact test compared to matching size random fragments) (Supplementary Fig. 3e) with an over representation of RUNX3 in the LCLs analysed (present in 85% of ATRX binding sites) (Fig. 2c). Motif analyses supported an enrichment of RUNX family motifs at ATRX binding sites in regions of open chromatin in LCLs (Fig. 2d and Supplementary Fig. 6b), further highlighted with the metagenomic analysis (Fig. 1e).
ChromHMM performed in LCLs (Supplementary Fig. 6d). As an example of enhancer-promoter interactions, we used the MYC promoter as viewpoint, and highlighted a complex pattern of enhancer-promoter interactions including those in which the regulatory elements are enriched in ATRX (Supplementary Fig. 6e). These sites were also enriched in active chromatin marks including open chromatin regions identified by ATAC-seq, and histone modifications such as H3K27ac or H3K4me1. Notably, these ATRX binding sites also displayed an enrichment in H3.3 ATRX3. Previous studies have identified enriched across this locus in LCLs (Fig. 3d). These genes are transcribed. In contrast, H3K27me3, a repressive mark on the α-globin gene cluster. Here we found that ATRX is also enriched at the TSS of the following downregulated genes in ATR-X cases: the activating transcription factor 3 (ATF3) and the α-globin silencing was depleted in erythroid cells, and fourth, genes are repressed in both cell types. As shown previously, ATRX enrichment at TSS is associated with transcription (Fig. 3i). As previous work has linked the recruitment of ATRX to transcription122, the relationship between these factors was further examined by comparing ATRX enrichment and gene expression across both cell types. ATRX signals in LCLs were visualised according to gene expression in four classes. First, genes are expressed in both LCLs and erythroid cells; second, genes only expressed in LCLs; third, genes only expressed in erythroid cells, and fourth, genes are repressed in both cell types. As shown previously, ATRX enrichment at TSS was correlated with the level of gene expression in LCLs (Fig. 3j). Our results thus show that enrichment of ATRX at gene regulatory regions correlates not only on their chromatin status but also on the level of gene expression.

ATRX binds to regulatory elements in a tissue-specific manner. To determine the distribution of ATRX in a different cell type and compare this with the distribution in LCLs, ATRX ChIP-seq was performed in erythroblasts using differentiated CD34+ human stem and progenitor cells (HSPCs) from normal human donors (Supplementary Fig. 7). Using our improved protocol and stringent peak calling, 12,659 ATRX binding sites were detected, a number more than ten times higher than previously described in erythroid cells3. As observed in LCLs, ATRX binding sites were mainly intragenic with a significant enrichment at TSS compared to matching size random fragments (p-value < 2.2e−16, odds ratio = 57.8, Fisher’s exact test) and overlapping with open chromatin in erythroblasts (p-value < 2.2e−16, odds ratio = 138.5, Fisher’s exact test) (Fig. 3a and b and Supplementary Fig. 3f and g). However, in erythroid cells, motif analysis showed that ATRX binding sites were enriched in GATA motifs in contrast to RUNX motifs in LCLs (Fig. 3c). ChIP-seq for histone marks and ATAC-seq showed that ATRX binds to active regulatory elements, confirming the similar observations in LCLs (Supplementary Fig. 8).

a-globin is a well-characterized, erythroid-specific target of ATRX3. Previous studies have identified ATRX predominantly at the G-rich ψC variable number tandem repeat (ψC VNTR) within the a-globin gene cluster. Here we found that ATRX is also significantly enriched at the CpG islands associated with the HBA2 and HBA1 genes and at the enhancers (R1, R2 and R4 in Fig. 3d). All of these sites showed the epigenetic features associated with active chromatin including chromatin accessibility, and enrichment in H3K4me3, H3K4me1 and/or H3K27ac (Fig. 3d). Furthermore, H3.3 enrichment was also observed at ATRX binding sites at HBA1, HBA2 and their major enhancers R1 and R2 and ATRX enrichment in this region was only seen in erythroid cells when the elements are active and the elements and genes are transcribed. In contrast, H3K27me3, a repressive mark associated with a-globin silencing was depleted in erythroid cells but enriched across this locus in LCLs (Fig. 3d). These observations at the a-globin cluster confirm that ATRX interacts with active regulatory elements specifically in cells in which these genes are transcribed and suggest that its binding depends on the transcriptional activity of such sequences.

ATRX binding is associated with changes in chromatin accessibility and gene expression. We further investigated the characteristics of ATRX binding sites shared in both erythroid and non-erythroid (LCLs) cell types. We identified 2251 ATRX binding sites in LCLs (27%) overlapping with ATRX peaks in erythroblasts (Fig. 3e). As expected, shared peaks were enriched for gene promoters, CpG islands, open chromatin and PQS (p-value < 2.2e−16, odds ratio = 73.9, p-value < 2.2e−16, odds ratio = 83.9, p-value < 2.2e−16, odds ratio = 50.7 and p-value < 2.2e−16, odds ratio = 7.2, respectively, Fisher’s exact test, compared to matching size random fragments) (Fig. 3f–h and Supplementary Fig. 3h–k), supporting the hypothesis that ATRX binding/recruitment occurs in such regions of the euchromatic genome. Almost 95% of ATRX binding sites found in both LCLs and erythroblasts showed similar chromatin accessibility in both cell types (Fig. 3g). Interestingly, most of the open chromatin sites were found in both LCLs and erythroblasts contrasting with the matching size random fragments in which the majority of the open chromatin regions were cell type specific (p-value < 2.2e−16, odds ratio = 33.1, Fisher’s exact test) (Supplementary Fig. 3j). As previous work has linked the recruitment of ATRX to transcription122, the relationship between these factors was further examined by comparing ATRX enrichment and gene expression across both cell types. ATRX signals in LCLs were visualised according to gene expression in four classes. First, genes are expressed in both LCLs and erythroid cells; second, genes only expressed in LCLs; third, genes only expressed in erythroid cells, and fourth, genes are repressed in both cell types. As shown previously, ATRX enrichment at TSS was correlated with the level of gene expression in LCLs (Fig. 3j). Our results thus show that enrichment of ATRX at gene regulatory regions correlates not only on their chromatin status but also on the level of gene expression.

