Genome Biology

RESEARCH Open Access

Widespread allele-specific topological domains in the human genome are not confined to imprinted gene clusters

Stephen Richer[1], Yuan Tian[1,2], Stefan Schoenfelder[3], Laurence Hurst[1], Adele Murrell[1*] and Giuseppina Pisignano[1*]

*Correspondence:
amm95@bath.ac.uk;
gp529@bath.ac.uk

[1] Department of Life Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK
[2] UCL Cancer Institute, University College London, Paul O'Gorman Building, London, UK
[3] Babraham Institute, Cambridge CB22 3AT, UK

Abstract

Background: There is widespread interest in the three-dimensional chromatin conformation of the genome and its impact on gene expression. However, these studies frequently do not consider parent-of-origin differences, such as genomic imprinting, which result in monoallelic expression. In addition, genome-wide allele-specific chromatin conformation associations have not been extensively explored. There are few accessible bioinformatic workflows for investigating allelic conformation differences and these require pre-phased haplotypes which are not widely available.

Results: We developed a bioinformatic pipeline, "HiCFlow," that performs haplotype assembly and visualization of parental chromatin architecture. We benchmarked the pipeline using prototype haplotype phased Hi-C data from GM12878 cells at three disease-associated imprinted gene clusters. Using Region Capture Hi-C and Hi-C data from human cell lines (1-7HB2, IMR-90, and H1-hESCs), we can robustly identify the known stable allele-specific interactions at the *IGF2-H19* locus. Other imprinted loci (*DLK1* and *SNRPN*) are more variable and there is no "canonical imprinted 3D structure," but we could detect allele-specific differences in A/B compartmentalization. Genome-wide, when topologically associating domains (TADs) are unbiasedly ranked according to their allele-specific contact frequencies, a set of allele-specific TADs could be defined. These occur in genomic regions of high sequence variation. In addition to imprinted genes, allele-specific TADs are also enriched for allele-specific expressed genes. We find loci that have not previously been identified as allele-specific expressed genes such as the bitter taste receptors (*TAS2Rs*).

Conclusions: This study highlights the widespread differences in chromatin conformation between heterozygous loci and provides a new framework for understanding allele-specific expressed genes.

Background

Higher-order chromatin conformation forms a scaffold upon which epigenetic mechanisms converge to regulate gene expression [1, 2]. Many genes are expressed in an allele-specific manner in the human genome, and this phenomenon is an important contributor to heritable differences in phenotypic traits and can be cause of congenital and acquired diseases including cancer [3, 4]. In most cases, allele-specific expression is driven by sequence variants located within gene regulatory elements to confer allele-specific preference for transcription factor binding. Genome-wide association studies (GWAS) have linked variants and diseases and have enabled insights into complex-trait genetics and important biological processes in gene regulation and mechanisms underlying disease. Allele-specific differences in chromatin conformation may be masked by chromatin capture approaches designed to provide a snapshot of a high number of dynamic interactions, averaged across both alleles in a heterogeneous cell population. Such approaches may therefore bias the interpretation of chromatin conformation at sites of allele-specific gene expression (ASE) [5]. Genome-wide association studies (GWAS) have linked variants and diseases and have enabled insights into complex-trait genetics and important biological processes in gene regulation and mechanisms underlying disease [6]. Methods that integrate GWAS data with expression quantitative trait loci (eQTL) data to identify associated genes [7], and approaches that combine epigenetic data such as DNA methylation [8], ChIP-seq, and DNase I hypersensitivity have been used to suggest functional hypotheses for variants associated with diseases [9]. More recently, chromatin interaction information has been used to link GWAS variants to target genes [10–12] and more tools are being developed to predict the functional effects of variants in disease including combining artificial intelligence and deep learning with Hi-C data [13].

Genomic imprinting is a special case of allele-specific expression, characterized by parent-of-origin monoallelic expression that is regulated by an array of epigenetic mechanisms rather than genetic sequence of the allele [14]. Epigenetic elements of imprinted gene regulation include sequences that are methylated on only one of the parental alleles (known as differentially methylated regions, DMRs). DMRs further have underlying allelic differences in post-translational histone modifications and "CCCTC-binding factor" (CTCF) occupancy. Where the DMRs have been shown to regulate imprinted gene expression in *cis*, they are referred to as imprinting control regions (ICRs) (reviewed [15].

Disturbances of the allelic dosage due to chromosomal rearrangements or the epigenetic disruption of co-regulated expression in imprinted genes, lead to defined clinical syndromes collectively known as imprinting disorders (reviewed [15]). The most common imprinting disorders include Beckwith-Wiedemann syndrome (BWS) [16, 17] with an incidence of 1 in 15,000 live births, Angelman syndrome (AS, 1:20,000); Prader-Willi syndrome (PWS, 1:25,000) [18], Silver-Russell syndrome (SRS, 1:100,000) [17], Temple (MatUPD14), and Kagami-Ogata (PatUPD14) syndromes [19, 20].

The *IGF2-KCNQ1* locus, implicated in BWS and SRS, divides into two imprinted gene clusters, each regulated by a separate imprinting control region (ICR). The first is *IGF2-H19*, with its ICR (*H19*-DMR) methylated on the paternal allele. The second cluster has a maternally methylated ICR (KvDMR) at the promoter of the long non-coding RNA (lncRNA) *KCNQ1OT1* gene [21] that when active silences *KCNQ1* and adjacent genes [22]. The *SNRPN* locus is implicated in Prader-Willi and Angelman syndromes. Imprinting at this region is regulated by a ~35-kb bipartite imprinting control region (PWS-IC and AS-IC). The PWS-IC section is methylated on the maternal allele and is the promoter for the pre-mRNA transcript for *SNRPN*, *SNURF*, and *SNHG14*, which is also a host transcript for several other long and short non-coding RNAs [23]. The *DLK1-DIO3* locus is implicated in Temple and Kagami-Ogata syndromes and includes *DLK1* and *RTL1* (paternally expressed genes) and several maternally expressed ncRNAs (*MEG3*, *RTL1-AS*, *MEG8*), snoRNAs, and miRNAs [24]. Two DMRs, IG-DMR (located in the intergenic region between *DLK1* and *MEG3*) and *MEG3*-DMR (at the *MEG3* promoter), regulate imprinted expression at this locus [25]. These DMRs are methylated on the paternal allele in most somatic tissues.

Imprinted genes are an excellent model system for analyzing epigenetic regulation of gene expression and the study of genomic imprinting has uncovered many paradigms that are generally relevant to gene expression [26]. One such paradigm is that the CTCF can act as a boundary element separating different regulatory elements that could be shared between genes. We and others have shown that differential binding of CTCF-cohesin complexes at the imprinted *IGF2-H19* locus regulates access to a series of enhancers through allele-specific differences in higher-order looping interactions [27–32].

These early studies used a chromosome conformation capture (3C) technique in which fixed chromatin is digested with a restriction enzyme followed by a ligation reaction that favors regions in close proximity. The principle of 3C technology is that interactions between distant regulatory regions that come close together in the 3D space will be more frequently detected than random interactions [33]. Newer technologies coupled to next-generation sequencing (Hi-C, Capture Hi-C) have enabled the detection of topologically associating domains (TADs), defined as local regions within a chromosome with a high density of interactions (contact clusters) that also exhibit insulation from one another [2, 34, 35].

TADs are thought to regulate gene expression by increasing the frequency of intra-domain promoter-enhancer interactions and insulating against spurious inter-domain interactions. TADs are formed via cohesin-mediated loop extrusion, whereby DNA is bidirectionally extruded through the ring-shaped cohesin complex until it is halted by convergently oriented CTCF to form a TAD boundary [36].

It is further assumed that CTCF and associated protein TAD boundaries compartmentalize the genome to implicitly prevent transcription read-through and spurious transcriptional activation of silent genes or constrain the spread of silencing chromatin [37–42]. Parameters such as CTCF density and orientation, as well as DNA methylation, have been shown to affect TAD direction, size and overall structure. Hi-C techniques have also identified that the higher-order structure is further shaped by nucleosome accessibility and divides into A- and B-compartments, each with distinctive chromatin and transcription features.

Mouse models in which an imprinted locus can be deleted and transmitted through either the male or female germline, have enabled allele-specific Hi-C profiles for the *Igf2-H19*, *Dlk1-Dio3* imprinted loci. For these loci in the mouse, it has been shown that the maternally and paternally imprinted genes are located together in large TADs that are similar in both parental alleles. Within the TADs, differential binding of CTCF creates allele-specific subTAD structures that provide the instructive or permissive context for imprinted gene activation during development [43].

A limitation to studying imprinted genes in humans has been the need for family studies to ascertain the parental origin of genes. Technologies that detect long-range *cis* interactions fortuitously link single-nucleotide polymorphism (SNP) variants within a chromosome and provide molecular haplotype information. One of the first studies to use haplotype phasing in Hi-C data from a human lymphoblastoid cell line, GM12878, detected allele-specific long-range interactions between a distal locus, termed HIDAD (Distal Anchor domain) and the promoters of the maternally expressed *H19* and the paternally expressed *IGF2* [44]. *IGF2-H19* has been studied in great depth as the archetypal locus for allele-specific interactions [44, 45]. However, the allele-specific-methylation-sensitive-CTCF-binding-for-alternative-looping paradigm as established for *IGF2-H19* is not universally true for all imprinted gene clusters.

In this study, we sought to examine how the higher-order chromatin conformation structures differ between the active and silent alleles at loci containing genes that are allele-specifically expressed in humans. To this end, we assembled a HiCFlow pipeline for processing raw Hi-C data for haplotype phasing and construction of allele-specific chromatin conformation profiles. A number of existing pipelines, including HiC-Pro [46], are capable of performing allele-specific Hi-C. However, these require a pre-phased haplotype as input as they cannot perform de novo haplotype assembly from input Hi-C data. Moreover, most do not have functionality to generate and visualize between-sample normalized differences in contact frequency. As such, we opted to assemble a custom pipeline that integrates the required functionality into a single workflow. Following this, we were able to characterize allelic differences at human imprinted gene clusters to establish the epigenetic framework for differential association frequencies.

Our analyses indicate that imprinted gene domains are not uniformly organized within a canonical higher-order structural profile regulated by elements within the ICRs. At the *IGF2-H19* locus, the ICR plays a direct role in directing allele-specific CTCF-mediated higher-order chromatin structures consistent with loop extrusion models, whereas at other loci, the ICR may have indirect or no specific effect. Allele-specific compartmentalization was observed in some cell lines at the *SNRPN* and *DLK1* loci. Rather than remaining spatially and temporally separated from their non-imprinted neighbouring genes, imprinted gene clusters share TADs with non-imprinted genes. Indeed, most allele-specific interactions occur within subTADs. Imprinted domain boundaries may be delimited by TAD structures, but some allele-specific associations can occur across TAD boundaries with concomitant effects of allele-specific expression. Allele-specific interactions were not confined to imprinted domains. In an unbiased genome-wide screen, we detected additional allele-specific TADs (ASTADs). The ASTAD distribution varied between cell lines. We found 8–32% of genes with allele-specific expression to be located within ASTADs. Regions of high genetic variability, such as olfactory receptor loci, the bitter taste receptor (*TAS2R*) gene cluster and the keratin gene (*KRT*) cluster, were found to be within ASTADs.

Results

A region capture data set with a HiCFlow pipeline provides an effective platform for haplotype phasing of imprinted gene loci

To examine imprinted loci at high resolution, we first generated a Region Capture Hi-C (RC-Hi-C) dataset in a human breast epithelial cell line, 1-7HB2. This diploid cell line has previously been used to examine allele-specific expression and imprinted methylation for several imprinted genes [31, 47–49] and been shown to have methylation and expression profiles consistent with the maintenance of normal imprinting in a somatic cell line. Using a tiled probe RC-Hi-C approach, combined with a frequent 4 base-cutter restriction enzyme (*MboI*), we generated capture regions (totalling 25Mb) at 5 imprinted chromosomal loci, the largest regions included *SNRPN*, *DLK1-DIO3*, and *IGF2-KCNQ1* (capture region IDs: CR_3, CR_2 and CR_4, respectively, Additional file 1: Table S1). In total, 34,399 probes were used covering approximately 4.1Mb (~16.1%) of the capture regions.

Our RC-Hi-C dataset yielded approximately 40 million valid read pairs with a mean coverage of almost 1700 read pairs per kilobase and was comparable to published high-resolution Hi-C datasets at the same genomic regions (Additional file 2: Fig. S1a). We assembled a Hi-C analysis pipeline (HiCFlow) to process raw Hi-C data to normalized matrices and for haplotype phasing (Additional file 2: Fig. S1b and Methods). This enabled the construction of allele-specific chromatin conformation profiles (alleles designated “A1” and “A2”). Separating the allelic profiles at imprinted loci and visually presenting them as A1 and A2 matrices showed subtle allelic differences in contact frequency. Therefore, we added a subtraction matrix function to the HiCFlow pipeline for highlighting interaction differences between alleles. To further emphasize regions with consistent directional bias, the subtraction matrices were denoised using a median filter (Additional file 2: Fig. S1c). We used the *IGF2-H19* locus to benchmark the allele-specific subtraction parameters and confirmed that the known allele associations could be robustly detected (Additional file 2: Fig. S1c). The subtraction matrix methodology provides a compromise between false positive calls in regions of low interaction density and missing interactions in regions of high interaction frequencies.

The RC-Hi-C library provided an excellent dataset to test and refine the HiCFlow pipeline. We included regions of non-imprinted genes that could be tested as negative controls. These showed similar profiles when separated into A1 and A2 alleles, and only slight indistinct differences in a subtraction matrix (Additional file 2: Fig. S1d).

Evaluation of bias during HiCFlow genotyping and haplotyping

The HiCFlow pipeline utilizes Hi-C data to perform both genotyping and haplotyping. A potential source of bias in this approach is loss of variant calling accuracy at differentially interacting sites, leading to mistaken homozygosity. Although missed heterozygous variants would not influence phasing accuracy, it would reduce the coverage of the haplotype. To assess the impact of this, we performed genotyping and phasing of GM12878 using HiCFlow and compared the results with the experimentally validated haplotype. In total, 4,267,624 and 4,049,512 variants were identified by

HiCFlow and the high-confidence dataset respectively. Of the 3,799,226 loci common to both datasets, there was 99.8% agreement in variant identity.

Phasing accuracy was similarly assessed; in total, 2,147,688 and 2,063,320 phased variants were identified by HiCFlow and the high-confidence dataset respectively. Of the 1,942,361 loci with informative phasing information in both datasets, there was 99.9% agreement in phasing. Visual comparison of the subtraction matrices at the *IGF2-H19* locus revealed no substantial differences (Additional file 2: Fig. S2). These results support the appropriateness of using Hi-C data to perform both genotyping and haplotyping. However, where available, the user should provide a list of SNPs generated from a non-biased assay to control for the possibility of missing rare but very strong allele-specific interactions. The HiCFlow pipeline includes a function to enable users to provide their own list of SNPs prior to phasing.

Imprinted gene clusters in human normal breast epithelial cell line exhibit variable patterns of allele-specific associations

We used HiCFlow to define allele-specific association profiles at three imprinted gene clusters in our 1-7HB2 RC-Hi-C library. These included the *IGF2-KCNQ1* (BWS/SRS locus, chromosome 11p15.5), *SNRPN* (Prader-Willi Angelman (PWS-AS) locus, chromosome 15q11-q13), and *DLK1-DIO3* (chromosome 14q32.2) loci (Fig. 1).

Our RC-Hi-C library captured a 1.2-MB section *H19-KCNQ1* domain and included the proximal enhancer to the *H19* gene [31], both imprinting control regions (*H19*-DMR and KvDMR, marked by arrows above the CTCF track in Fig. 1a) and CTCF sites upstream of the *OSBPL5* promoter. It does not include the HIDAD region described by Rao et al. [44]. In Fig. 1a, the diploid contact matrix for the region shows a series of dense interactions forming several TADs, against a backdrop of cross-TAD interactions, reminiscent of contiguous TAD cliques [51, 52]. The *IGF2* gene promoters are located at the TAD boundary on both alleles (Fig. 1a). The allele-specific matrices display marked differences between A1 and A2, particularly at the *H19*-DMR and the KvDMR regions. In the A1-allele, the *H19*-DMR falls within a TAD, whereas in the A2-allele the DMR subdivides the TAD. By contrast, the ICR at the *KCNQ1OT1* promoter region (KvDMR) forms a weak subTAD boundary. The difference in allele-specific enhancer interactions with *IGF2* (labelled 1, in Fig. 1a) and *H19* (labelled 2 in Fig. 1a) is most clearly seen as a blue signal for the A1 allele and a red signal for the A2 allele in the subtraction matrix. This fits with the known shared enhancer model for *IGF2* and *H19* promoters being regulated by the CTCF sites in the *H19*-DMR [27–32].

The subtraction matrix further indicates a mosaic pattern throughout the region rather than an enrichment of associations of one allele over another such as a complete single colored TAD. Instead, allele-specific bidirectional stripes from the boundary at the *IGF2* gene and KvDMR regions are evident. At short distances, these stripes show a bias towards the A2 allele, whereas at longer range the bias is towards the A1 allele. At the KvDMR, the A2-allele has bidirectional interactions towards the CTCF sites upstream of the *KCNQ1* and CTCF site within the *KCNQ1* gene, as well as towards the CTCF sites at the *KCNQ1DN*, *CDKN1C*, and *PHLDA2* genes.

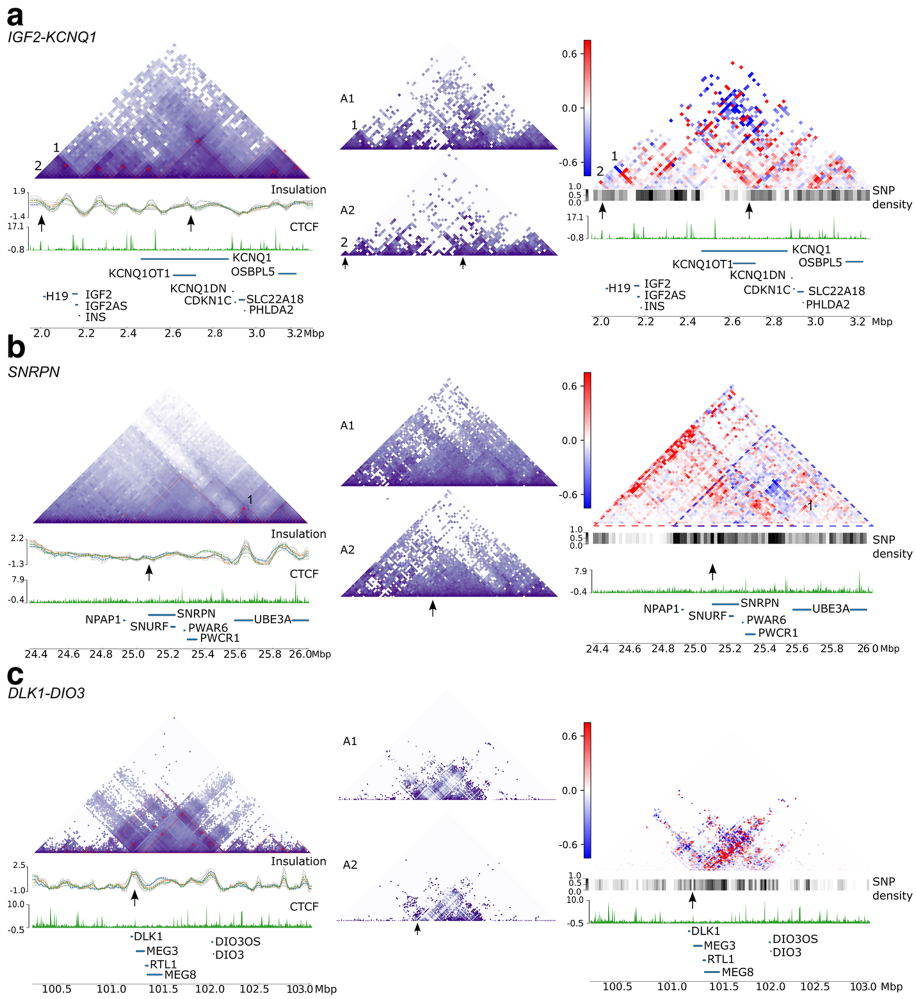

Fig. 1 Allele-specific associations in 1-7HB2 cells with Region Capture Hi-C (RC-Hi-C) for selected imprinted loci. **a** *IGF2-KCNQ1*, with the known *IGF2* and *H19* enhancer interactions labelled 1 and 2. **b** *SNRPN* locus. **c** *DLK1* locus. For each locus, the full (diploid) contact matrix (binned at 10kb resolution) is presented, showing the average of all interactions. Below the diploid contact matrix is the TAD insulation score [50], the CTCF track, and the gene track, with imprinted genes annotated. The positions of DMRs with imprinting control functions are highlighted with arrows above the CTCF track for each matrix. Adjacent to the diploid matrices are the haplotype phased allele-specific matrices (A1 above and A2 below) and a subtraction matrix highlighting the differences between the alleles. Enrichment of A1 relative to A2 (blue), enrichment of A2 relative to A1 (red), the scale bar represents distance-normalized differences between A1 and A2. A SNP density track is included to indicate areas of reduced SNP densities that cannot be haplophased. Allelic differences in these regions cannot be called. Coordinates refer to genome build GRCh37/hg19

At the *SNRPN* locus, our RC-Hi-C library captured a 1.6-MB region including *NPAP1* to *UBE3A*. The bipartite ICR (marked with an arrow in Fig. 1b) is flanked by two CTCF sites. Neither of these sites forms a strong TAD boundary as seen by the weak insulation score below the matrix in Fig. 1b. The strongest insulation score in the region is at the *UBE3A* transcription unit corresponding to a small TAD. Overall, TADs are weak/not clearly defined, which could be explained by the low number of CTCF binding peaks at the locus. When the matrix is split into A1 and A2 profiles according to the phased haplotypes, the A2 allele has a fewer long-range associations. The subtraction matrix shows that the region has directional allelic biases: most A2-allele associations (red) are towards the left, whereas A1 (blue) are towards the right. We have highlighted these as triangles with dotted outlines (Fig. 1b). The scarcity of CTCF binding leads us to postulate that the TAD-like structures at this locus may be formed through phase condensation of heterochromatic compartments. However, region capture data is not suitable for analysis of compartments by current available methodologies.

The 1-7HB2 RC-Hi-C library captured a region of ~ 2.5MB around the *DLK1-DIO3* locus. In the diploid contact matrix, we identified a TAD domain overlapping the imprinted region (Fig. 1c). The DMRs (both IG-DMR/*MEG3*-DMR, marked by an arrow) are within this TAD and appear to form a subTAD boundary (Fig. 1c). However, at this resolution, it is also likely that the subTAD boundary is formed by a CTCF site upstream of *DLK1*. In the individual allele matrices, the A1-allele forms a slightly larger subTAD with CTCF bound region upstream of the *DLK1*, while on the A2-allele, the subTAD is more clearly anchored at the ICR (Fig. 1c). The subtraction matrix shows a V-shape above the DMR. Towards the *DIO3* locus, it forms a predominantly red A2-allele stripe whereas upstream of *DLK1* it forms a blue A1-allele stripe.

It has been proposed that ICRs such as the above DMRs, mediate their epigenetic functions by directing allele-specific chromatin conformation. However, not all imprinted genes contain CTCF sites at their ICRs [48]. Allele-specific chromatin conformation in the 1-7HB2 cell line for the above imprinted gene clusters indicate that while the *IGF2* locus is very clearly shaped by the *H19*-DMR (ICR) that regulates CTCF site availability, this is not invariably the case at other loci. Indeed, at the *SNRPN* locus, TAD-like structures are assembled in the absence of CTCF sites, and independent of the ICR.

Do imprinting control regions directly participate in allele-specific interactions?

To further examine whether known ICRs are anchor points for allele-specific chromatin interactions, we conducted allele-specific viewpoint analyses using haplotype phased Hi-C data from the following cell lines GM12878, IMR-90, and H1-hESC, alongside our 1-7HB2 library. We compared these to the relevant subtraction matrices, to which we also added Peakachu loops [53]. Viewpoint analyses enable focused examination of associations with the ICRs in high-resolution data sets, unobscured by the intrinsic density of Hi-C data. Adding the additional cell line data sets enabled us to investigate how expression, methylation, and heterochromatin compartments correlate with allele-specific chromatin conformation. EBV transformed lymphoblastic cell lines such as GM12878 retain DNA methylation profiles consistent with monoallelic expression for several imprinted genes [54]. GM12878 has been haplotype phased previously, as one of the original International HapMap Project cell lines. The remaining cell lines are karyotypically normal diploid and the publicly available Hi-C data have suitable read depth (Additional file 2: Fig. S1a), to suggest that they could be haplotype phased in our HiC-Flow pipeline.

We first examined the *IGF2-H19* and the *KCNQ1* loci. In GM12878 cells, the maternal origin of the *H19* interactions with the downstream HIDAD locus and the reciprocal paternal *IGF2*-HIDAD interactions have previously been demonstrated [44]. We used this information to set the paternal interactions as A1 (blue) at the *H19*-DMR and maternal interactions at the *IGF2* promoter regions as A2 (red) in the subtraction

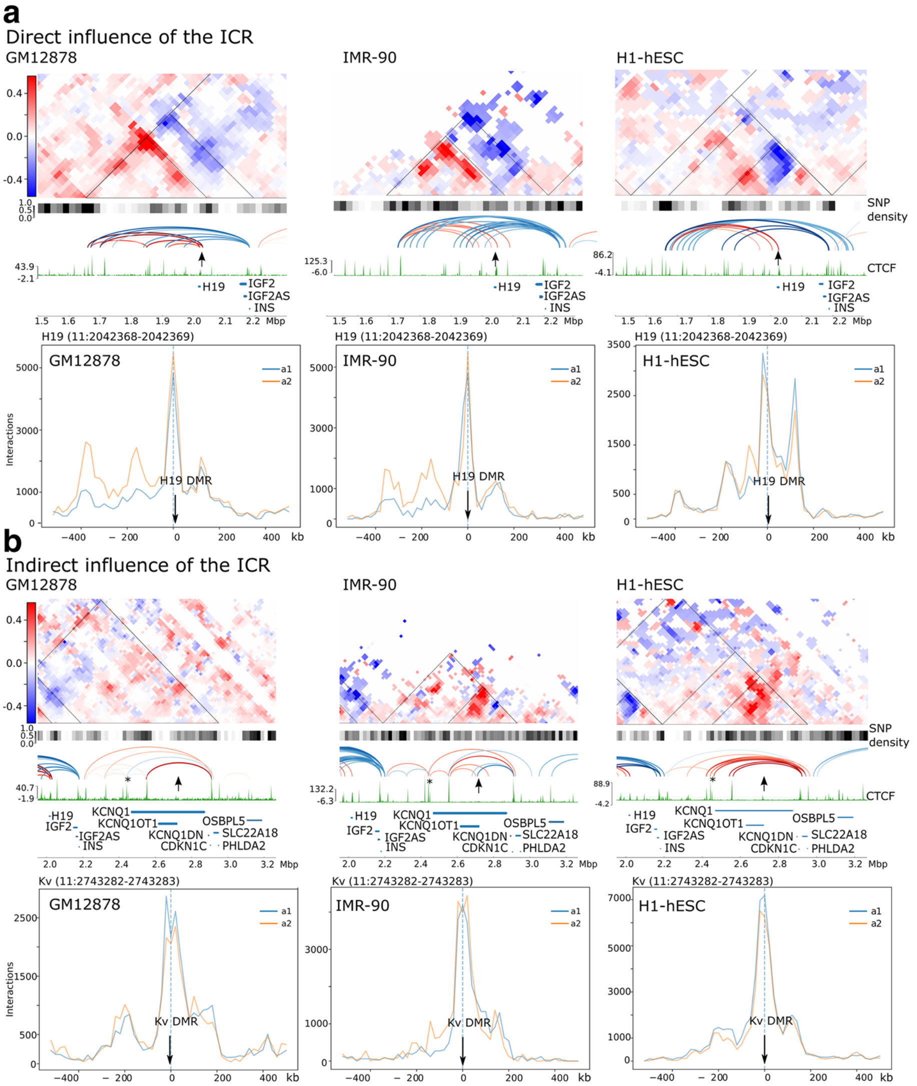

Fig. 2 Imprinting control region (ICR) conformation at the Beckwith-Wiedemann syndrome locus impact differently on chromatin conformation. **a** Direct influence of the ICR in structuring local allele-specific chromatin conformation at the *IGF2-H19* locus, in GM12878, IMR-90 and H1-hESC. Denoised subtraction matrices show that the CTCF regulated *H19*-DMR (arrow) subdivides the region (10kb resolution, blue areas correspond to A1, paternal allele, red areas to A2, maternal allele). Below the matrices we have the SNP density, and allele-specific loops, generated by Peakachu (red maternal, blue paternal), that corresponds to maternal and paternal expression of *H19* and *IGF2* respectively. The color of loops matches the underlying value of the subtraction matrix. The lower panels represent the viewpoint interaction traces for each cell line showing interactions between the *H19*-DMR and loci 400kb in both directions (blue trace A1, paternal, orange A2, maternal). The *H19*-DMR is methylated on the paternal allele in normal cells. **b** Indirect influence of the ICR in structuring local allele-specific chromatin conformation at the *KCNQ1* locus, in GM12878, IMR-90 and H1-hESC. Subtraction matrices and viewpoint interaction traces as in **a** above but focused on the *KCNQ1* locus and the KvDMR ICR (arrow), normally methylated on the maternal allele. Subtraction matrices and the allele-specific loops below display variable structural effects on chromatin conformation by the KvDMR. The viewpoint interaction traces mostly show weak biallelic interaction traces for the KvDMR associations. The CTCF site highlighted with an * is "region 3," previously been shown to be important for allele-specific (maternal) expression of *KCNQ1* [21]

matrices for all three cell lines. This enabled us to also assign the parental origin to the nearby *KCNQ1* locus. Figure 2 demonstrates the effects of the ICRs on higher-order structures at the *H19* and *KCNQ1* loci. We note that the *H19*-DMR strongly anchors the maternal allele-specific interactions as can be seen in the subtraction matrices (Fig. 2a). This association is independent of the *H19* RNA levels which are high in H1-hESCs, and relatively low in GM12878 and IMR-90 (Additional file 2: Fig. S3a). *IGF2* transcript levels in IMR-90 and H1-hESC are higher than in GM12878 cells (Additional file 2: Fig. S3a). Interestingly, the three cell lines vary for the CTCF interactions with the *H19*-DMR, possibly indicating different tissue-specific enhancer associations. Both GM12878 and IMR-90 show this ICR associating with HIDAD region (~0.3Mbp downstream of the ICR), while in H1-hESCs the ICR interacts with regions further downstream (Fig. 2a). At the *H19*-DMR viewpoints (Fig. 2a, bottom), IMR-90, and GM12878 show peaks of high-frequency A2 (maternal) associations, 50–200kb downstream of this ICR, which correlates with *H19*-enhancer sites. *H19*-DMR also forms weaker biallelic associations at sites up to 150kb upstream. H1-hESCs in contrast have a stronger biallelic association peak upstream of the ICR and fewer allele-specific enhancer peaks downstream, which may reflect less stable imprinted expression previously reported in human ESCs [55]. Overall, despite the variable expression, the subtraction matrices show similar “stripe” structures in all three of the cell lines, which would be consistent with an allele-specific loop extrusion between CTCF sites. Thus, the *H19*-DMR is an anchor point for a scaffold of stable allele-specific associations at the *IGF2-H19* that ostensibly only depend on whether this ICR is correctly methylated.