Pathogenic ATRX mutations are associated with changes in the chromatin environment at active regulatory elements. To date, only a small number of genes have been identified that are affected by pathogenic ATRX mutations in humans. We performed microarray experiments on 20 LCLs from normal individuals and 28 LCLs from individuals with diverse pathogenic ATRX mutations, to identify more genes regulated by ATRX. When possible, we selected control-case pairs composed of ATR-X cases (all males) and their unaffected fathers thus reducing background effects (Supplementary Fig. 9). In total, 388 probes displayed an adjusted p-value < 0.05 (red dots in Fig. 4a), representing 234 differentially expressed genes (DEGs). The previous work3 had shown that NME4 (a nucleoside diphosphate kinase) was downregulated in LCLs derived from ATR-X cases and that this was correlated with the size of an intragenic G-rich tandem repeat which is a PQS. This downregulation was replicated in the current study (logFC = −1.34 compared to control) (Fig. 4a) and ATR-X cases displayed widely spread signals compared to controls (Supplementary Fig. 10a). Microarray data were further validated by qRT-PCR on 20/21 candidate genes including 17/18 DEGs (Supplementary Fig. 10b). Gene ontology analyses revealed a significant enrichment in biological processes linked to transcription (Table 1). Expression data were cross-referenced with the ATRX ChIP-seq data to analyse the distribution of the ATRX binding sites in the environment of the DEGs. Considering the complete set of DEGs and ATRX binding sites, we found a total of 83 DEGs containing at least one intragenic ATRX binding site and a total of 98 DEGs (10 kb window), 128 DEGs (50 kb window) and 156 DEGs (100 kb window) with at least one ATRX binding sites within these windows (p-value = 3.45e−06, odds ratio = 1.95, p-value = 1.51e−05, odds ratio = 1.81, p-value = 0.009595, odds ratio = 1.41, p-value = 0.01836, odds ratio = 1.39, respectively, Fisher’s exact test) that could include genes directly affected by ATRX mutations (Supplementary Fig. 10c).

By way of example and supporting a role in the maintenance of chromatin integrity at regulatory elements, we observed that ATRX is enriched at the TSS of the following downregulated genes in ATR-X cases: the activating transcription factor 3 (ATF3) and the α-globin gene cluster. Here we found that ATRX is also enriched at the TSS of the following downregulated genes in ATR-X cases: the activating transcription factor 3 (ATF3) and the α-globin gene cluster. Here we found that ATRX is also enriched at the TSS of the following downregulated genes in ATR-X cases: the activating transcription factor 3 (ATF3) and the α-globin...
and three ZNFs (ZNF555, ZNF57 and ZNF718). Their promoter regions were also characterised by the presence of CpG islands and PQS (Fig. 4b, Supplementary Fig. 11a and b). At these loci, ATRX deficiency was associated with a change in the chromatin environment including a reduction in chromatin accessibility; a reduction in active chromatin marks including H3K4me3 and H3K27ac; reduced RUNX3 binding; a reduction in H3.3 at the promoter and decreased H3K36me3, a mark associated with transcriptional elongation.

In addition, we observed hypermethylation at the ATF7IP2 and ZNF718 CpG island promoters in LCLs from ATR-X cases supporting previous data performed in peripheral blood cells from ATR-X cases23 (Fig. 4b, Supplementary Figs. 11b and 12a). By contrast, the methylation status of the ZNF555 and ZNF57
CpG islands remained unaffected in ATR-X cases (Supplementary Figs. 11a and 12b) suggesting that DNA methylation is not the primary event affecting the chromatin environment in these cases.

Perturbation of chromatin was also observed at putative enhancers as illustrated at the PBX4 locus (Fig. 4c). This gene, encoding a putative transcription factor, was downregulated in ATRX deficient cells and contained an ATRX binding site within.

Fig. 3 ATRX enrichment at regulatory elements varies with the activity of the locus and ATRX binding sites conserved across cell lines show conserved chromatin accessibility states. a Distribution of the ATRX binding sites relative to genes in erythroblasts. b Distribution of the ATRX binding sites depending on their chromatin accessibility as assessed by ATAC-seq in erythroblasts. c Motif analysis of ATRX binding sites in erythroblasts (p-values HOMER findMotifsGenome.pl)58. d Representative image of the α-globin locus active in erythroblasts and silenced in LCLs. The signals represent an average of the independent replicates (in LCLS, n = 3 for ATRX ChIP-seq, H3K4me1, H3K4me3, H3K27me3 and H3.3 ChIP-seq, n = 2 for H3K27ac ChIP-seq and n = 4 for ATAC-seq. In erythroblasts, n = 3 for ATRX-Chip-seq, n = 1 for H3K4me1, H3K4me3, H3.3 and H3K27me3 and n = 4 for ATAC-seq). e Venn diagram showing the cell type-specific and conserved ATRX binding sites in LCLs and erythroblasts. f Distribution of the conserved ATRX binding sites based on their position in relation to genes. g Chromatin accessibility status in LCLs and erythroblasts at the conserved ATRX binding sites. h Distribution of the conserved ATRX binding sites based on the presence or absence of PQS. i Genome-wide ATRX enrichment across genes in LCLs depending on the gene expression status in LCLs and erythroblasts. The signals represent an average of the independent replicates (n = 3). j Genome-wide ATRX enrichment across genes depending on the level of gene expression in LCLs. The signals represent an average of the independent replicates (n = 3).

Fig. 4 Pathogenic ATRX mutations are associated with changes in chromatin environment at regulatory elements in LCLs. a Volcano plot of the microarray data comparing LCLs derived from ATR-X cases and unaffected donors. In red (or blue if highlighted with gene name), the probes that are significantly differentially expressed (ATR-X cases relative to controls) with an adjusted p-value ≤ 0.05 (horizontal grey dot line). adj.P.Val = adjusted p-values based on lmFit and eBayes (empirical Bayes moderated t-statistics test) from the limma packages39 and adjustment with Benjamini Hochberg. Adjusted p-values for probes highlighted in blue: ATF7IP2 (adj.P.Val = 9.14E−06 and 3.19E−06), LPAR6 (adj.P.Val = 1.37E−03), NME4 (adj.P.Val = 3.67E−02), PBX4 (adj.P.Val = 2.96E−02), RASGEF1A (adj.P.Val = 2.22E−08), ZNF555 (adj.P.Val = 4.65E−03), ZNF595 (adj.P.Val = 2.36E−04 and 1.05E−03) and ZNF718 (adj.P.Val = 1.08E−04). logFC = log fold change of cases over controls (n = 20 for the controls and n = 28 for the cases). b ATF7IP2 locus. c PBX4 locus. Signals of chromatin marks, Bio-CAP, ATAC, and Runx3, for ATR-X individuals (Case) and controls (Ctr). The signals represent an average of the independent replicates (n = 3 except for ATAC (n = 4) and H3K27ac and runx3 (n = 2) and ATRX (n = 6)).
Table 1. Gene ontology analysis based on the microarray data analysis from LCLs, ATR-X cases vs unaffected donors showing the top enriched pathways based on the count with a p-value < 0.05 (p-values based on EASE Score DAVID 6.8).

| Term                                      | Count | p-value | Genes                                                                 |
|--------------------------------------------|-------|---------|----------------------------------------------------------------------|
| Transcription, DNA-templated               | 32    | 0.044   | ZNF55, ZNF57, ZNF347, CNOT6, ZNF32, ZBTB38, ZNF30, HOXA1, L3MBTL4,   |
|                                           |       |         | HOXAI, IPRN, SATX, LBH, TEAD4, VP36, ZNF36, ARID5A, ZNF649, MAPK11,  |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
|                                           |       |         | ZNF32, ZNF30, HOXA1, L3MBTL4, TRNP1, ZNF57, ADNP2, TFAP2B, MBD1,    |
|                                           |       |         | ZNF595, ZBTB38, MBD2, CTR9, AP1F2, ARID5A, ZNF32, ZNF57, CNOT6,    |
|                                           |       |         | LMBTL1, TFAP2B, MBD2, ZNF62, MBD1, HD, ADNP, ZNF506, MBD2, MBD1,    |
its gene body. In control cells, these ATRX binding sites showed marks of active enhancers (open chromatin, RUNX3, H3K4me1 and H3K27ac) and enrichment in H3.3 (Fig. 4c). Capture C confirmed a specific interaction between this site and the PBX4 promoter, which is consistent with a role as an enhancer of PBX4 (Supplementary Fig. 13). Strikingly, ATRX deficient cells displayed a decrease in chromatin accessibility, active chromatin marks (H3K27ac and H3K4me1) and H3.3 signals and a marked diminution in RUNX3 binding at this putative enhancer. Together, these results show that ATRX plays a role in initiating and/or maintaining chromatin accessibility and transcription factor occupancy at a subset of regulatory elements.

**Altered chromatin accessibility associated with a reduced α-globin expression in ATR-X syndrome.** By contrast with the common types of α-thalassaemia, in individuals with ATR-X syndrome and α-thalassaemia, the α-globin cluster is structurally intact. In these individuals downregulation of genes in the α-globin cluster (HBA1/2 and HBM) results from a deficiency in ATRX. The previous work³ showed that the length of the adjacent G-rich VNTR accounts for almost 60% of the variance in the severity of the α-thalassaemia as reflected in the number of red cells exhibiting HbH (β₄) inclusions. The mechanism by which the interaction between ATRX/DAXX/H3.3 and these upstream PQS repeats affect α-globin gene expression is unknown. However, this alone is insufficient to fully explain the phenotype, suggesting that a deficiency in ATRX downregulates α-globin expression via more than one of its various nuclear activities. To investigate this further, we isolated CD34⁺ HSPCs from two brothers diagnosed with the ATR-X syndrome (case 1 and case 2). These individuals had the same ATRX mutation but only one of them (case 1) had a detectable α-thalassaemia phenotype with 3% of red cells exhibiting HbH (β₄) inclusions.

CD34⁺ HSPCs cells from these ATR-X individuals differentiated along the erythroid pathway in a similar manner to normal controls24. Due to the limited amount of material from these ATR-X cases, we prioritized the analysis of chromatin accessibility and assessment of H3K4me1, the most perturbed epigenetic features in ATR-X X LCLs. In the affected individuals, we observed a depletion of H3K27ac and chromatin accessibility at HBM, the HBA1/2 loci and the R1 enhancer (Fig. 5a). In addition, we showed that binding of the erythroid-specific transcription factor GATA1 was reduced at the ATRX bound R1 enhancer in the affected individuals. Together, these findings confirmed that ATRX deficiency was associated with changes in chromatin accessibility at active regulatory elements in this model locus (Fig. 5a). Capture C with the HBA promoters as a viewpoint did not show any drastic differences between affected individuals and normal controls suggesting that ATRX deficiency does not affect gene expression via significant changes in the 3D organisation of the locus (Supplementary Fig. 14). Consequently, it appears that changes in chromatin accessibility and associated chromatin modifications of regulatory elements within the α-globin locus are likely to contribute to the α-thalassaemia phenotype.

**Analysing the mechanism by which ATRX regulates chromatin and gene expression in single cells.** Although deficiency of ATRX causes α-thalassaemia, the red cell phenotype is quite different to that seen in the common forms of this anaemia. In particular, in ATR-X syndrome, most red cells have a normal haemoglobin content and volume (mean corpuscular haemoglobin and mean corpuscular volume) whereas both parameters are uniformly reduced in all cells in the common forms of α-thalassaemia. Furthermore, we have found that individuals (such as case 1 and case 2) with exactly the same mutation in the ATRX gene have variable degrees of α-thalassaemia. This suggests that α-globin gene expression may not be affected in the majority of erythroid cells in ATR-X syndrome and that the presence of α-thalassaemia may be variable even with the same underlying mutation.

To determine if there might be cellular heterogeneity associated with changes in chromatin accessibility we performed single-cell ATAC-seq (scATAC-seq) in erythroblasts. As in bulk ATAC-seq, in a composite of the single-cell signals we observed a reduction in chromatin accessibility across the HBA1/2 loci and at the R1 enhancer in the ATR-X cases case1 and case2 (Fig. 5a and Supplementary Fig. 15a).

At the single-cell level, the chromatin accessibility in erythroid cells from ATR-X cases was not reduced at the HBB locus compared to controls (p-value = 0.9995, Wilcoxon tests), whereas it was at the HBA and HBM loci (p-value < 2.2e−16, Wilcoxon tests) (Fig. 5b–d). β₄ inclusions result from an excess of HBB relative to HBA expression. We therefore assess the chromatin accessibility at the HBA or HBM loci relative to HBB in each cell in affected individuals and normal controls via a t-SNE analysis. The results revealed that most of the cells clustered in a major population (consistent with our differentiation protocol (Supplementary Fig. 7)) and identified a subpopulation of cells which displayed low accessibility at the HBA and HBM loci despite their high chromatin accessibility at the HBB locus (encircled with a red line in Fig. 5e–h). This subpopulation was predominantly composed of cells from ATR-X cases (p-value < 2.2e−16, odds ratio = 1.6, Fisher’s exact test) most of which were derived from the ATR-X case 1 with α-thalassaemia (p-value < 2.2e−16, odds ratio = 2.3, Fisher’s exact test)). This contrasts with subpopulations where there is an association between the high levels of chromatin accessibility at all three loci (HBA, HBM and HBB) in both ATR-X cases and controls (pointed to by arrows in Fig. 5e,g, Supplementary Fig. 15b). The subpopulations which displayed high chromatin accessibility at the HBB locus also displayed high chromatin accessibility at the SLC4A1 locus (also known as Band3 or CD233), a marker which is upregulated in late erythroblasts (Supplementary Fig. 15c).