In contrast at the *KCNQ1* locus, the ICR (KvDMR) has a weaker and more variable effect on the higher-order structure (Fig. 2b). Viewpoint analyses at this ICR showed that it formed associations about 200kb upstream and downstream, in all the cell lines tested, albeit weaker than that seen for the *H19*-DMR and not as allele-specific. In IMR-90 cells, there is a maternal-specific peak (A2-allele) approximately −100kb of the KvDMR (Fig. 2b, bottom) that is also seen in 1-7HB2 (Additional file 2: Fig. S3b). For GM12878 and H1-hESC, the associations with the KvDMR viewpoint are biallelic (Fig. 2b, bottom). This ICR is a promoter for *KCNQ1OT1* for which we confirmed the expression in these cell lines as well as in 1-7HB2 (Additional file 2: Fig. S3a). Median levels of methylation at KvDMR are about 50% for IMR-90 and GM12878 (Additional file 2: Fig. S3c) with a pattern of methylated and unmethylated alleles consistent with an expected pattern for allele-specific methylation and expression and confirms our previous reports for 1-7HB2 and IMR-90 [48]. For H1-hESC, overall methylation levels were less than 25% (Additional file 2: Fig. S3c), but still separated into a pattern of methylated and unmethylated alleles, which together with the overall levels suggests loss of methylation and biallelic expression of *KCNQ1OT1*. Thus, the viewpoint analysis for the IMR-90 cells fits with allele-specific regulation of loops by the ICR. However, for the other two cell lines, there is more ambiguity regarding regulation by ICR.

In all three cell lines, the subtraction matrices indicate that associations surrounding *KCNQ1* are predominantly on the maternal allele (Fig. 2b). The CTCF signal is weak at the KvDMR in these cell lines (Fig. 2b), and there is controversy in the literature about whether this ICR contains CTCF binding sites [56–60]. A paternal-specific loop connecting the KvDMR and other CTCF sites was only present in IMR-90. In IMR-90, the

KvDMR is the anchor point of a small subTAD and bidirectional allele-specific loops with CTCF sites in *KCNQ1* (A2, maternal) and *CDKN1C* (A1, paternal) (Fig. 2b). We note that in these cell lines, several maternal loops were formed between an intragenic CTCF site within the *KCNQ1* gene (marked with an asterisk in Fig. 2b) and other sites at the locus. Recently, this CTCF site has been described as "region 3" by Naveh et al. [21], and it was proposed that interactions between this site and surrounding CTCF sites drive transcription of *KCNQ1* and *CDKN1C* on maternal alleles and are required for normal methylation of the KvDMR. SNPs within this CTCF binding site have previously been reported to be associated with a risk for loss of methylation at this ICR [61]. If this model is correct, then the maternal conformation would prevent *KCNQ1OT1* expression. *KCNQ1OT1* transcription from the paternal allele could potentially displace the intragenic CTCF binding at *KCNQ1* on the paternal allele to reciprocally prevent *KCNQ1* and *CDKN1C* transcription. Thus, unlike the *H19*-DMR, the KvDMR has an indirect effect on chromatin conformation at the locus.

At the *SNRPN* locus, viewpoint analysis at the bipartite imprinting control region indicates a low frequency of associations within a +400-kb window from the ICR (Additional file 2: Fig. S4). GM12878 showed an allele-specific interaction with the PWS-AS-IC viewpoint at about +400kb that corresponded with a small A1 enriched TAD-like area on the subtraction matrix. However, allele-specific loops between the ICR and other sites were not identified by the Peakachu algorithm, as shown below the subtraction matrix (Additional file 2: Fig. S4, left). We did not detect similar interactions in the viewpoint plots for IMR-90, 1-7HB2 and H1-hESC, despite the subtraction matrices indicating a strong accumulation of allele-specific associations especially in H1-hESC (IMR-90 due to reduced SNP density at this region is uninformative). In the embryonic cells, there is more CTCF binding at this locus, compared to other cells and a more distinct TAD structure. Unexpectedly, despite this strong difference in the subtraction matrices, and the clonal methylation patterns indicating an allelic split between methylation and unmethylated alleles, the overall methylation data for CpG sites at the ICR indicate that H1-hESC is hypermethylated, suggesting that there is a prevalence of methylated alleles (Additional file 2: Fig. S3c). *SNRPN* RNA (normally paternally expressed) is present in all these cell lines, but at a lower level in IMR-90. The bipartite imprinting center at the *SNRPN* locus therefore seems to affect allele-specific chromatin conformation at the wider locus. This effect is not reliant on the methylation status of the DMR region within the ICR, at least not in H1-hESC cells. The substantial allelic interaction differences in H1-hESC, despite hypermethylation of the DMR, may reflect the stability of imprinted gene expression of this locus in embryonic stem cells which occurs independently of DNA methylation [55, 62].

At the *DLK1-DIO3* locus, only IMR-90 and 1-7HB2 cells showed allele-specific associations with the IG-DMR viewpoints (Additional file 2: Fig. S5, right). The subtraction matrices for IMR-90, 1-7HB2 and H1-hESC show an allele-specific stripe of interactions from the IG-DMR/*MEG3*-DMR with CTCF sites up to *DIO3* (and beyond in the case of H1-hESC). There may also be a stripe in the opposite direction; however, this is less consistent between cell lines. The *MEG3*-DMR contains CTCF sites, which have been reported to be important in maintaining imprinting in somatic tissues [25]. The strength of allele-specific loops (shown below the subtraction matrices) does not correlate with mRNA levels for *DLK1, MEG3, MEG8 RTL1,* or *DIO3* in these cell lines. Indeed, 1-7HB2 and IMR-90 which had the lowest level of expression for these genes showed the strongest allele-specific loops, and more intense differences between A1 and A2 on the subtraction matrix (Additional file 2: Fig. S5). IMR-90 cells which showed the most distinct difference in allelic conformation was also the only cell line with overall methylation levels likely to support allele-specific methylation (Additional file 2: Fig. S3c). However, in this case, the intermediate levels of methylation could not be validated, as clonal analysis showed that the methylation patterns do not separate into allelic differences. This may be as a result of well-known experimental allele-drop out and clonal artifacts of allelic bisulphite sequencing.

The methylation analysis was done bioinformatically using published whole genome bisulfite sequencing (WGBS) data to provide the mean methylation level per CpG within the region (each dot is a CpG in the boxplot in Additional file 2: Fig. S3c), averaged for all for the region (median line in the boxplot in Additional file 2: Fig. S3c). For an imprinted region, we expect the median methylation score to be 50% if the alleles are methylated on one allele only. However, 50% methylation can also be the result of a heterogeneous mix of methylated CpGs across both alleles. Clonal bisulphite analysis (depicted as circle plots in Additional file 2: Fig. S3c) reveals the methylation state of a CpG in single PCR amplicons. At imprinted loci methylation should separate into patterns of methylated and unmethylated amplicons, but without analyzing large numbers of amplicons, the percentage of methylated to unmethylated alleles cannot be determined. Thus, where alleles separate into methylated and unmethylated amplicons on a circle plot, the box plot could indicate that the ratio of methylated to unmethylated alleles is skewed towards hypomethylation (e.g., KvDMR in H1-hESC and IG-DMR in GM12878, Additional file 2: Fig. S3c) or hypermethylation (e.g., PWS-AS-IC in H1-hESC).

The analyses of these four imprinted ICRs indicate that they participate in chromatin conformation to variable degrees in normal cell lines. At the *H19* locus, the *H19*-DMR robustly directs allele-specific chromatin conformation in keeping with a CTCF-mediated methylation-sensitive enhancer competition model. Here it seems that the chromatin conformation is a stable scaffold even in the absence of *H19* or *IGF2* expression. The IG-DMR seems to direct the chromatin conformation, when normally methylated. At other loci, the ICRs can have indirect effect on chromatin conformation such as the *KCNQ1* and *SNRPN* loci. At the *SNRPN* locus, where there is low amount of CTCF binding, allele-specific associations are present but do not seem to be driven by the ICR. These results suggest ICRs utilize a variety of mechanisms in addition to CTCF insulation to facilitate allele-specific chromatin conformation at imprinted loci.

Allele-specific compartment differences and effects of imprinting domains on neighbouring loci in normal cells

It is not yet understood how imprinted domains are contained locally and why they do not spread across an entire chromosome. It is expected that chromatin structural elements and compartmentalization confine imprinted genes to TADs or subTADs to prevent allele-specific associations spreading beyond their domains. To examine how far allele-specific associations spread and to detect A/B-compartments, we added the

CscoreTool (v1.1) [63] to our HiCFlow pipeline and examined a wider 3-6Mb window around each imprinted cluster.

At the *IGF2-H19* locus, allele-specific interactions did not extend beyond the HIDAD region (chr11:1,500,000) in our cell lines (Additional file 2: Fig. S6a). Interestingly, the recently identified associations between *KRTAP5-6* and *INS* are present within this region [64]. At the *KCNQ1* locus, we identified looping interactions extending from the *KCNQ1* region to *NUP98* and *RRM1* in an adjacent TAD in H1-hESC. *NUP98* and *RRM1* are both monoallelic in this cell line, but are not known to be imprinted, which suggests that monoallelic interactions can and do extend beyond a TAD containing imprinted genes (Fig. 3a, Additional file 2: Fig. S6a). The Cscore analysis for this locus indicates that it is located within a 4-Mb active A-compartment on both alleles in all three cell lines shown as a red bar below the allele-specific matrices in Additional file 2: Fig. S6a.

The *SNRPN* locus is within a heterochromatin B-compartment that starts upstream of the imprinted *MKRN3* locus and extends 5Mb towards the telomeric end of chromosome 15 in the three cell lines (Additional file 2: Fig. S6b and Fig. 3b, top). In GM12878, this is a bifold B-compartment that splits into two sections at a point of insulation just after the *UBE3A* gene (Additional file 2: Fig. S6b). Subtle allelic differences are noted in the matrices for the A1 and A2 alleles. In A1, just above the PWS-IC, there is a break within the B-compartment which is not present in the A2-allele, which remains within the B-compartment (Additional file 2: Fig. S6b). A similar bifold pattern is seen for this region in IMR-90 cells, except for a region just above the *ATP10A* locus which shows this gene to be in an A-compartment on both alleles in allele-specific matrices (Additional file 2: Fig. S6b). The A2 allele shows further small interruptions in the B-compartment just above the *SNRPN* cluster of genes. The A1 allele does not show these breaks (Additional file 2: Fig. S6b). The most striking difference in allele structure for this locus is seen in the embryonic cells (H1-hESC), where the Cscore analyses returns a similar B-compartment for the full matrix as for the other cells, but in the separate alleles the region above the imprinted genes is distinctly located within a wider active A-compartment in one allele (A1), whereas on the A2-allele the imprinted locus remains in an inactive B-compartment (Fig. 3b, top). Interestingly, on the A1 allele, the active compartment seems to spread slightly beyond the boundary upstream of *MKRN3.* H1-hESC seems to have more distinct TAD structures in the separate allele matrices compared to the other cell lines (Fig. 3b, top, and Additional file 2: Fig. S6b). The subtraction matrices and Peakachu loop algorithm confirm the presence of only a few allele-specific loops in GM12878 and IMR-90 (Additional file 2: Fig. S6b), whereas H1-hESCs have several allele-specific loops, forming a TAD above the *SNRPN* region (Fig. 3b). There are also several cross-TAD associations between the *SNRPN* TAD and the adjacent *MKRN3* TAD. These results suggest that the *SNRPN* locus is shaped by phase condensation in conditions of low CTCF binding [65], and that when present, CTCF can enhance and stabilize compartmentalization.

The *DLK1-DIO3* locus, which showed the clearest allele-specific differences surrounding the ICR in IMR-90 (Additional file 2: Fig. S5), was found to have allelic differences in Cscores in these cells (Fig. 3b, bottom). The A1-allele of this locus was in a B-compartment whereas the A2 was in an active A-compartment. In GM12878