Subsequently, we performed single-cell RNA-seq (scRNA-seq) on cells from the ATR-X case with α-thalassaemia (case1) and one control sample (Ctr3), and analysed the expression of HBA, HBM and HBB in single cells from both samples. Similarly to the scATAC-seq analysis, we observed a large population containing most of the cells (Supplementary Fig. 16a–d). Cells were negative for CD34 (a marker of human stem and progenitors) and positive for GPA (a marker expressed during terminal erythroid differentiation, such as at day 10 of differentiation) (Supplementary Fig. 16a). By highlighting only the cells which have a signal above quantile 5, we observed that the cells at the left side of the main population (black arrow in Supplementary Fig. 16b) tended to be slightly less differentiated. By contrast, the cells on the right part of the main population (red arrow in Supplementary Fig. 16b) tend to be more differentiated based on the expression of GPA, SLC4A1 and HBB. In agreement with a previous report focusing on a homogenous population of cells, we observed that the cells clustered mainly based on their cell cycle phase rather than cell identity (Supplementary Fig. 16c). These results identified the minor population as slightly more differentiated G1 phase cells (Supplementary Fig. 16a–c). Splitting the cells between control and case, highlighted the presence of cells from the ATR-X case with a high expression signal for HBB but a reduce signal for HBA and HBM (pointed out by a red arrow in Supplementary Fig. 16d).

Using a scatter plot approach allowed to deeper assess the correlation between the expression of HBB and HBA/HBM. In the control, HBA and HBB expression levels were proportional,
maintaining a constant HBA/HBB ratio between different cells throughout differentiation (Fig. 5i). Equivalent levels of HBA and HBB were also observed in partially differentiated ATR-X cells from case1 with low to medium HBB expression (i.e. relative expression of HBB below 8 in Fig. 5i). Conversely, in the most differentiated cells with a high level of HBB (relative expression above 8), we detected a subpopulation of ATR-X cells with reduced HBA expression in case 1 (Fig. 5i). The reduced expression of HBA was supported by RNA-FISH data performed on erythroblasts from case 1 at day 10 and day 13 with the lower expression of HBA being most apparent at the later stages of differentiation (Supplementary Fig. 16e). Interestingly, ATR-X
Fig. 5 ATRX loss of function is associated with perturbation of the chromatin environment and gene expression in a subpopulation of erythroblasts.

a Comparison of the chromatin environment in erythroblasts across the α-globin locus between unaffected donors and ATR-X cases (n = 1). H3K27ac ChIP-seq were performed as ChIP-Rx and normalised based on the Drosophila melanogaster S2 cells spiked in. b–d Tukey based box plots of scATAC-seq data showing the distribution of chromatin accessibility in controls (n = 2: Ctr2 and Ctr3, each of them composed of 4000 cells) and ATR-X cases (n = 2: case1 and case2, each of them composed of 4000 cells). Tukey based box plots showing the 25th and 75th percentiles (lower and upper bounds of the box, respectively), the median (centre line highlighted by an arrow), the minimum value lower than the 25th percentile minus 1.5*IQR (lower whisker) and the maximum value greater than the 75th percentile plus 1.5*IQR (upper whisker), any values beyond the whiskers boundaries are represented as dots. e–h t-SNE analysis of scATAC-seq data showing each individual cell (n = 4: Ctr2, Ctr3, Case1 and Case2) and encircled in red a subpopulation with contrasting chromatin accessibility scores for e HBB, f HBA1, g HBM; in h highlighting the cells belonging to either controls or cases that are included in or surrounding the subpopulation encircled with a red line, in i–g the black arrows point to examples of subpopulations where there are high levels of chromatin accessibility at HBB, HBA and HBM loci, i–j scRNA-seq data showing each individual cell colour based on HBB expression in control (n = 1: Ctr3) and in case (n = 1: Case1) and showing in i the relative gene expression of HBA relative to HBB, and j the relative gene expression of HBM relative to HBB. k Model of the loss of chromatin integrity observed in ATR-X cases during cell differentiation/locus activity and the associated effect on gene expression.

cells with a relatively high level of HBB expression also showed an even more pronounced reduction in expression of the HBM gene (Fig. 5k) which lies close to the G-rich ψζ VNTR (Fig. 5a).

Taken together, our results suggest that when ATRX is mutated, changes in chromatin accessibility and gene expression predominantly affect a subset of late differentiating erythroid cells rather than causing a subtle change in the whole population (Fig. 5k). This provides an in vitro model recapitulating the α-thalassaemia associated with ATRX deficiency, and using this, we show that ATRX deficiency at α-globin and elsewhere in the genome, disrupts the epigenetic environment of active regulatory elements correlating with an impairment in gene expression.

Discussion

It has previously been shown that the ATRX/DAXX/H3.3 complex binds repetitive sequences and putative G4 quadruplex forming sequences (PQS) in heterochromatin, where it plays an important ancillary role in multiple nuclear processes including replication, DNA repair, recombination, and transcription. Deposition of H3.3 in mammalian euchromatin has been generally considered to depend on the HIRA complex although deposition of H3.3 in euchromatin27,28.

Here using a substantially improved ChIP-seq protocol we have comprehensively and definitively shown that ATRX plays a role at multiple sites in euchromatic regions, in particular at enhancers and promoters including those located within the context of CpG islands. Importantly, many of these sequences are G-rich regions with the potential to form G4 quadruplex sequences, so called PQSs. The presence of ATRX at its binding sites in euchromatin is correlated with levels of histone H3.3, indicating that in addition to the HIRA complex, the ATRX/DAXX/H3.3 complex also plays a role in the deposition of H3.3 within euchromatic regions of the genome, in particular at active regulatory elements.

Analysis of the patterns of ATRX enrichment at regulatory elements in different cell types shows that recruitment of ATRX occurs in a tissue-specific rather than a gene-specific or transcription factor-specific manner. Furthermore, we show here that the level of ATRX enrichment is correlated with the level of gene expression. These findings are consistent with our previous observations that ATRX is preferentially recruited to actively transcribed regions22. In cells with loss of function mutations in ATRX, gene expression dysregulation is associated with variations in chromatin accessibility, active chromatin marks and H3.3 deposition at actively transcribed genes and their regulatory elements. At some loci, these changes predominantly affect promoters whereas at other loci enhancers were affected and, at both

promoters and enhancers, transcription factor binding is also reduced. Therefore it is possible that ATRX is required to maintain chromatin in an accessible state to facilitate a variety of nuclear processes during interphase by maintaining the presence or the positioning of the labile histone variant H3.3 at active genes and regulatory elements29–31.

Methylation at CpG islands has previously been reported in blood cells from individuals with ATR-X syndrome.23 We observe this at a subset of downregulated genes and it seems most likely that this CpG island methylation is secondary to other epigenetic changes rather than methylation changes dictating the epigenetic landscape at these loci. It has previously been shown that TF binding is necessary to prevent methylation of a CpG island promoter32 and it is possible that at loci which become methylated when ATRX is mutated, chromatin remodelling by ATRX is necessary to facilitate TF binding33 and keep these regions free of methylation.