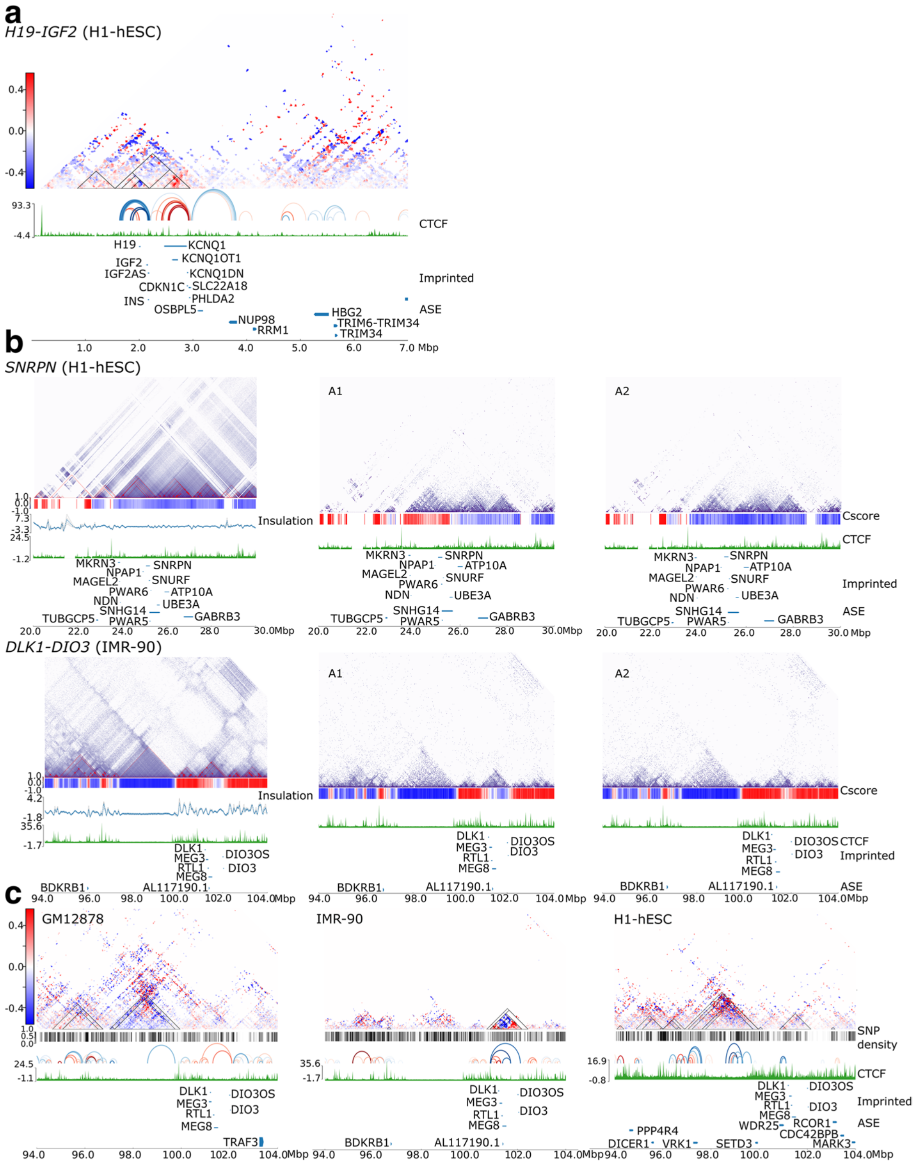

Fig. 3 Effects of imprinting domains on neighbouring loci and allele-specific compartment differences in normal cells. **a** An example of cross-TAD associations from an imprinted gene region. The subtraction matrix at the *H19-KCNQ1* locus with allele-specific loops in H1-hESC demonstrating cross-TAD association between *KCNQ1* region to *NUP98* and *RRM1* which are allele-specifically expressed (ASE), but not known to be imprinted. Gene density is shown in blue below the CTCF track, with imprinted genes below, and genes with ASE below. **b** Examples of allele-specific compartmentalization at *SNRPN* and *DLK1-DIO3* loci. The diploid contact matrix (10kb resolution) with a Cscore below (blue for B-compartment, red A-compartment), followed by TAD insulation score, CTCF track, imprinted genes, and ASE genes. Adjacent to the diploid matrices are the haplotype phased allele-specific matrices (A1 and A2). Note the allele-specific differences in the Cscore track between A1 and A2 alleles at both loci. See Additional file 2: Fig. S6 for a comparison of the other cell lines, and subtraction matrices. **c** Allele-specific cross-TAD associations and additional TAD domains enriched for allele-specific associations near the *DLK1-DIO3* locus. Subtraction matrices, SNP densities, allele-specific loops, imprinted and ASE genes are as described. The *DLK1-DIO3* domain in H1-hESC and GM12878 forms several cross-TAD associations and has weak TAD boundaries. In H1-hESC several genes adjacent to the imprinted domain have allele-specific expression. In GM12878, a nearby TAD (labelled v-TAD) has stronger enrichment for allele-specific associations than *DLK1-DIO3* locus

and H1-hESC cells, the locus has no allelic differences in Cscore with both alleles either in a B-compartment (GM12878, Additional file 2: Fig. S7a) or A-compartment (H1-hESC, Additional file 2: Fig. S7a). In IMR-90, the TADs containing the imprinted genes seem to be sharply defined and separate from neighbouring TADs with no overlapping interactions, especially in the B-compartment (Fig. 3b, bottom). In H1-hESC, there seem to be more cross-TAD interactions and less sharp TAD borders, although no loops extend from the imprinted region into other TADs. Several allele-specific expressed genes in this cell line are detected in the A-compartment, including *WDR25* and *SETD3* located upstream of *DLK1* in an adjacent TAD (Additional file 2: Fig. S7a, left). This 1Mb sized TAD contains several ncRNAs as well as coding genes, none of which have yet been reported to be imprinted in human, although there is evidence that one or more of the orthologous genes are tissue-specifically imprinted in mice [66].

All cell lines have a large 3-Mb-sized TAD corresponding to a B-compartment that is located 2Mb upstream of the *DLK1* cluster (Fig. 3b, bottom, Additional file 2: Fig. S7a). The subtraction matrices show strong allelic bias for predominantly A1 associations in GM12878 cells, and to a lesser extent in H1-hESCs (Fig. 3c, right). We have named this the "v-TAD", after a single coding gene, *VRK1* near the TAD boundary. We examined this region to see whether structural variations were present, specifically duplications that can skew the ratio of allelic associations and found two duplications of 750 and 89bp (nssv16165643, chr14:98,934,427-98,935,179 and nssv16173610, chr14:97,441,932-97,442,020), that were not associated with any genes and a 304-bp duplication in an intron of the *VRK1* gene (nssv16177248, chr14:97,284,401-97,284,704), listed in the NCBI database. In our cell lines, we find no variation in copy number within the v-TAD domain. However, both GM12878 and H1-hESC possess different Indel mutations within the region corresponding to the nssv16165643 duplication. It is unclear to what extent these variants are responsible for allele-specific associations in the v-TAD. Since they do not overlap CTCF sites, we do not anticipate they are responsible for such large-scale allelic changes in GM12878 and H1-hESC.

VRK1 encodes a Serine/Threonine Kinase and is associated with pontocerebellar hypoplasia, Type 1A and Microcephaly-Complex Motor and Sensory Axonal Neuropathy Syndromes. It is widely expressed in several tissues and has roles in cell cycle, mitosis and DNA damage responses. It has never been reported to have monoallelic expression. We found it to be highly expressed in all cell lines (Additional file 2: Fig. S6b) and monoallelic in H1-hESC (Fig. 3c). The v-TAD region in H1-hESC seemed to form several cross-TAD interactions with adjacent TADs, and further genes (*PPP4R4* and *DICER1*) were found to have allele-specific expression (Fig. 3c).

In summary, these results indicate that imprinted regions can have allele-specific associations confined within TADs, without differences in compartmentalization such as at the *IGF2-H19* and *KCNQ1* loci. We have also seen that compartmentalization can be detected allele-specifically and that imprinted regions when present in active A-compartments can form looping associations that extend beyond their own TAD regions, with the potential of allele-specifically activating genes outside the imprinted locus.

Clusters of allele-specific interactions occur throughout the genome as allele-specific TADs

The detection of the above v-TAD prompted us to examine the frequency in which differences in allele-specific associations can be found within TADs genome-wide. We therefore performed an unbiased ranking of all TADs to assess the allelic association differences genome-wide in GM12878, IMR-90, and H1-hESC cells (Fig. 4a). We defined allele-specific TADs (ASTADs) as having higher than expected absolute differences in A1 and A2 associations. TADs containing imprinted genes (Additional file 3: File S1) had *Z*-scores of 2.9–5.8 (*IGF2-H19*, in all three cell lines), 5.5–9.1

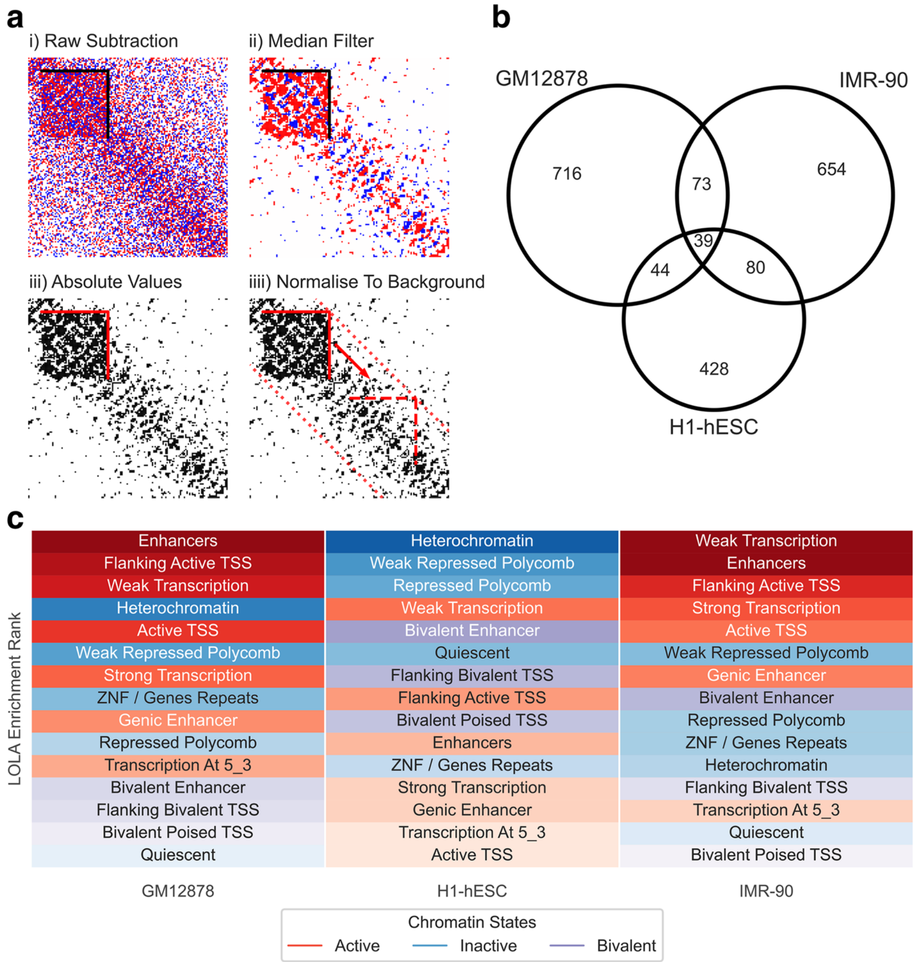

Fig. 4 Several TADs genome-wide are enriched for allele-specific associations. **a** Illustration of ASTAD detection methodology. (i) Reference TAD domains are aligned with the raw Hi-C subtraction matrix. (ii) A median filter is applied to remove background noise and emphasize regions of consistent directional bias. (iii) The absolute sum of intra-domain allelic differences is calculated. (iv) A *Z*-score is calculated by comparing against the chromosome-wide background level of absolute differences for a domain of equivalent size. TADs with *Z*-score > 2 are considered ASTADs. **b** Venn diagram of conserved ASTADs between cell lines. Conserved ASTADs were defined as any set of domain intervals, between cell lines, that shared 90% reciprocal overlap. **c** ASTAD enrichment across different chromatin states. Chromatin states are ordered, per cell line, according to their enrichment level as determined by LOLA (max rank). ASTADs possess contrasting enrichment characteristics in H1-hESCs compared the differentiated cell lines. IMR-90 and GM12878 were significantly enriched in active chromatin relative to all TAD domains. H1-hESC were enriched for inactive and bivalent states

(*DLK1-DIO3* in IMR-90), and the *SNRPN* locus (2.2-3.1 in H1-hESCs). The v-TAD described above had a *Z*-score >4 in GM12878 and H1-hESC.

Each cell line had its own unique profile of ASTADs distributed across the genome (Additional file 2: Fig. S8a and b), although there was a small overlap between cell lines (Fig. 4b). Allelic imbalances due to abnormalities resulting in trisomy show up as large blocks of allelic bias, and this was found at the 11q chromosomal region in GM12878 cells (Additional file 2: Fig. S8c). Imbalances due to monosomy will be masked as these regions cannot be phased. Random monoallelic effects such as X-inactivation are not expected to show up as ASTADs because most normal tissues have a 50% mix of cells with either of the parental X-chromosomes inactivated. Indeed, in IMR-90 cells where X-inactivation is random (presumably because it was derived from primary lung tissue rather than cloned from a single cell), no allelic bias was found on the X-chromosome (Additional file 2: Fig. S8d). GM12878 cells, in comparison, have skewed X-inactivation [44], and a large number of allele-specific associations on the X-chromosome (Additional file 2: Fig. S8d). The H1-hESC cell line is male and thus these cells do not register heterozygosity for phasing on the X-chromosome. A full description of all ASTADs detected is available in Additional file 4: File S2. Overall, ASTADs vary in size comparable to non-ASTADs (Additional file 2: Fig. S8e).

To determine whether ASTADs were associated with specific chromatin (heterochromatin compartments, DNAse I sensitivity, histone signatures associated with regulatory elements) or genomic (heterozygous SNPs, lncRNA, CpG islands) features, we carried out locus overlap analysis (LOLA) [67] for 15 states previously imputed using chromHMM [68]. This revealed that the ASTADs had different characteristics in H1-hESCs compared the differentiated cell lines. IMR-90 and GM12878 are significantly enriched in active chromatin relative to all TAD domains (Fig. 4c). H1-hESC were enriched for inactive and bivalent states.

CNVs are known to influence chromatin architecture [69] and therefore the detection of ASTADs. CNVs may also influence mappability to the reference genome and introduce artifacts in variant discovery and haplotype phasing. Although HiCFlow implicitly removes CNV-driven biases through HiCcompare, we carried out an independent analysis of CNV using QDNAseq [70] to assess whether ASTADs were associated with non-normal copy numbers. With the exception of a low magnitude, but significant, enrichment of gain CNV in IMR-90, non-normal CNV regions were not substantially over-represented in ASTADs relative to TADs (Additional file 1: Tables S5 and S7).

Conserved ASTADs

The three cell lines tested came from three different individuals and represent three different cellular lineages (lymphoblastic, fetal lung, and embryonic stem cells). Based on these genetic and the tissue differences, we expect only a few ASTADs would be common to all three cell lines. From the above analysis, we found 39 conserved ASTAD. We hypothesized that common ASTADs would fall into two categories. The first category being sequence directed, such that ASTADs present at the same location in the three cell lines would share identical SNPs variants. The second category being ASTADs with stable epigenetic mechanisms directing allele-specific chromatin conformation as in genomic imprinting.

We first tested whether conserved ASTAD boundaries possessed the same, i.e., identical genetic variants, across all three cell lines that may influence chromatin structure in a consistent manner. Randomization testing with repeat random sampling ($n = 10{,}000$) of conserved TAD versus ASTAD boundaries (see Methods, Additional file 1: Table S8), indicated significant enrichment (Z-score = 2.13, $p = 0.016$) of identical genetic variants at conserved ASTAD boundaries compared to conserved TADs.

A similar analysis, examining the distribution of allele-specific methylation (ASM) within conserved TAD versus ASTAD boundaries was performed for each cell line. We did not detect any significant enrichment of in H1-hESC and IMR-90. Enrichment was marginally higher in GM12878 (Z-score = 1,67, $p = 0.048$). These results suggest that conserved ASTADs are primarily sequence directed due to sharing similar haplotypes.

The conserved ASTAD with the highest ranking allelic difference (Z-score = 4.9–9.7), was identified at chr3:195,270,000–195,730,000 (Additional file 5: File S3). However, this region was found to overlap an ENCODE Blacklist region usually associated with anomalously high read mapping signal [71], possibly due to unannotated repeats in the reference gene sequence. We therefore cannot exclude this ASTAD being an artifact. The next highest ranking conserved ASTADs (chr12:52,530,000-53,390,000, Z-score = 3.6–5.3 and chr7:2,910,000–4,810,000, Z-score = 3.6–5.6) are not within blacklisted regions.

The ASTAD at the chr12:52,530,000–53,390,000 region contains a cluster of keratin type II cytoskeletal orthologues, involved in hair and epithelial keratin synthesis, and a high association with disease-associated variants. This region has been described as an EAFD locus (genetic variants with extreme allele frequency differences) and a feature of such loci are that the SNPs have longer linkage disequilibrium (LD) ranges than random SNPs [72]. *KRT1* has been reported to be expressed allele-specifically as a result of cis-regulatory polymorphisms [73], and a more in-depth analysis has shown allele-specific expression is a complex trait of multiple SNPs having a cumulative effect on gene expression [74]. In GM12878, IMR-90 and H1-hESCs, none of the *KRT* genes were listed as having allele-specific expression, despite this region showing strong allelic differences in association frequencies (Additional file 2: Fig. S9).