α-thalassaemia is a hallmark of ATR-X syndrome and is due to the reduced expression of α-globin in erythroid cells. This is associated with a subsequent imbalance in globin chain synthesis and the production of excess β-globin chains leading to the formation of β-tetramers (HbH: β4). This well-characterised locus provides an ideal model for understanding how ATRX normally plays a role in regulating gene expression. Here we have shown that in cells with a deleterious mutation in ATRX, chromatin accessibility at the α-globin cluster is reduced in only a subset of cells at intermediate and late stages of differentiation when the globin loci are most active. In single cells from this subgroup in which chromatin accessibility of α-globin is low, the accessibility of the β-globin gene is relatively high. This same subset of cells has an even greater reduction in the expression of the α-like gene HBM. These observations do not exclude the possibility that underlying defects associated with ATRX deficiency are present at earlier stages of differentiation but are below the level of detection and accumulate with rounds of cell division and differentiation.

Given the role of the ATRX/DAXX/H3.3 complex in facilitating a diverse range of nuclear processes, it is likely that any associated changes in gene expression due to the downregulation of components of this complex will result from different and even multiple effects. To address this we have used the human α-globin gene cluster as an experimental model.

Previously we have shown that the severity of α-thalassaemia in ATR-X syndrome is determined, to a large extent, (accounting for ~60% of the variance) by the length of a G-rich variable number tandem repeat (ψζ) lying upstream of the α-globin cluster to which ATRX binds. This G-rich sequence can form G4 structures in vitro and it has been speculated that the longer the repeat the more likely G4 is to form. ATRX binds to G4 in vitro and in the absence of ATRX, G4 structures are more likely to form.
RNA extraction. RNA was extracted using RNeasy Mini kit from Qiagen. 4 × 10^6 cells were used as starting material. lysates were homogenised using QIAshredder spin column (79658, Qiagen). Samples were submitted to on-column DNA digestion using RNAase-Free set DNase Set (79254, Qiagen) and RNA quality assessed by NanoDrop and integrity was assessed using Agilent RNA ScreenTape assay system (3067-5576/7, Agilent). Only samples displaying a RINe score higher than 8 were selected. Note that RNA extractions for the microarray experiments were carried out in batches of 7–12 cell-well plates and this batch effect for further analysis. All the samples present in the same batch were processed to the cell from the microarray experiment.

Reverse transcription. cDNA was produced using High Capacity RNA-to-cDNA Kit (4387406, Applied Biosystems) using 1.5 μg of RNA as starting material and following the manufacturer’s instructions.

Primeview affymetrix microarray. The microarray experiment was performed using PrimeView™ Human Gene Expression Array Plate and following the 100 or 81/4-format 3′IVT PLUS Reagent Kit User Manual (Affymetrix). 250 ng of RNA were used as the starting material. In short, RNA samples were used for cDNA synthesis followed by the production of biotinylated-labelled cRNA. Labelled cRNA was then purified, fragmented and hybridised on arrays. Quality controls were performed during the process using Agilent Technologies Agilent RNA ScreenTape System. After staining, the arrays were scanned. Results were further analysed using Affymetrix® Transcriptome Analysis Console (TAC) and Affymetrix® Expression Console® Softwares (Affymetrix).

Power calculation and correction for the batch effects in microarray experiments analyses. The power calculation was determined using the same and size packages from R. In addition to the microarray analysis as per the manufacturer’s instruction (described above), a parallel analysis was performed allowing the correction of the batch effect generated during RNA extraction using R script with limma39, affy40 and genefilter41 packages. To increase the power, half of the probes were removed after normalisation based on their very weak signal or weak variation across samples42.

High-throughput qRT-PCR by Fluidigm Biomark HD system. High-throughput qRTPCR was performed with preliminary extracted RNA using a Fluidigm 192-24 Gene Expression Chip and Taqman assay (ThermoFisher Scientific) listed in Supplementary Table 2. Data were analysed using the Fluidigm Real-time PCR analysis 3.1.3 software and were normalised to the mean of two housekeeping genes (RPL13A and GAPDH).

Quantitative-PCR. qPCR based on the Taqman assay was performed using Taqman probes with a 2x Taq mastermix (Applied Biosystems) and Taqman expression assay (ThermoFisher Scientific) listed in Supplementary Table 2. qPCR experiments based on SYBR green assay were performed using SYBR green master mix (4309155, Applied Biosystems). The primers used for these experiments are listed in Supplementary Table 3.

Cell culture. LCLs derived from ATR-X cases and healthy controls were stored in liquid nitrogen in 10% Dimethyl-Sulphoxide (DMSO). Upon recovery, LCLs were thawed and washed with RPMI medium supplemented with 15% FCS, 1% Pen-Strep and 1% glutamine. Then, LCLs were incubated with supplemented RPMI media at 37 °C and 5% CO2 until reaching the desired number which depended on the experiment.

CD34⁺ HSPCs were isolated from either 50 mL of healthy adult peripheral blood or leucocyte reduction filter (obtained from the NHSBT) using Histopaque-1077 Hybri-Max (Sigma) density centrifugation and positive bead selection. No change in enhancer-promoter interaction was detected in ATR-X erythroblasts but rather there was a clear reduction in chromatin accessibility, H3K27acetylation and GATA1 binding at R1, one of the major enhancers in the globin cluster. In future studies to determine the relative contribution of ATRX at the α-globin regulatory elements could be assessed by excising the G-rich tandem repeat in erythroblasts.

The general observations that when ATRX has downregulated the chromatin at sites normally bound by ATRX becomes less accessible, less enriched in H3.3, less bound by their cognate transcription factors and less transcribed, together with the finding that ~50% of these sites are putative G4 quadruplexes, suggests that the mechanisms underlying downregulation of the α-genes may also be relevant to many other targets in euchromatin. These mechanisms together or independently may therefore explain many of the general effects of ATRX deficiency on gene expression at other loci.

Methods

Ethical compliance. Use of blood cells from human participants was approved under REC reference: 07/MRE60/70 by Scotland A Research Ethics Committee. Informed consent was obtained for all participants. The study was authorized by the UK Medical Research Council, and the study design and conduct complied with all relevant regulations regarding the use of human study participants and was conducted in accordance to the criteria set by the Declaration of Helsinki.

Cell culture. LCLs derived from ATR-X cases and healthy controls were stored in liquid nitrogen in 10% Dimethyl-Sulphoxide (DMSO). Upon recovery, LCLs were thawed and washed with RPMI medium supplemented with 15% FCS, 1% Pen-Strep and 1% glutamine. Then, LCLs were incubated with supplemented RPMI media at 37 °C and 5% CO2 until reaching the desired number which depended on the experiment.

CD34⁺ HSPCs were isolated from either 50 mL of healthy adult peripheral blood or leucocyte reduction filter (obtained from the NHSBT) using Histopaque-1077 Hybri-Max (Sigma) density centrifugation and positive bead selection. No change in enhancer-promoter interaction was detected in ATR-X erythroblasts but rather there was a clear reduction in chromatin accessibility, H3K27acetylation and GATA1 binding at R1, one of the major enhancers in the globin cluster. In future studies to determine the relative contribution of ATRX at the α-globin regulatory elements could be assessed by excising the G-rich tandem repeat in erythroblasts.

The general observations that when ATRX has downregulated the chromatin at sites normally bound by ATRX becomes less accessible, less enriched in H3.3, less bound by their cognate transcription factors and less transcribed, together with the finding that ~50% of these sites are putative G4 quadruplexes, suggests that the mechanisms underlying downregulation of the α-genes may also be relevant to many other targets in euchromatin. These mechanisms together or independently may therefore explain many of the general effects of ATRX deficiency on gene expression at other loci.
lysis buffer 1 (100 mM HEPES, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 and 0.25% Triton X-100), then with lysis buffer 2 (200 mM NaCl, 1 mM EDTA, 0.5 M Tris-HCl pH 8, 0.25 M glycerol, 50 µL of magnetic beads per 50 µL of lysate). The samples were cleared with magnetic beads and the nuclei were collected by centrifugation. The nuclei were then resuspended in lysis buffer 3 (100 mM HEPES pH 7.9 and 0.3 M NaCl) and 2 µL of Proteinase K and 2 µL of 1 M LiCl were added. After overnight incubation at 37 °C, the DNA was extracted using a QIAQuick PCR Purification kit (28104, Qiagen). The purified DNA was then eluted with 1 µL of TE buffer (10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.1% SDS) and stored at −80 °C.

Single step cross-linking ChIP. The single step cross-linking ChIP-seq was performed using the ChIP Assay kit (Millipore) according to the manufacturer's instructions. The lysis buffer 2 was omitted and 400 µg of sonicated DNA were incubated with 10 µg of the antibody of interest overnight at 4 °C. The DNA was then purified using a QIAQuick PCR Purification kit (28104, Qiagen) and eluted in 1 µL of TE buffer (10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.1% SDS) and stored at −80 °C.

Biotinylated CxxC affinity purification (Bio-CAP). Bio-CAP experiments were performed as described above. In brief, genomic DNA was extracted from 2.5 x 10^7 cells using 950 µL of extraction buffer (20 mM Tris HCl (pH 8.0), 1 mM EDTA, 0.1% Triton X-100, 0.5% SDS) followed by RNaseA/TE (EN0551, Thermo Scientific) treatment and phenol/chloroform extraction. The DNA was sonicated to an average size of 200 bp and diluted to 17.5 µg/mL in CAP100 buffer (50 mM Hepes, pH 7.9 and 100 mM NaCl). 100 µL aliquot was used as input. NeutrAvidin Agarose Resin (29201, Thermo Scientific) was conjugated with biotinylated KDM2B ZF-CxxC protein (kindly provided by Rob Klose). 500 µL of diluted sonicated DNA were incubated with conjugated CxxC resin for 1 h at 4 °C. Serial washes and elution methods using standard protocol. Biotinylated capture oligonucleotides were designed to the ends of the viewpoint fragments (list of Capture-C oligos in Supplementary Table 5). Where possible 1-2 µg of each adapter-ligated library were hybridized with the biotinylated capture oligonucleotides, using the Nimblegen SeqCap reagents and an adapter protocol. The quality of the resultant captured library was assessed by Agilent tape station or bioanalyzer (D1000).

Library preparation. Unless specified otherwise, library preparation was performed using NEBNext Ultra DNA Library Prep Kit for Illumina (E7370L, NEB) and the NEBNext Multiplex Oligos for Illumina (E7353, NEB) following the manufacturer’s protocol.

Single cell ATAC-seq. sc-ATAC-seq was performed using the 10X Genomics Chromium® V1 1 Multiplex Kit N Set A (1000084). Chromium® V1 Chip Single Cell 3’ Reagent Kit v3 (1000076) was used to perform single cell capture. Chromium® Single Cell 3’ Reagent Kits were analysed using cellranger/3.0.2. The scRNA-seq data were then analysed using scRNAseq on aggregated samples with the following parameters: α = 50, β = 0.1, iterations = 500 and a number of topics between 2 and 100 (2, 10, from 20 to 60, 1 by 1; from 70 to 100, 10 by 10). The chromatin accessibility around the HB Mayer, HBB and HBA genes were analysed based on the gene score activity.

Single cell RNA-seq. scRNA-seq was performed using the 10X Genomics Chromium® V1 Single Cell 3’ Reagent Kits v3 (1000075). 50,000 cells to start material following the manufacturers’ instructions and sequenced as described below in Library quantification and High throughput sequencing. Sequencing data were analysed using Cellranger/3.0.2. The scRNA-seq data were then analysed using the Seurat package and Signacular extension46. Low count cells (passed filters > 500) were filtered and 4000 cells per sample were randomly selected. Normalised signals were compared by boxplots. In parallel, the samples were analysed with cisTopics47 on aggregated samples with the following parameters: α = 50, β = 0.1, iterations = 500 and a number of topics between 2 and 100 (2, 10, from 20 to 60, 1 by 1; from 70 to 100, 10 by 10). The chromatin accessibility around the HB Mayer, HBB and HBA genes were analysed based on the gene score activity.
platforms (KR0405, KAPA Biosystems). 4 nL libraries with compatible indexes were pooled and sequenced on a NextSeq 500 sequencer (Illumina) using either 300 cycles (Capture-C only), 150 cycles or 75 cycles NextSeq 500/550 High Output v2 kit (Illumina).

**Analysis of raw sequencing data from ChiP-seq, ATAC-seq and Bio-CAP-seq experiments.** Fastq files were aligned to the human genome hg19 using an in-house pipeline described in79 (https://github.com/Hughes-Genome-Group/NGSequencingBasic/releases) using Bowtie 2. In addition, –atc flag was used for ATAC-seq analysis. Data quality was assessed using fastQC reports59. Trimming of the adapter sequences was performed using trim-galore52. PCR duplicates were removed using Samtools53. Signal artefact blacklisted regions54 were excluded. For normalisation, with the exception of ChiP-Rx data which were normalised following60 the (the normalization factor was adjusted to take into account the ratio of mapped human (hg19) and Drosophila (dm3) reads in the ChiP and input samples), the total number of mapped reads or a total number of mapped reads under peak regions was determined after the removal of PCR duplicates and excluding the reads mapping to the mitochondrial chromosome. Reads were normalised per 100 million reads. Data were visualised on the UCSC genome browser (http://genome.ucsc.edu/)55. Versions of software packages used for the analysis include FASTQC 1.0.19, HOMER 4.10. McDowell, T. L. et al. Localization of a putative transcriptional regulator cause X-linked mental-retardation with alpha-defects. J. Cell Biol. 180, 315–324 (2008).