Imprinted genes in ASTADs

Only 5 Imprinted genes are present in conserved ASTADs. Four of these (*H19*, *IGF2*, *IGF2-AS* and *INS*) are part of the same imprinted gene cluster on chromosome 11, the other is the recently identified pseudo-gene *ATP5F1EP2* [75] on chromosome 13 (Additional file 3: File S1). The *IGF2-H19* cluster is the most studied of imprinted genes, and perhaps this is due to its robust and stable CTCF-mediated imprinting mechanism. Little is known about imprinting mechanisms for *ATP5F1EP2.* However, a close look at this locus showed allele-specific expression of further genes within this ASTAD, including *POLR1D*, *MTIF3*, and *USP12* in H1-hESC, and *RPL21* in GM12878. These genes have not been reported to be imprinted. Gene mutations in *POLR1D* underlie autosomal dominant inheritance of Treacher Collins Syndrome (TCS, OMIM 154500).

In our targeted analyses, we found that the *KCNQ1*, *SNRPN*, and *DLK1* loci had several differences between the cell lines and therefore unlikely to be within conserved ASTADs (Additional file 3: File S1). We screened 115 genes reported to be imprinted in humans [76], to determine if they were within ASTADs in any of the cell lines as opposed to being within a conserved ASTAD (Additional file 3: File S1). In H1-hESC, 45 (39%) of the imprinted genes are in ASTADs. IMR-90 and GM12878 have 38 (33%) and 42 (37%) respectively.

Randomization testing (see Methods) revealed that imprinted genes were significantly enriched ($p \leq 0.001$) within ASTADs compared to non-ASTADs (Additional file 1: Table S9). However, the observation that so few imprinted genes were in conserved ASTADs suggest that somatic differences in imprinted gene expression, methylation, and other epigenetic effects influence the density of allele-specific interactions such that their TADs do not consistently meet the threshold of a conserved ASTAD.

Allele-specific gene expression in ASTADs

ASTADs are domains with high frequency of allele-specific contacts. Therefore, we examined whether genes located within ASTADs have allele-specific expression and downloaded RNAseq ASE data for GM12878 (3099 biallelic, 480 monoallelic) [77] and IMR-90 (409 monoallelic) / H1-hESC (2398 monoallelic) [78]. Only the GM12878 dataset included genes with confirmed biallelic expression. In GM12878, 153 of 480 (32%) of ASE genes were found to be within ASTADs. For IMR-90, this was 73 out of 409 (18%) and for H1-hESC, it was 182 out of 2398 (8%). Randomization testing (see Methods) revealed that ASE genes were significantly enriched ($p < 0.001$) within ASTADs in GM12878 and IMR-90, but not in H1-hESC (GM12878 (Z-score $= 4.64$, $p = 1.7\text{e}{-6}$), IMR-90 (Z-score $= 3.43$, $p = 0.0003$) and H1-hESC (Z-score $= -1.78$, $p = 0.962$).

It has been suggested that polymorphisms within enhancers are more likely to disrupt chromatin architecture and influence gene expression. We therefore assessed whether the heterozygotic variation in enhancers (using public available data from [79]) are enriched in ASTADs and found a significantly higher than expected proportion of heterozygous variants in enhancers overlapping ASTADs (chi-square test, $p < 0.001$) in GM12878 and IMR-90, but not in H1-hESC. In addition, we find that enhancers associated with ASE genes are significantly over-represented in ASTADs (chi-square test, $p < 0.001$) in GM12878 (Additional file 6: File S4).

We further found that a cluster of *TRIM* genes on chromosome 11 contained allele-specific (paternal allele) expressed genes (*TRIM5*, *TRIM22* in GM12878, *TRIM6*, *TRIM6-TRIM34* in IMR-90 and *TRIM34*, *TRIM6-TRIM34* in H1-hESC) and were within an ASTAD in GM12878, and in IMR-90 (Additional file 2: Fig. S6). The ASTAD containing the *TRIM* cluster also contains a cluster of olfactory receptor genes (*OR52 -OR56*), which typically express only one allele, but did not feature in the lists of ASE in these cell lines. Olfactory genes have been shown to form interchromosomal associations and aggregate in foci within the nucleus when they are repressed, with the expressed allele localized outside of such foci [80, 81].

One ASTAD region of interest included the *TAS2R* gene cluster that encodes an array of Bitter Taste Receptor genes on chromosome 12p13.2. The *TAS2R* gene cluster overlaps an ASTAD in both GM12878 (chr12:10,915,000 - 11,405,000) and in IMR-90 (chr12:10,900,000–11,370,000) corresponding to an overlap of approximately 90%, but this may be due to a lack of SNP density in IMR-90 (Fig. 5). The ~400kb ASTAD seems to originate from a weak CTCF binding site as a subTAD within a 800-kb-wide CTCF defined TAD. We further found that the region had allele-specific differences in Cscores,

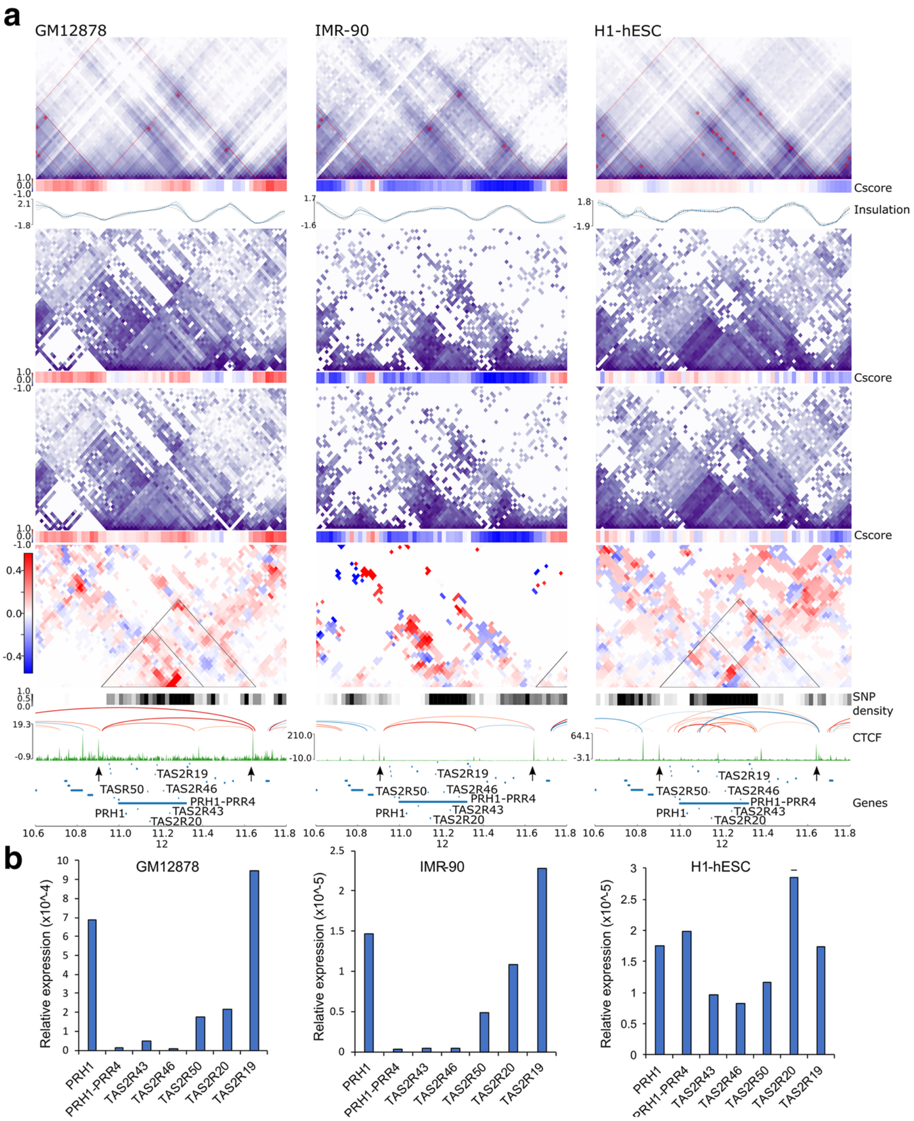

Fig. 5 The Bitter Taste Receptor (*TAS2R*) cluster on Chromosome 12p13.2 is within an ASTAD. **a** Contact matrices for diploid, haplotype phased alleles and their subtraction matrices at 10kb resolution in GM12878, IMR-90, and H1-hESC. Arrows below the CTCF track indicate boundaries of the ~800kb TAD which hosts the ASTAD (as a subTAD). ASTAD shown as triangles. The ASTAD is identified in GM12878 (chr12:10,915,000–11,405,000, *Z*-score = 4.06) and IMR-90 (chr12:10,900,000–11,370,000, *Z*-score = 4.20) with >90% overlap. H1-hESC was not called as an ASTAD (*Z*-score = 1.56). Cscores below the contact matrices show clear allelic differences that varied between cell lines. **b** Quantitative RT-PCR analysis of RNA transcript levels for *TAS2R*s genes and the lncRNA *PRH1-PRR4*

such that in H1-hESC, one allele was more enriched for the A-compartment, while the other was divided into several smaller A- and B-compartments (Fig. 5a, right). In GM12878, there was allelic variation within A-compartments, whereas in IMR-90 there was allelic variation within B-compartment. *TAS2R* genes have not previously been identified as imprinted or to have allelic expression. We examined RNA transcript levels for several of the *TAS2R*s genes at the locus by PCR and confirmed that these were expressed in all three cell lines (Fig. 5b). These genes are usually very small single exon genes and despite being in a region of high sequence variability, most of the SNPs are intergenic. Even the lncRNAs, of which there are several at the locus, have very small exons and thus we found no informative SNPs to enable us to verify whether genes at this locus have monoallelic expression. Taste receptors, like olfactory receptors, are G-protein coupled receptors and they may similarly have monoallelic expression due to allelic exclusion. We examined the *TASR2* clusters on chromosomes 5 (*TAS2R1*) and 7 (*TAS2R3-TAS2R38)* and found that these did not overlap ASTADs. However, to our knowledge this is the first time that taste receptors on chromosome 12 have been reported to be present in an allele-specific chromatin conformation.

Discussion

The recent attempts to link chromatin interaction information with GWAS variants to target genes has led to several tools being developed to predict the functional effects of variants in disease [13]. However, aside from allele-specific differences at imprinted loci, the occurrence of allele-specific TAD structures and their association with allele-specific gene expression has not been extensively documented at a genome-wide level.

In this study, we assembled a novel bioinformatics pipeline, HiCFlow, which combines variant calling and haplotype phasing with allele-specific Hi-C analysis to enable robust investigation and visualization of allele-specific associations. Since genomic imprinting is a phenomenon of epigenetically established parent-of-origin monoallelic gene expression, that is independent of the genetic sequence of the expressed or silenced allele, we initially focused on these genes as domains known to be allele-specifically regulated.

We focused on three imprinted gene clusters that exemplified imprinted domains known to be allele-specifically regulated. The canonical boundary model exemplified by *IGF2-H19* is consistent with an allele-specific loop extrusion model, which shows up as a pair of parallel stripes in a subtraction matrix. In this case, CTCF sites anchored at the HIDAD locus (or other sites downstream in a tissue-specific manner) interact with CTCF sites at the *H19*-DMR on the maternal allele, and with CTCF sites near the *IGF2* promoter on the paternal allele.

The key finding at imprinted loci is that although the canonical boundary model for regulating imprinted genes expression as exemplified by *IGF2-H19* is consistent with an allele-specific loop extrusion chromatin conformation model, this is not invariably the rule at all imprinted loci. Allele-specific loop extrusion shows up as a pair of parallel stripes in a subtraction matrix at the *IGF2-H19* locus and it is the *H19*-ICR that determines the allele-specific chromatin scaffold. Here, CTCF sites at the HIDAD region and 9 other intervening CTCF pause-sites engage in loop extrusion, either with the CTCF-rich *H19*-DMR on the maternal allele, or with CTCF sites near the *IGF2* promoter on the paternal allele. A large body of literature exists for the *IGF2-H19* locus, demonstrating that deletion of the ICR, or its inactivation through DNA methylation, results in loss of imprinting, with over expression of *IGF2* causing BWS [21, 49, 82–86]. It has previously been shown by us and others that allele-specific deletion of CTCF sites in the ICR results in allele-specific conformational changes and that cells from BWS patients with loss of imprinting have allele-specific profiles consistent with “paternalized” conformation structures and that perturbation of methylation profiles using 5-Azacytidine results in restructuring of the chromatin conformation of this locus [21, 49, 82–85]. Thus, the impact of deleting/modifying the ICR on chromatin conformation is known for this locus. In contrast, the adjacent imprinted gene cluster clearly shows that the ICR (KvDMR) regulating imprinting at the *KCNQ1* locus has weak if any effects on 3D chromatin structure. If there is allele-specific loop extrusion, this occurs at a CTCF site previously identified as "region 3" and shown to be required for setting up a maternal allele-specific loop to facilitate *KCNQ1* expression [21]. The paternally expressed *KCNQ1OT1* lncRNA transcript is regulated by the KvDMR, and its transcription may disrupt CTCF binding at region 3 on the paternal allele to prevent this loop. RNA polymerase 2 has previously been reported to displace CTCF occupancy [87]. However, we have little evidence that this may be the case at the *KCNQ1* locus, and even if it did, the ICR would still have an indirect effect on 3D structure at this locus.

Phase condensation models have not previously been tested at imprinted gene loci. Data from chromatin immunoprecipitation for post-translational histone modification profiles at imprinted loci in mouse and humans [88] predict that the silent allele should be in a heterochromatic configuration. Early studies using fluorescent in situ hybridization at the *SNRPN* locus demonstrated that active and silent alleles occupy different nuclear compartments [89]. Adding the Cscore to the HiCFlow pipeline enabled us to confirm allele-specific differences in compartments at imprinted loci. These differences were undetectable at the *IGF2-H19* and the *KCNQ1* loci in the cell lines tested. Strong differences in compartmentalization were observed at the *DLK1* locus in IMR-90 cells. This correlated with ICR methylation levels that would be consistent with normal imprinting and strong allele-specific association differences. A striking allelic difference in compartment structure was detected at the *SNRPN* locus in H1-hESC which also showed clear evidence of CTCF occupancy at the locus. Current models suggest that CTCF stabilizes TAD structures and that compartmentalization processes counteract the formation of TADs [90]. Our observations at the *SNRPN* locus suggest that CTCF stabilized higher-order structures may also be initiated by phase condensation-mediated compartmentalization. Indeed, recent analysis suggests that CTCF may have an instructive function in the formation of condensates [90]. Within imprinted domains, there are also non-imprinted genes that escape the effects of neighbouring gene silencing and we do not yet understand how imprinting is contained locally and why it does not spread across an entire chromosome. By examining the loci from a wider perspective, we found that each locus occurred within a larger TAD structure, which is similar to observations made by the Feil group in mice for the *Igf2* and *Meg2* loci [43]. The loci we examined suggest that imprinted genes are present within interacting subTADs with weak boundaries. We showed evidence that allele-specific looping associations in imprinting domains that are present in A-compartments can extend into neighbouring TADs. Such observations are rare, and it is feasible that this "neighbor effect" requires that the non-imprinted neighbor is already in an active state and able to be allele-specifically upregulated by associations with enhancers within the imprinted domain.