**Gene ontology and motif analyses.** Gene ontology analysis for the differentially expressed genes has been performed using GOTERM_BP_DIRECT DAVID 6.860. Gene and genome ontology analyses based on the ATRA binding sites not encoded in LCLs and erythroblasts as well as the ATRX binding sites not encoded in LCLs were performed using R scripts based

**Graphical analysis.** Heat maps and profile plots analyses were performed using Deeptools61. Volcano plots and single cell analysis plots were plotted on R. Box plots and associated statistical analyses were performed with GraphPad Prism 9 or R.

**Repeatome analysis.** Repeat enrichment in ATRA ChiP-seq data (read one of the paired end only) was performed using a repeat analysis pipeline based on62 and mapped using the hg18 assembly to compare a union of the fastq reads of ChiP vs Input.

**RefSeq and CpG island annotation.** RefSeq annotation for hg19 was used to characterise the promoter, gene body and intergenic regions identified at ATRA binding sites80. The CpG island annotation for hg19 was retrieved from the cpgIsland Ext table in the UCSC database.

**GenoSTAN analysis.** GenoSTAN analysis was performed using R scripts based on67. Analyses were performed on 10000 bp fragments covering the ATAC-seq peaks identified in LCLs and erythroblasts as well as the ATRA binding sites not overlapping with ATAC-seq peaks. For the annotation in LCLs, H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K27me3 ChiP-seq datasets were used. In addition, CTCF ChiP-seq (GEO accession number GSM733752) was used in the analysis. 12 states were initially used and analogous states were subsequently pooled together resulting in seven distinct states: active promoter (P), poised Promoter (Pp), active enhancers (E), enhancer (En), CTCF binding site (EC), CTCF site (C), repressed region (R) and background (B). For the annotation in erythroblasts, five chromatin marks, H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K9me3, were used. 10 states were initially used and analogous states were subsequently pooled together resulting in five distinct states: active promoter (P), poised Promoter (Pp), active enhancers (E), repressed region (R) and background (B).

**Transcription factor binding site analysis.** TFBS analysis was performed using the Uniform TFBS (wgEncodeAwgTfbsUniform) dataset associated with LCLs64.

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

**Data availability.** The data that support this study are available from the corresponding authors upon reasonable request. The microarray and sequencing data generated for this study have been deposited at Gene Expression Omnibus (GEO) under the accession numbers: GSE192767, GSE193038, GSE193310, GSE193511, GSE193312, GSE193314 and GSE193315. Additional data used in this study are available under the GEO accession numbers GSM733752, GSM758559 and GSM125924856 as well as in the Uniform TFBS (wgEncodeAwgTfbsUniform) dataset associated with LCLs64. Genomic assemblies: Homo sapiens (human) genome assembly NCBI36 (hg18) (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.12), Homo sapiens (human) genome assembly GRCh37 (hg19) (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13). Source data are provided with this paper.