Our novel HiCFlow pipeline is suitable for processing multi-sample Hi-C, RC-Hi-C and Micro-C datasets. It is implemented as a user-friendly Snakemake pipeline and, to our knowledge, is the first workflow to combine haplotype phasing with allele-specific-Hi-C analysis. Testing it on a RC-Hi-C dataset provides proof-of-concept data for a diagnostic assay that can be used to detect allele-specific chromatin conformation in humans with imprinting disease. In this capture data set, we only examined a small set of imprinted gene clusters, but these could be expanded to the full set of imprinted genes and would still provide the required read depth when sequenced in-house on standard sequencing platforms. A limitation of our RC-Hi-C analysis was that we focused on the core regions of the imprinted domains, which enriched for the detection of shorter-range interactions. We would recommend utilizing tools, such as CHiCANE and Peaky, to detect long-range interactions [91, 92]. For the purposes of this study, our focused analysis was sufficient to show that different chromatin conformation structures are present at imprinted loci, rather than a "standard ICR-centered" structure. In the last decade, it has been shown that a subset of patients diagnosed with an imprinting disorder, have multi-locus imprinting disturbances (MLID), characterized by loss of methylation at multiple imprinted loci across the genome (reviewed [15]). RC-Hi-C and HiCFlow will be useful tools in the comprehensive and integrative analysis of MLID.

To investigate the properties of genome-wide allele-specific interactions, we scored each TAD based on the absolute allelic differences observed. The set of top-ranked TADs, denoted "allele-specific TADs (ASTADs)," were assessed for over-representation of various genomic annotations relative to non-ASTADs. Most strikingly, ASTADs were found to be enriched with polymorphic variants. This is unsurprising since heterozygous SNPs are necessary to distinguish alleles during allelic assignment of read pairs. However, enrichment of INDELs in ASTADs, which were not used for allelic assignment, suggests that high genetic variability plays a role in influencing allele-specific chromatin conformation. Indeed, regions of high variability are prone to allele-specific gene expression as demonstrated by the large body of GWAS studies that correlate genetic variants with allele-specific binding of transcription factors, DNA methylation patterns and gene expression (reviewed [93]). This is further supported by our finding that heterozygous variants are more likely to be associated with ASTADs if they overlap a known enhancer.

We also found that allele-specific expressed genes were significantly over-represented in ASTADs in GM12878 (32%) and IMR-90 (18%), although not in H1-hESC (8%). Despite enrichment, only a small absolute proportion of ASE genes overlapped ASTADs. In some cases, it is likely that short-range allele-specific interactions may be indistinguishable at the 20kb resolution of our data. In addition, the absence of informative SNPs can prevent ASE detection and detection of allele-specific chromatin interactions.

As expected, imprinted genes were significantly over-represented in ASTADs compared to non-ASTADs. However, only the *H19/IGF2* locus was consistently identified as an ASTAD in all cell lines. Allele-specific associations at imprinted loci also did not correlate with levels of detectable transcripts. It has been suggested that the chromatin conformation forms a scaffold upon which transcription factors can dock and activate gene expression [94]. Such a scaffold would therefore not correlate with expression levels if the right transcription factors are not present. However, while this may be true for some loci, given the overall conserved CTCF binding across multiple tissues and cell types, there is enough variation in TAD structures, and reported experimental evidence indicating that active transcription modulates higher-order chromatin structures [95]. A lack of correlation between looping associations and transcripts could therefore be due to the post transcriptional effects that affect RNA stability. The strength of chromatin loops between an enhancer and promoter and transcription factor binding kinetics has recently been correlated with transcriptional bursting [96, 97]. Thus, the frequency and length of time that a gene promoter and enhancer interact affects the frequency and length of a transcriptional burst. Technologies in which such models can be tested experimentally on a genome-wide scale are not yet available.

We were able to confirm that regions of high variability and known to have allelic imbalances in expression such as the olfactory receptor genes were within ASTADs. The *TAS2R* family of receptors similar to olfactory receptors are G-protein coupled receptors. We were intrigued to find that the bitter taste receptor (*TAS2R*) clusters were present within ASTADs. To our knowledge, *TAS2R* genes have not been reported to be subject to allelic exclusion, unlike olfactory receptors. They occur in regions of high genetic variability similar to olfactory receptors. There are about 25 functional *TAS2R* genes and 11 pseudogenes spread between chromosomes 5, 7, and 11. Extensive population studies have utilized the high variation to examine evolutionary origins of different haplotypes and to identify the selection pressures that an ability to distinguish bitter toxic substances has had on the genetic evolution of this gene family. The functional effects of the variants on G-protein receptor protein structures and the mechanisms whereby they convey taste perception to the brain have been elucidated. However, limited information on their transcriptional regulation exists. It has been recently shown that *TAS2R* genes are not only expressed on the tongue but that they are more widespread and present in heart and respiratory epithelia as well as in the gut and that they may have further sensing functions unrelated to bitter taste. An in situ hybridization study has indicated that humans co-express a heterogeneous mix of between 4 and 11 taste receptors per cell in papillae of the tongue. Although we do not know if this is due to allele-specific expression, our data indicate that this may be the case as ASTADs are commonly associated with allele-specific expression.

Conclusions

This study exemplifies the utility of the bioinformatics HiCFlow tool for combined variant calling and haplotype phasing with allele-specific Hi-C analysis for investigation of allele-specific associations at regions subjected to epigenetic silencing such as genomic imprinting as well as sequence-mediated influences on expression. Overall, this study highlights how genetic sequence variation and the regulatory mechanisms behind allele-specific gene regulation culminate in widespread allelic differences in chromatin organization that are not confined to imprinted gene loci.

Methods

Cell lines

Human mammary epithelial cell (1-7HB2) line were purchased from the ECACC (catalog no 10081201) (Culture Collections). 1-7HB2 were cultured in RPMI-1640 (Sigma-Aldrich), supplemented with 5% fetal bovine serum (Sigma-Aldrich, B6917), 10 ml/l penicillin-streptomycin solution (Gibco, 15140122), 5 μg/ml insulin (Sigma-Aldrich, I0516), and 1 μg/ml hydrocortisone (Sigma-Aldrich, H0888 – 1G). GM12878 cells (B-Lymphocyte) were purchased from Coriell Institute and cultured in RPMI-1640 (Sigma-Aldrich) supplemented with 15% fetal bovine serum (Sigma-Aldrich, B6917). IMR-90 cells (normal lung tissue derived from a 16-week-old female) were obtained from ATCC (CCL-186™). Cell lines were cultured at 37°C. All cell lines have been periodically tested in-house for mycoplasma contamination.

RC-Hi-C library preparation

3–4×10^7 number of 1-7HB2 cells were crosslinked on plate with formaldehyde (Agar Scientific R1026), followed by a quenching step with 1.25M glycine, scraping for cell detachment and two washes with cold PBS 1×. The cell pellet was re-suspended in 50ml freshly prepared ice-cold lysis buffer (10mM Tris-HCl pH 8, 10mM NaCl, 0.2% Igepal CA-630 (Sigma-Aldrich, I8896-50ML), one protease inhibitor cocktail tablet (Roche complete, EDTA-free 11873580001)). Cells were lysed on ice for a total of 30min, with 2 × 10 strokes of a Dounce homogenizer with a 5-min break between Douncing to minimize cell clumping. Following lysis, the nuclei were pelleted and washed with cold 1.25xNEB Buffer 2 (NEB, B7002S) then re-suspended in 1.25xNEB Buffer 2 to make two aliquots of 10–15×10^6 cells for digestion. Hi-C libraries were digested using 1500U MBOI (NEB, R0147M) at 37°C overnight while orbital shaking. Following digestion, the restriction fragment overhangs are filled in for 1h with dNTPs including biotin-14-dATP (Life Technologies, 19524-016). Fragments are then blunt-end ligated with 1U/μl T4 DNA ligase (Invitrogen, 15224-025) under dilute conditions to favor ligation between crosslinked fragments (in a 15-ml tube for overnight at 16°C). DNA crosslinks were then reversed with 10mg/ml proteinase K (Roche, 03115879001) for 6–8 h at 65°C followed by RNase A (Roche, 10109142001) treatment at 37°C for 60 min. Two rounds of DNA extraction/purification were carried out with phenol pH 8.0 (Sigma-Aldrich, P4557) and phenol: chloroform: isoamyl alcohol (Sigma-Aldrich, P2069) followed by precipitation with 3M sodium acetate pH 5.2 (Sigma-Aldrich, S7899) and 2.5× volume of ice-cold 100% ethanol on wet ice for 1–2 h. The Hi-C Library's quantity and quality were assessed by running 50–100ng of the Hi-C libraries on a 0.8% agarose gel. Hi-C marking and Hi-C ligation efficiency was verified by PCR digest assay using MBOI and CLAI (NEB, R0197S) enzymes. Biotin is then removed from the ends of un-ligated fragments using the exonuclease properties of T4 DNA polymerase (NEB, M0203L), and DNA was sheared to obtain DNA fragments with a peak concentration around 400 bp. DNA ends were repaired and fragments, with internally incorporated biotin, are pulled down using magnetic Dynabeads MyOne Streptavidin T1 beads (Life Technologies 65601). After PE adaptor ligation ((5′-P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′ and 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3), pre-Capture amplification was performed with eight cycles of PCR on multiple parallel reactions from Hi-C libraries immobilized on Streptavidin beads, which were pooled post PCR and SPRI Ampure XP beads (Beckman Coulter, A63881) purified. The final Hi-C library was re-suspended in 25μl of Tris low-EDTA and quantified by the Qubit™ dsDNA BR Assay Kit (Thermo Fisher, Q32853). The size distribution of the library was assessed by Tapestation D1000 (Agilent). The Hi-C capture regions were enriched via hybridization with biotin-RNA probes (Agilent Technologies). The capture regions of interest were then pulled down with Dynabeads MyOne Streptavidin T1 beads (Life Technologies 65601) and purified using SPRI Ampure XP beads (Beckman Coulter, A63881). Finally, Hi-C Capture library was amplified and then sequenced with Illumina 50 bp paired-end sequencing.

Capture biotinylated RNA oligos design

Capture biotinylated 120-mer RNA oligos (25–65% GC, <3 unknown (N) bases) were designed to target either one or both sides of MBOI site and within 4–500bp as close as possible to the ends of the targeted restriction fragments using a custom genome-wide Perl script made available from the Babraham Institute and then submitted to the Agilent eArray software (Agilent) for manufacture.

RNA extraction and qPCR

RNA was extracted from cell lines using TRI Reagent (Sigma-Aldrich) and 1μg of total RNA was converted to cDNA using QuantiTect Reverse Transcription Kit (Qiagen). Quantitative PCR (qPCR) was performed using a ¼ dilution of cDNA with SYBR Green PCR Master Mix (Thermo Fisher Scientific) on ABI Step One Plus (Applied Biosystems) and specific primers (Sigma-Aldrich) for target genes (see Additional file 1: Table S6).

Public Hi-C data

Human Hi-C data for GM12878 [98, 99], IMR-90 [98, 99], and H1-hESC [100, 101] was downloaded from the 4D Nucleome project [33, 44].

Public gene lists

Allele-specific gene expression data for IMR-90 and H1-hESC [102–108] were obtained from the UCSD Human Reference Epigenome Mapping Project [78] via the Allele-specific Methylation database (ASMdb) [109]. Allele-specific expression data for GM12878 (Additional file 1: Table S2) were obtained from previously published work [77, 110]. Imprinted gene list was obtained from https://www.geneimprint.com. All genes in our study were associated with their Ensembl ID in Gencode v38 (GRCh37) [111]. Genes with ambiguous or unknown Ensembl mapping were excluded from the analysis.

Other public datasets

Allele-specific methylation datasets for GM12878 [112], IMR-90 [113], and H1-hESC [112] (Additional file 1: Table S3) were downloaded from ENCODE [114] and the ASMdb [109]. Publicly available CTCF ChIP datasets were downloaded for GM12878 [115–117], IMR-90 [117, 118], and H1-hESC [117, 119] as described in "Availability of data and materials" section. CTCF ChIP dataset for 1-7HB2 was downloaded from ERX115548, Illumina [120, 121]. Chromatin state data was obtained from the Roadmap Epigenomics Project [78, 122]. This dataset represents a core 15-state chromatin state model, built using ChromHMM (v1.10.0) [68], based on 5 epigenetic marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me) (see Additional file 1: Table S4). Chromatin loop data processed by Peakachu [53] were downloaded from the 3D Genome Browser [123].

Data processing

Hi-C data was processed using our in-house pipeline, HiCFlow. Read adapters were trimmed, using Cutadapt (v3.5) [124] and truncated using HiCUP (v0.7.4) [125] to remove sequences overlapping putative ligation sites. Processed reads were mapped independently to the GRCh37/h19 reference assembly using Bowtie2 (v2.4.4) [126]. The GRCh37 reference assembly was chosen as we observed mappability issues at the *IGF2-H19* locus in GRCh38. Since this locus is an essential control, we have herein presented all results using the GRCh37/hg19 reference. Alignment files were re-merged to paired-end files using Samtools (v1.1.0) [127]. Reads were deduplicated and processed to raw contact matrices using HiCExplorer (v3.7.1) [128]. Finally, contact matrices were corrected using the KR balancing algorithm [129].

For allele-specific analysis, a phased haplotype for IMR-90 and H1-hESC was generated from the raw Hi-C data. Variants were called using the GATK (v4.2.4.1) best practises pipeline [130]. In brief, base quality scores were recalibrated using GATK BaseRecalibrator before potential variant sites were called using GATK HaplotypeCaller. Following this, joint genotyping was performed using GATK GenotypeGVCFs. The quality of raw variant calls was scored against a set of high-confidence variants obtained from the GATK Resource Bundle using GATK VariantRecalibrator. Finally, low-quality variants were filtered using a truth sensitivity filter of 99.5%. Haplotype assembly was then performed using HapCUT2 (v1.3.2) [131]. For GM12878, a phased haplotype was obtained from the Platinum Genomes phased variant truthset [132, 133]. The total number of phased variants identified in each cell line were as follows: IMR-90 (2,231,685), H1-hESC (1,643,225), and GM12878 (2,147,688). Prior to alignment, the reference genome was masked at site of phased variants using BEDTools (v2.29.2) to avoid reference bias during mapping [134]. Finally, allelic assignment of reads was performed using SNPsplit (v0.5.0) [135]. Visualizations were created using pyGenomeTracks (v3.6) [136]. The total number of informative valid pairs identified in each cell line were as follows: IMR-90 (0.13e9), H1-hESC (0.44e9), and GM12878 (0.85e9).

Explanation of subtraction matrices

Visual comparison of allelic matrices (A1 vs. A2) was performed using "subtraction matrices." Joint-normalization of raw A1 and A2 matrices was first performed using HiCcompare (v1.6.0) [137]. This provides implicit correction of between-sample bias. Normalized matrices were then transformed using the "Observed / Expected" method to correct for genomic distance and more effectively resolve changes in long-range interactions. Normalized counts were subtracted (A2–A1), and the resulting subtraction matrices were denoised using a median filter (Scikit-Learn v1.7.3) [138]. This emphasizes regions with consistent directional bias and which are more likely to represent signals of interest (Additional file 2: Fig. S1c).

Bioinformatic processing of RC-Hi-C

Processing of RC-Hi-C was as described above, but analysis was restricted to read pairs mapping within a single capture region. All other read pairs were filtered from the analysis.