Received: 27 August 2021; Accepted: 7 June 2022; Published online: 17 June 2022

**References**

1. Clynes, D., Higgs, D. R. & Gibbons, R. J. The chromatin remodeler ATRX: a repeat offender in human disease. Trends Biochem. Sci. 38, 466 (2013).
2. Gibbons, R. J. et al. Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. Nat. Genet. 24, 368–371 (2000).
3. Law, M. J. et al. ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell 143, 367–378 (2010).
4. Clynes, D. et al. ATRA dysfunction induces replication defects in primary mouse cells. PLoS ONE 9, e92915 (2014).
5. Clynes, D. et al. Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRA. Nat. Commun. 6, 7358 (2015).
6. Ritchie, K. et al. Loss of ATRA leads to chromosome cohesion and congression defects. J. Cell Biol. 180, 315–324 (2008).
7. De La Fuente, R., Viveiros, M. M., Wigglesworth, K. & Eppig, J. A. ATRX, a member of the SNF2 family of helicase/ATPases, is required for chromosome alignment and meiotic spindle organization in metaphase II stage mouse oocytes. Dev. Biol. 272, 1–14 (2004).
8. Gibbons, R. J., Picketts, D. J., Villard, L. & Higgs, D. R. Mutations in a putative global transcriptional regulator cause X-linked mental-retardation with alpha-thalassemia (ATRX-X Syndrome). Cell 80, 837–845 (1995).
9. Heaphy, C. M. et al. Altered telomeres in tumors with ATRX and DAXX begin to uncover the ATRX role in regulating telomere length. Genome Res. 13, 753–761 (2003).
10. McDowell, T. et al. Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes. Proc. Natl Acad. Sci. USA 96, 13983–13988 (1999).
11. Wong, L. H. et al. ATRA interacts with H3.3 in maintaining telomere structural integrity in pluripotent embryonic stem cells. Genome Res. 20, 351–360 (2010).
12. Vosn, H. F. et al. ATRA plays a key role in maintaining silencing at interstitial heterochromatic loci and imprinted genes. Cell Rep. 11, 405–415 (2015).
13. Sadic, D. et al. ATRA promotes heterochromatin formation at retrotransposons. Embo Rep. 16, 836–850 (2015).
14. Valtchev, A. et al. ATRA binds to atypical chromatin domains at the 3' exons of zinc finger genes to preserve H3K9me3 enrichment. Epigenetics-U5, 0, https://doi.org/10.1080/15592294.2016.1169351 (2016).
15. Deaton, A. M. et al. Enhancer regions show high histone H3.3 turnover that changes during differentiation. Elife 5, https://doi.org/10.7554/eLife.15316 (2016).
16. Truch, J., Telenius, J., Higgs, D. R. & Gibbons, R. J. How to tackle challenging ChiP-Seq, with long-range cross-linking, using ATRA as an example. Methods Mol. Biol. 1832, 105–130 (2018).
17. Blackledge, N. P. & Klose, R. J. CpG island chromatin A platform for gene regulation. Epigenetics-U5 6, 147–152 (2011).
18. Eustermann, S. et al. Combinatorial readout of histone H3 modifications specifies localization of ATRA to heterochromatin. Nat. Struct. Mol. Biol. 18, 777–782 (2011).
19. Iwasue, S. et al. ATRA ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. Nat. Struct. Mol. Biol. 18, 769–776 (2011).
20. Zacher, B. et al. Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. Cell Rep. 12, 1373–1383.e1 (2015).  
21. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–U52 (2011).  
22. Nguyen, D. T. T. et al. The chromatin remodelling factor ATRX suppresses R-loops in transcribed telomeric repeats. *Embo Rep.* **18**, 914–928 (2017).  
23. Schenkel, L. C. et al. Identification of epigenetic signatures associated with alpha thalassaemia/mental retardation X-linked syndrome. *Epigenet. Chromatin* **10**, https://doi.org/10.1186/s13048-017-0118-4 (2017).  
24. Scott, C. et al. Recapitulation of erythropoiesis in congenital dyserythropoietic anaemia type I (CDA-I) identifies defects in differentiation and nucleolar abnormalities. *Haematologica* https://doi.org/10.3324/haematol.2020.260158 (2020).  
25. Johansson, P. A. et al. A cis-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development. *Cell Stem Cell* **29**, 52–69 e58 (2022).  
26. Goldberg, A. D. et al. Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell** 140, 678–691 (2010).  
27. Schneiderman, J. I., Orsi, G. A., Hughes, K. T., Loppin, B. & Ahmad, K. Nucleosome-depleted chromatin gaps recruit assembly factors for the H3.3 histone variant. *Proc. Natl Acad. Sci. USA* **109**, 19721–19726 (2012).  
28. Danussi, C. et al. Atrx inactivation drives disease-defining phenotypes in glioma cells of origin through global epigenomic remodeling. *Nat. Commun.* **10**, https://doi.org/10.1038/s41467-019-13671-3 (2019).  
29. Jin, C. & Felsenfeld, G. Distribution of histone H3.3 in hematopoietic cell lineages. *Proc. Natl Acad. Sci. USA* **103**, 574–579 (2006).  
30. Jin, C. Y. & Felsenfeld, G. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Gene Dev.* **21**, 1519–1529 (2007).  
31. Kernohan, K. D., Vernimmen, D., Gloor, G. B. & Berube, N. G. Analysis of R-loops in transcribed telomeric repeats. *Mol. Cell** **57**, 539–552 (2015).  
32. Macleod, D., Charlton, J., Mullins, J. & Bird, A. P. Sp1 sites in the mouse Aprt gene are required for G-quadruplex DNA-driven transcriptional activation. *Genet. Biol.* **8**, 50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).  
33. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).  
34. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).  
35. Kundaje, A. A comprehensive collection of signal artifact blacklist regions in the human genome. https://personal.broadinstitute.org/ashual/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf (2013).  
36. Kern, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 1160–1168 (2002).  
37. Telenius, J. M. et al. CaptureCompendium: a comprehensive toolkit for 3C analysis. Preprint at https://www.biorxiv.org/content/10.1101/2020.02.17.952572v1 (2020).  
38. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R13 (2008).  
39. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell** **38**, 576–589 (2010).  
40. McGowan, S. J., Hughes, J. R., Han, Z. P. & Taylor, S. MIG: multi-image genome viewer. *Bioinformatics* **29**, 2477–2478 (2013).  
41. Gentleman, R., Carey, V. J., Huber, W. & Hahne, F. gene 1.74.0. R package version 1.74.0.  
42. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Genome Res.* **25**, 1074–1082 (2015).  
43. Wang, Y. X. et al. G-quadruplex DNA drives genomic instability and acyclic regions. *Nucleic Acids Res.* **37**, 3897–3911 (2009).  
44. Raymire, P. D., Furey, D., Diehl, S., Gruning, B. A. & Manke, T. DeepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).  
45. Day, D. S., Luqueett, L. J., Park, P. J. & Kharchenko, P. V. Estimation of enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.* **11**, R69 (2010).  
46. O’Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).  
47. Consortium, E. P. et al. An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57–74 (2012).  
48. Downes, D. J. et al. An integrated platform to systematically identify causal variants and genes for polygenic human traits. Preprint at https://www. biorxiv.org/content/10.1101/813618v2 (2020).  

Acknowledgements

We are very grateful to the ATR-X families and normal controls who donated samples that made this work possible. We are grateful to A. Pellagatti, S. Roy and V. Steeles for their help with the microarray assays, to the Klose group (Department of Biochemistry, University of Oxford) for providing the material to perform the Bio-CAP assay, to the Fulga group for the production of S2 cells and to the Miéne group for the crosslinked S2 cells and their valuable advice. We thank C. Fisher for the HbH inclusion assay and his advice, to S. Butler for her precious help with cell culture, to B. Xella and D. Clynes for their help and advice, to R. Beagrie for his help with the erythroblast experiments and T. Rostroon (WIMM screening core) provided sequencing services. The single-cell experiments were performed in the WIMM single-cell facility with the help of N. Ashley. The Computational Biology Research Group, WIMM helped with the bioinformatics. The cell sorting and analysis were performed in the WIMM single cells facility with the help of J.T., R.J.G. and D.R.H. conceptualised the study and conceived the experimental design; J.M.B. carried out the RNA FISH. J.T and D.J.D. performed most of the differentiation and FACS analyses. M.G. contributed to the bulk ATAC-seq on erythroblasts. J.M.T. provided bioinformatic pipelines and supervision. E.R.G. performed the t-SNE analysis of the scATAC-seq data. E.R. performed the power cal-

Author contributions

J.T., R.J.G. and D.R.H. conceptualised the study and conceived the experimental design; J.T. generated most of the data and analyses. D.J.D. helped to carry out experiments and analyses generated the RNA-seq and Capture-C data from erythroblasts. C.S. and P.L.C. performed the CD34+ HSPC purification. C.S. performed the CD34+ HSPC differentiation and FACS analyses. M.G. contributed to the bulk ATAC-seq on erythroblasts. J.M.R. carried out the RNA FISH. J.T and D.J.D. performed most of the bioinformatics analyses. J.M.T. provided bioinformatic pipelines and supervision. E.R.G. performed the t-SNE analysis of the scATAC-seq data. E.R. performed the power calculation and the analysis of the microarrays raw data. S.T. provided the pipeline for the CEBP analysis. R.S. provided the pipeline and guidance for the GenoSTAN analyses. J.R.H. provided supervision for the bioinformatic analyses and gave advice. J.M.T., J.R.H., D.R.H. and R.J.G. advised on the interpretation of the data. J.T., D.R.H. and R.J.G wrote the original manuscript; all authors contributed to reviewing and editing the manuscript. D.R.H. and R.J.G. provided supervision and funding.

Competing interests

The authors declare no competing interests.
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31194-7.

Correspondence and requests for materials should be addressed to Douglas R. Higgs or Richard J. Gibbons.

Peer review information Nature Communications thanks Bartek Wilcynski, Ivan Costa and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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.