Quality control

Following sequencing, quality control of the raw FASTQ data sets was performed using FastQC (v0.11.9). Screening for potential sequence contamination was performed using

FastQ Screen (v0.5.2) [139]. Reproducibility of the RC-Hi-C data set was assessed using HiCRep (v1.10.0) [140]. Correlation between biological replicates was high (0.97–0.98 at 5kb resolution). Replicates were therefore merged in downstream analyses to improve resolution for haplotype phasing and allele-specific analysis.

Compartment analysis

HiCFlow performs compartment analysis using CscoreTool (v1.1) [63]. Compartment analysis was performed on the full Hi-C datasets for each cell line at a 20kb resolution. The sign of the Cscore was oriented such that positive and negative scores represented "A-" and "B-" compartments respectively. Results were intersected with chromatin state data such that negative scores were associated with heterochromatin and positive scores were associated with active transcription ("activeTSS").

Allele-specific TAD (ASTAD) classification

TAD domain detection was performed, using OnTAD (v1.2) [141], on the full Hi-C dataset binned at 10kb resolution. The set of TADs were then used as a reference set to identify TADs with substantial differences in contact frequency between allele-specific matrices. A1 and A2 matrices were first jointly normalized using the LOESS method described in HiCcompare. A median filter was then applied to remove spurious or noisy background interactions. Following this, the absolute sum of differences is calculated within the relevant TAD domain. For each TAD, a Z-score is calculated by comparing against the chromosome-wide background level of absolute differences for a domain of equivalent size. The methodology of ASTAD detection is illustrated in Fig. 4a.

Enrichment analysis

To determine if ASTADs were enriched for particular genomic features, we performed enrichment analysis using LOLA (v1.22) [67]. ASTAD enrichment was tested against a background set of all identified TAD domains in a given cell line. To facilitate cell line comparison, only autosomal regions were tested for enrichment. A full list of genomic features tested and enrichment status is available in Additional file 1: Table S5.

Overlap analysis

Overlap analysis was performed to identify ASTADs that were conserved between cell lines. Conserved ASTADs were defined as sets of ASTAD intervals, between cell lines, with at least a 90% reciprocal overlap. A 10% difference in overlap allows for slight error in domain positioning due the loss of resolution during matrix binning. Given a bin size of 20kb, a 10% difference equates to a shift of approximately one bin length for a median size domain interval. Interval overlap was calculated using BedTools (v2.29.2).

Randomization testing

In each of the following randomization tests, any TAD domain overlapping a region with non-normal copy number or overlapping an Encode Blacklist region were removed from the enrichment analysis.

Comparison of identical heterozygous variants in conserved ASTAD boundaries compared to conserved TAD boundaries

Conserved identical heterozygous variants were defined as any heterozygotic variant (SNPs or Indels) present in all three cell lines. The total number of observed conserved variants overlapping the set of conserved ASTAD boundaries was compared against a null distribution of repeat random samples ($n = 10{,}000$) of conserved TAD boundaries. A TAD boundary was defined as the outermost 20kb of a TAD domain, equivalent to 1 bin size on the AS-Hi-C matrices. This analysis was repeated for the allele-specific methylation (ASM) data for each cell line (see Additional file 1: Table S8).

Enrichment of imprinted (or ASE) genes in ASTADs relative to all TAD domains

The total number of observed imprinted genes overlapping ASTADs was compared against an expected distribution of randomly sampled genes. Genes were selected via randomly stratified sampling to match the sample size and distribution of imprinting gene types. Stratification was used due to the imbalance in gene type distributions between the imprinted / ASE gene sets and the total gene sets. Genes were excluded from the analysis if they did not overlap a TAD domain or if they overlapped a blacklisted region or a region with non-normal copy number. A total of 10,000 samples were taken per analysis to build a null distribution and to calculate a Z-score from the observed overlap. This analysis was repeated for each cell line and for ASE genes (see Additional file 1: Table S9).

Methylation status at CpG islands

To check methylation status within regions of interest, preprocessed whole genome bisulfite sequencing (WGBS) data, corresponding to CpG methylation in ENCODE bed bedMethyl format, were obtained from publicly available datasets (GM12878 [142], IMR-90 [143], H1-hESC [144]) as described in "Availability of data and materials" section. Filtering was performed using the methylKit R package [145]. For each cell line, all CpGs overlapping a CpG island were selected. For each target CpG island, boxplot was generated for comparison among three cell lines using the ANOVA statistics.

Bisulfite analysis of allelic methylation

The methylation patterns were studied using a bisulphite conversion kit (EZ DNA Methylation Gold Kit – D5005 Zymo Research). PCR amplification products of the bisulphite template using previously published primers for IG-DMR (hg38_chr14: 100810848-100811276) [25], KvDMR (hg38_chr11:2699867-2700238) [146] and PWS-AS-IC (hg38_chr15:24954788-24955196) [146] were cloned in pGEM-T easy (Promega) and then sequenced with T7 primer. A total of approximately 20–30 clones per DMR per cell line was sequenced. CG methylation was assessed by Multiple Sequence Alignment with CLustalX EMBL-EBI tool.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02876-2.

Additional file 1: Table S1. Capture Regions for RC-HiC in 1_7HB2 Cells. **Table S2.** Allele Specific Gene Expression Data. **Table S3.** Allele Specific Methylation Data. **Table S4.** Core 15-state model (5 marks). **Table S5.** LOLA Enrichment "Max Rank" Score among 20 genomic features. **Table S6.** Oligo primers used in qPCR. **Table S7.** Proportion of ASTAD / non-ASTAD domains overlapping Normal CNV. **Table S8.** Observed vs. Expected Overlap of Features in Conserved ASTADs. **Table S9.** Observed vs. Expected Overlap of Genes in ASTADs.

Additional file 2: Fig. S1. HiCFlow pipeline and Region Capture HiC (RC-HiC) library. **Fig. S2.** Comparison of subtraction matrices, at *IGF2-KCNQ1* locus, between experimental validated and HiCFlow inferred haplotype in GM12878. **Fig. S3.** Supporting information relevant to Fig. 2: DNA methylation data and expression levels of imprinted genes. **Fig. S4.** The effect of the PWS-AS imprinting control region on allele-specific chromatin conformation. **Fig. S5.** The effect of the IG-DMR/*MEG3* imprinting control region on allele-specific chromatin conformation. **Fig. S6.** Compartment analysis of *H19-KCNQ1* and *SNRPN* loci. **Fig. S7.** Compartment analysis of *DLK1-DIO3* locus. **Fig. S8.** Features and distribution of ASTADs. **Fig. S9.** *KRT* gene cluster on chr12 is within a conserved ASTAD.

Additional file 3: File S1. All Gene Statistical Analysis.

Additional file 4: File S2. All TADs Statistical Analysis.

Additional file 5: File S3. Conserved ASTADs Statistical Analysis.

Additional file 6: File S4. Chi-square analysis.

Additional file 7: File S5. Conserved SNPs.

Additional file 8. Review history.

Acknowledgements

We acknowledge Dr Cameron Osborne (King's College London) and Dr Valeriya Malysheva (University of Antwerp, Belgium) for guidance on preparing RC-Hi-C libraries and for critically reading the manuscript, Dr Zongling JI (University of Manchester) for providing H1-hESC DNA and RNA samples, and Dr Simon Andrews (The Babraham Institute Cambridge) for supplying the script to generate Hi-C probes.

Peer review information

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

Review history

The review history is available as Additional file 8.

Authors' contributions

A.M. and L.H. conceived and obtained grant funding for the project and supervised the research. A.M., G.P., and S.R. designed the study, G.P. acquired the experimental data and S.R. carried out bioinformatic and statistical analysis. A.M., L.H., G.P., S.R., and Y.T. analyzed data. S.S. design, troubleshooting, and logistics. A.M., G.P., and S.R. wrote the manuscript. All the authors have read and approved the final manuscript.

Funding

This work has been supported by the Medical Research Council (MR/P000711/1 to A.M. and L.H.), the Leverhulme Trust (RPG-2020-327 to A.M.), and the EPSRC DTP studentship (2106811 to S.R.).

Availability of data and materials

The Region Capture Hi-C datasets that we generated in this work are available in NCBI repository at the accession number PRJNA926951 [147].
Scripts used for downstream bioinformatics analysis are available under MIT license at Github: https://github.com/StephenRicher/HiCFlow [148] and https://github.com/StephenRicher/AS-HiC-Analysis [149]. These scripts are also deposited in Zenodo: https://zenodo.org/record/7563515 [150] and https://zenodo.org/record/6510198 [151].
Further details of the HiCFlow workflow are provided below.

- Project name: HiCFlow
- Project home page: https://github.com/StephenRicher/HiCFlow
- Archived version: 10.5281/zenodo.7563515
- Operating system: Unix-based operating systems
- Programming language: Snakemake (Python)
- Other requirements: Snakemake 7.3.1 or higher, Conda
- License: MIT License
- Any restrictions to use by non-academics: None

Datasets supporting the conclusions of this study include public available Hi-C Data (GSE63525 [98, 99] GSE163666 [100])/ Phased Variant Data (PRJEB338 [133])/ CTCF ChIP Data (GSE30263 [115, 116], GSE31477 [118], GSE29611 [119], PRJEB3073 [121], GSE51334 [117])/ CpG Data ((GSE86765 [142], GSE17312 [143], GSE80911 [144])/ Allele-Specific Expression Data (NA12878 [110], GSE16256 [102–108])/Allele-Specific Methylation Data (GSE40832 [112, 113])/ Chromatin Loop Data (http://3dgenome.fsm.northwestern.edu/downloads/loops-hg19.zip) [53]/ Chromatin State Data (15-core) (https://egg2.wustl.edu/roadmap/web_portal/) [122].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 18 May 2022 Accepted: 13 February 2023
Published online: 03 March 2023

References

1. Dorsett D, Merkenschlager M. Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans. Curr Opin Cell Biol. 2013;25(3):327–33.
2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80.
3. PCAWG Transcriptome Core Group, Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578(7793):129–36.
4. Przytycki PF, Singh M. Differential allele-specific expression uncovers breast cancer genes dysregulated by Cis non-coding mutations. Cell Syst. 2020;10(2):193–203.
5. Buckland PR. Allele-specific gene expression differences in humans. Hum Mol Genet. 2004;13(suppl_2):R255–60.
6. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science (1979). 2012;337(6099):1190–5.
7. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
8. Giambartolomei C, Liu JZ, Zhang W, Hauberg M, Shi H, Boocock J, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics. 2018;34(15):2538–45.
9. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat Commun. 2018;9(1):3121.
10. Javierre BM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016;167(5):1369–84.
11. Gorkin DU, Qiu Y, Hu M, Fletez-Brant K, Liu T, Schmitt AD, et al. Common DNA sequence variation influences 3-dimensional conformation of the human genome. Genome Biol. 2019;20:1–25.
12. Yu J, Hu M, Li C. Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. BMC Genet. 2019;20(1):1–9.
13. Meng XH, Xiao HM, Deng HW. Combining artificial intelligence: deep learning with Hi-C data to predict the functional effects of non-coding variants. Bioinformatics. 2021;37(10):1339–44.
14. Reinius B, Sandberg R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nat Rev Genet. 2015;16(11):653–64.
15. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet. 2019;20(4):235–48.
16. Soejima H, Higashimoto K. Epigenetic and genetic alterations of the imprinting disorder Beckwith-Wiedemann syndrome and related disorders. J Hum Genet. 2013;58(7):402–9.
17. Azzi S, Habib WA, Netchine I. Beckwith-Wiedemann and Russell-Silver Syndromes: from new molecular insights to the comprehension of imprinting regulation. Curr Opin Endocrinol Diabetes Obes. 2014;21(1):30–8.
18. Rabinovitz S, Kaufman Y, Ludwig G, Razin A, Shemer R. Mechanisms of activation of the paternally expressed genes by the Prader-Willi imprinting center in the Prader-Willi/Angelman syndromes domains. Proc Natl Acad Sci U S A. 2012;109(19):7403–8.
19. Ogata T, Kagami M. Kagami-Ogata syndrome: a clinically recognizable upd(14)pat and related disorder affecting the chromosome 14q32.2 imprinted region. J Hum Genet. 2016;61(2):87–94.
20. Prasasya R, Grotheer KV, Siracusa LD, Bartolomei MS. Temple syndrome and Kagami-Ogata syndrome: clinical presentations, genotypes, models and mechanisms. Hum Mol Genet. 2020;29(R1):R107–16.
21. Naveh NSS, Deegan DF, Huhn J, Traxler E, Lan Y, Weksberg R, et al. The role of CTCF in the organization of the centromeric 11p15 imprinted domain interactome. Nucleic Acids Res. 2021;49(11):6315–30.
22. Korostowski L, Sedlak N, Engel N. The Kcnq1ot1 long non-coding RNA affects chromatin conformation and expression of Kcnq1, but does not regulate its imprinting in the developing heart. PLoS Genet. 2012;8(9):e1002956.
23. Horsthemke B, Wagstaff J. Mechanisms of imprinting of the Prader-Willi/Angelman region. Am J Med Genet A. 2008;146:2041–52.
24. Benetatos L, Hatzimichael E, Londin E, Vartholomatos G, Loher P, Rigoutsos I, et al. The microRNAs within the DLK1-DIO3 genomic region: Involvement in disease pathogenesis. Cell Mol Life Sci. 2013;70:795–814.
25. Kagami M, O'Sullivan MJ, Green AJ, Watabe Y, Arisaka O, Masawa N, et al. The IG-DMR and the MEG3-DMR at human chromosome 14q32.2: hierarchical interaction and distinct functional properties as imprinting control centers. PLoS Genet. 2010;6(6):e1000992.
26. Adalsteinsson BT, Ferguson-Smith AC. Epigenetic control of the genome — lessons from genomic imprinting. Genes. 2014;5(3):635–55.

27. Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, et al. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc Natl Acad Sci U S A. 2006;103(28):10684–9.
28. Engel N, Raval AK, Thorvaldsen JL, Bartolomei SM. Three-dimensional conformation at the H19/Igf2 locus supports a model of enhancer tracking. Hum Mol Genet. 2008;17(19):3021–9.
29. Szabó PE, Tang SHE, Rentsendorj A, Pfeifer GP, Mann JR. Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Curr Biol. 2000;10(10):607–10.
30. Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature. 2000;405(6785):486–9.
31. Nativio R, Wendt KS, Ito Y, Huddleston JE, Uribe-Lewis S, Woodfine K, et al. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. Bickmore WA, editor. PLoS Genet. 2009;5(11):e1000739 Available from: https://dx.plos.org/10.1371/journal.pgen.1000739. Cited 2018 Dec 12.
32. Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature. 2000;405(6785):482–5.
33. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, et al. The 4D nucleome project. Nature. 2017;549(7671):219–26.
34. Sexton T, Cavalli G. The role of chromosome domains in shaping the functional genome. Cell. 2015;160(6):1049–59.
35. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381–5.
36. Golfier S, Quail T, Kimura H, Brugués J. Cohesin and condensin extrude DNA loops in a cell-cycle dependent manner. Elife. 2020;9:e53885.
37. Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science (1979). 2015;347(6225):1017–21.
38. Khoury A, Achinger-Kawecka J, Bert SA, Smith GC, French HJ, Luu PL, et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat Commun. 2020;11(1):54.
39. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res. 2016;26(1):70–84.
40. Heurteau A, Perrois C, Depierre D, Fosseprez O, Humbert J, Schaak S, et al. Insulator-based loops mediate the spreading of H3K27me3 over distant micro-domains repressing euchromatin genes. Genome Biol. 2020;21(1):1–9.
41. Witcher M, Emerson BM. Epigenetic Silencing of the p16INK4a tumor suppressor is associated with loss of CTCF binding and a chromatin boundary. Mol Cell. 2009;34(3):271–84.
42. Gentile C, Kmita M. Polycomb repressive complexes in Hox gene regulation: silencing and beyond. BioEssays. 2020;42(10):1900249.
43. Llères D, Moindrot B, Pathak R, Piras V, Matelot M, Pignard B, et al. CTCF modulates allele-specific sub-TAD organization and imprinted gene activity at the mouse Dlk1-Dio3 and Igf2-H19 domains. Genome Biol. 2019;20(1):1–7.
44. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
45. Ye T, Ma W. ASHIC: hierarchical Bayesian modeling of diploid chromatin contacts and structures. Nucleic Acids Res. 2020;48(21):e123.
46. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16(1):1–1.
47. Niemczyk M, Ito Y, Huddleston J, Git A, Abu-Amero S, Caldas C, et al. Imprinted chromatin around DIRAS3 regulates alternative splicing of GNG12-AS1, a long noncoding RNA. Am J Hum Genet. 2013;93(2):224–35.
48. Woodfine K, Huddleston JE, Murrell A. Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. Epigenetics Chromatin. 2011;4(1):1–3.
49. Ito Y, Nativio R, Murrell A. Induced DNA demethylation can reshape chromatin topology at the IGF2-H19 locus. Nucleic Acids Res. 2013;41(10):5290–302.
50. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189.
51. Liyakat Ali TM, Brunet A, Collas P, Paulsen J. TAD cliques predict key features of chromatin organization. BMC Genomics. 2021;22(1):499.
52. Paulsen J, Liyakat Ali TM, Nekrasov M, Delbarre E, Baudement MO, Kurscheid S, et al. Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. Nat Genet. 2019;51(5):835–43.
53. Salameh TJ, Wang X, Song F, Zhang B, Wright SM, Khunsriraksakul C, et al. A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat Commun. 2020;11(1) http://3dgenome.fsm.northwestern.edu/downloads/loops-hg19.zip.
54. Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, et al. The landscape of genomic imprinting across diverse adult human tissues. Genome Res. 2015;25(7):927–36.
55. Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA. Status of genomic imprinting in human embryonic stem cells as revealed by a large cohort of independently derived and maintained lines. Hum Mol Genet. 2007;16(R2):R243–51.
56. Schultz BM, Gallicio GA, Cesaroni M, Lupey LN, Engel N. Enhancers compete with a long non-coding RNA for regulation of the Kcnq1 domain. Nucleic Acids Res. 2015;43(2):745–59.
57. Zhang H, Zeitz MJ, Wang H, Niu B, Ge S, Li W, et al. Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the Kcnq1 locus. J Cell Biol. 2014;204(1):61–75.
58. Fitzpatrick GV, Pugacheva EM, Shin JY, Abdullaev Z, Yang Y, Khatod K, et al. Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. Mol Cell Biol. 2007;27(7):2636–47.
59. Prickett AR, Barkas N, McCole RB, Hughes S, Amante SM, Schulz R, et al. Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. Genome Res. 2013;23(10):1624–35.

60. Lin S, Ferguson-Smith AC, Schultz RM, Bartolomei MS. Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci. Mol Cell Biol. 2011;31(15):3094–104.
61. Demars J, Shmela ME, Khan AW, Lee KS, Azzi S, Dehais P, et al. Genetic variants within the second intron of the KCNQ1 gene affect CTCF binding and confer a risk of Beckwith–Wiedemann syndrome upon maternal transmission. J Med Genet. 2014;51(8):502–11.
62. Kim KP, Thurston A, Mummery C, Ward-Van Oostwaard D, Priddle H, Allegrucci C, et al. Gene-specific vulnerability to imprinting variability in human embryonic stem cell lines. Genome Res. 2007;17(12):1731–42.
63. Zheng X, Zheng Y. CscoreTool: Fast Hi-C compartment analysis at high resolution. Bioinformatics. 2018;34(9):1568–70.
64. Jian X, Felsenfeld G. Large parental differences in chromatin organization in pancreatic beta cell line explaining diabetes susceptibility effects. Nat Commun. 2021;12(1):4338.
65. Hildebrand EM, Dekker J. Mechanisms and functions of chromosome compartmentalization. Trends Biochem Sci. 2020;45(5):385–96.
66. Tierling S, Gasparoni G, Youngson N, Paulsen M. The Begain gene marks the centromeric boundary of the imprinted region on mouse chromosome 12. Mamm Genome. 2009;20(9–10):699–710.
67. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics. 2016;32(4):587–9.
68. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6.
69. di Gregorio E, Riberi E, Belligni EF, Biamino E, Spielmann M, Ala U, et al. Copy number variants analysis in a cohort of isolated and syndromic developmental delay/intellectual disability reveals novel genomic disorders, position effects and candidate disease genes. Clin Genet. 2017;92(4):415–22.
70. Scheinin I, Sie D, Bengtsson H, Van De Wiel MA, Olshen AB, Van Thuijl HF, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Res. 2014;24(12):2022–32.
71. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9(1):9354.
72. Sulovari A, Chen YH, Hudziak JJ, Li D. Atlas of human diseases influenced by genetic variants with extreme allele frequency differences. Hum Genet. 2017;136:39–54.
73. Pant PVK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. Analysis of allelic differential expression in human white blood cells. Genome Res. 2006;16(3):331–9.
74. Tao H, Cox DR, Frazer KA. Allele-specific KRT1 expression is a complex trait. PLoS Genet. 2006;2(6):e93.
75. Santoni FA, Stamoulis G, Garieri M, Falconnet E, Ribaux P, Borel C, et al. Detection of imprinted genes by single-cell allele-specific gene expression. Am J Hum Genet. 2017;100(3):444–53.
76. Ginjala V. Gene imprinting gateway. Genome Biol. 2001;2(8):1–3.
77. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16(12):1297–305.
78. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Roadmap Epigenomics Consortium: integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317.
79. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593(7858):238–43.
80. Markenscoff-Papadimitriou E, Allen WE, Colquitt BM, Goh T, Murphy KK, Monahan K, et al. Enhancer interaction networks as a means for singular olfactory receptor expression. Cell. 2014;159(3):543–57.
81. Clowney EJ, Legros MA, Mosley CP, Clowney FG, Markenskoff-Papadimitriou EC, Myllys M, et al. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell. 2012;151(4):724–37.
82. Freschi A, del Prete R, Pignata L, Cecere F, Manfrevola F, Mattia M, et al. The number of the CTCF binding sites of the H19/IGF2:IG-DMR correlates with DNA methylation and expression imprinting in a humanized mouse model. Hum Mol Genet. 2021;30(16):1509–20.
83. Rovina D, la Vecchia M, Cortesi A, Fontana L, Pesant M, Maitz S, et al. Profound alterations of the chromatin architecture at chromosome 11p15.5 in cells from Beckwith-Wiedemann and Silver-Russell syndromes patients. Sci Rep. 2020;10(1):1–9.
84. Nativio R, Sparago A, Ito Y, Weksberg R, Riccio A, Murrell A. Disruption of genomic neighbourhood at the imprinted IGF2-H19 locus in Beckwith-Wiedemann syndrome and Silver-Russell syndrome. Hum Mol Genet. 2011;20(7):1363–74.
85. Vu TH, Nguyen AH, Hoffman AR. Loss of IGF2 imprinting is associated with abrogation of long-range intrachromosomal interactions in human cancer cells. Hum Mol Genet. 2010;19(5):901–19.
86. Liao J, Zeng TB, Pierce N, Tran DA, Singh P, Mann JR, et al. Prenatal correction of IGF2 to rescue the growth phenotypes in mouse models of Beckwith-Wiedemann and Silver-Russell syndromes. Cell Rep. 2021;34(6):108729.
87. Lefevre P, Witham J, Lacroix CE, Cockerill PN, Bonifer C. The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription. Mol Cell. 2008;32(1):129–39.
88. Farhadova S, Gomez-Velazquez M, Feil R. Stability and lability of parental methylation imprints in development and disease. Genes. 2019;10(12):999.
89. Mahy NL, Perry PE, Gilchrist S, Baldock RA, Bickmore WA. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. J Cell Biol. 2002;157(4):579–89.
90. Lee R, Kang MK, Kim YJ, Yang B, Shim H, Kim S, et al. CTCF-mediated chromatin looping provides a topological framework for the formation of phase-separated transcriptional condensates. Nucleic Acids Res. 2022;50(1):207–26.
91. Holgersen EM, Gillespie A, Leavy OC, Baxter JS, Zvereva A, Muirhead G, et al. Identifying high-confidence capture Hi-C interactions using CHiCANE. Nat Protoc. 2021;16(4):2257–85.
92. Eijsbouts CQ, Burren OS, Newcombe PJ, Wallace C. Fine mapping chromatin contacts in capture Hi-C data. BMC Genomics. 2019;20(1):1–3.
93. Wang H, Lou D, Wang Z. Crosstalk of genetic variants, allele-specific DNA methylation, and environmental factors for complex disease risk. Front Genet. 2019;9:695.
94. Ray J, Munn PR, Vihervaara A, Lewis JJ, Ozer A, Danko CG, et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. Proc Natl Acad Sci U S A. 2019;116(39):19431–9.

95. Li G, Reinberg D. Chromatin higher-order structures and gene regulation. Curr Opin Genet Dev. 2011;21(2):175–86.
96. Yokoshi M, Segawa K, Fukaya T. Visualizing the role of boundary elements in enhancer-promoter communication. Mol Cell. 2020;78(2):224–35.
97. Ma L, Gao Z, Wu J, Zhong B, Xie Y, Huang W, et al. Co-condensation between transcription factor and coactivator p300 modulates transcriptional bursting kinetics. Mol Cell. 2021;81(8):1682–97.
98. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. GSE63525: Gene Expression Omnibus; 2014. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525
99. Sanborn AL, Rao SSP, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. GSE63525: Gene Expression Omnibus; 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525
100. Akgol Oksuz B, Yang L, Abraham S, Venev SV, Krietenstein N, Parsi KM, et al. Systematic evaluation of chromosome conformation capture assays. Nat Methods. 2021;18(9):1046–55.
101. Akgol Oksuz B, Yang L, Abraham S, Venev SV, Krietenstein N, Parsi KM, et al. Systematic evaluation of chromosome conformation capture assays. GSE163666: Gene Expression Omnibus; 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163666
102. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. GSE16256: Gene Expression Omnibus; 2009. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
103. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. GSE16256: Gene Expression Omnibus; 2010. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
104. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. GSE16256: Gene Expression Omnibus; 2011. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
105. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. GSE16256: Gene Expression Omnibus; 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
106. Micheletti R, Plaisance I, Abraham BJ, Sarre A, Ting CC, Alexanian M, et al. The long noncoding RNA Wisper controls cardiac fibrosis and remodeling. GSE16256: Gene Expression Omnibus; 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
107. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. GSE16256: Gene Expression Omnibus; 2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
108. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. GSE16256: Gene Expression Omnibus; 2010. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256
109. Zhou Q, Guan P, Zhu Z, Cheng S, Zhou C, Wang H, et al. ASMdb: a comprehensive database for allele-specific DNA methylation in diverse organisms. Nucleic Acids Res. 2022;50(D1):D60–71.
110. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. NA12878: GitHub; 2019. https://github.com/nanopore-wgs-consortium/NA12878
111. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
112. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. GSE40832: Gene Expression Omnibus; 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40832
113. Kacmarczyk TJ, Fall MP, Zhang X, Xin Y, Li Y, Alonso A, et al. 'same difference': comprehensive evaluation of four DNA methylation measurement platforms. GSE83595: Gene Expression Omnibus; 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83595
114. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
115. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. GSE30263: Gene Expression Omnibus; 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30263
116. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. GSE30263: Gene Expression Omnibus; 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30263
117. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. Topologically associating domains are stable units of replication-timing regulation. GSE51334: Gene Expression Omnibus; 2014. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51334
118. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. GSE31477: Gene Expression Omnibus; 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31477
119. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. GSE29611: Gene Expression Omnibus; 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29611
120. Zuin J, Franke V, van IJcken WFJ, van der Sloot A, Krantz ID, van der Reijden MIJA, et al. A cohesin-independent role for NIPBL at promoters provides insights in CdLS. PLoS Genet. 2014;10(2):e1004153.
121. Zuin J, Franke V, van IJcken WFJ, van der Sloot A, Krantz ID, van der Reijden MIJA, et al. A cohesin-independent role for NIPBL at promoters provides insights in CdLS. PRJEB3073: Bioproject; 2014. https://www.ncbi.nlm.nih.gov/sra/ERX115548/

122. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Roadmap Epigenomics Consortium: integrative analysis of 111 reference human epigenomes: Roadmap Epigenomics Project; 2015. https://egg2.wustl.edu/roadmap/web_portal/
123. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. 2018;19(1):151.
124. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10.
125. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res. 2015;4:1310.
126. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
127. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
128. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19(1):217.
129. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal. 2013;33(3):1029–47.
130. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1):11–0.
131. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017;27(5):801–12.
132. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 2017;27(1):157–64.
133. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. PRJEB3381: European Nucleotide Archive; 2012. https://www.ebi.ac.uk/ena/browser/view/PRJEB3381
134. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
135. Krueger F, Andrews SR. SNPsplit: allele-specific splitting of alignments between genomes with known SNP genotypes. F1000Res. 2016;5:1479.
136. Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Grüning B, et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics. 2021;37(3):422–3.
137. Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. BMC Bioinformatics. 2018;19(1):1–0.
138. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
139. Wingett SW, Andrews S. FastQ Screen: a tool for multi-genome mapping and quality control. F1000Res. 2018;7:1338.
140. Yang T, Zhang F, Yardımcı GG, Song F, Hardison RC, Noble WS, et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939–49.
141. An L, Yang T, Yang J, Nuebler J, Xiang G, Hardison RC, et al. OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. Genome Biol. 2019;20(1):1–6.
142. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. GSE86765: Gene Expression Omnibus; 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86765
143. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. GSE17312: Gene Expression Omnibus; 2010. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17312
144. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. GSE80911: Gene Expression Omnibus; 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80911
145. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13(10):1–9.
146. Peng HH, Chang SD, Chao AS, Wang CN, Cheng PJ, Hwang SM, et al. DNA methylation patterns of imprinting centers for H19, SNRPN, and KCNQ1OT1 in single-cell clones of human amniotic fluid mesenchymal stem cell. Taiwan J Obstet Gynecol. 2012;51(3):342–9.
147. Richer S, Tian Y, Schoenfelder S, Hurst L, Murrell A, Pisignano G. HiC capture imprinted gene panel. PRJNA926951: Bioproject; 2023. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA926951
148. Richer S, Tian Y, Schoenfelder S, Hurst L, Murrell A, Pisignano G. HiCFlow: Github; 2022. https://github.com/StephenRicher/HiCFlow
149. Richer S, Tian Y, Schoenfelder S, Hurst L, Murrell A, Pisignano G. HiCFlow: Zenodo; 2022. https://github.com/StephenRicher/AS-HiC-Analysis
150. Richer S, Tian Y, Schoenfelder S, Hurst L, Murrell A, Pisignano G. AS-HiC-analysis: Zenodo; 2022. https://zenodo.org/record/7563515
151. Richer S, Tian Y, Schoenfelder S, Hurst L, Murrell A, Pisignano G. AS-HiC analysis: Zenodo; 2022. https://zenodo.org/record/6510198

Publisher's Note

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