Dynamic changes in the epigenomic landscape regulate human organogenesis and link to developmental disorders

Dave T. Gerrard1,6, Andrew A. Berry1,6, Rachel E. Jennings1,2, Matthew J. Birket1, Peyman Zarrineh1, Myles G. Garstang1, Sarah L. Withey1, Patrick Short3, Sandra Jiménez-Gancedo4, Panos N. Firbas4, Ian Donaldson1, Andrew D. Sharrocks1, Karen Piper Hanley1,5, Matthew E. Hurles3, José Luis Gomez-Skarmeta4, Nicoletta Bobola1 & Neil A. Hanley1,2

How the genome activates or silences transcriptional programmes governs organ formation. Little is known in human embryos undermining our ability to benchmark the fidelity of stem cell differentiation or cell programming, or interpret the pathogenicity of noncoding variation. Here, we study histone modifications across thirteen tissues during human organogenesis. We integrate the data with transcription to build an overview of how the human genome differentially regulates alternative organ fates including by repression. Promoters from nearly 20,000 genes partition into discrete states. Key developmental gene sets are actively repressed outside of the appropriate organ without obvious bivalency. Candidate enhancers, functional in zebrafish, allow imputation of tissue-specific and shared patterns of transcription factor binding. Overlaying more than 700 noncoding mutations from patients with developmental disorders allows correlation to unanticipated target genes. Taken together, the data provide a comprehensive genomic framework for investigating normal and abnormal human development.
Organogenesis is the key phase when the body’s tissues and organs are first assembled from rudimentary progenitor cells. In human embryos, this is the critical period during weeks 5–8 of gestation when disruption can lead to major developmental disorders. While ∼35% of developmental disorders are explained by damaging genetic variation within the exons of protein-coding genes, de novo mutation (DNM) in the non-coding genome has been associated increasingly with major developmental disorders. The non-coding genome also harbours over 80% of single-nucleotide polymorphisms (SNPs) implicated in genome-wide association studies (GWAS) for developmental disorders, or in GWAS of later onset disease, such as schizophrenia and type 2 diabetes, where contribution is predicted from early development. These genetic alterations are presumed to lie in enhancers for developmental genes or in other regulatory elements, such as promoters for non-coding RNAs that may only be active in the relevant tissue at the appropriate stage of organogenesis. Aside from rare examples, this has remained unproven because of lack of data in human embryos. While regulatory data are available in other species at comparable stages, extrapolation is of limited value because the precise genomic locations of enhancers are poorly conserved even allowing for enriched sequence conservation around developmental genes. Sequence conservation alone is also uninformative for when and in what tissue a putative enhancer might function. Comprehensive regulatory information is available from later fetal development via initiatives such as NIH Roadmap, but these later stages largely reflect terminally differentiated, albeit immature cells rather than progenitors responsible for organ formation. In contrast, a small number of studies on a handful of isolated tissues, such as limb bud, craniofacial processes, pancreas, or brain, have demarcated regulatory elements directly during human organogenesis. However, most organs remain unexplored. Moreover, nothing is known about patterns of regulation, including both activation and repression, deployed across tissues, which is an important factor because tissues are often co-affected in developmental disorders. To address these gaps in our knowledge, we set out to build maps of genome regulation integrated with transcription during human organogenesis at comprehensiveness currently unattainable from single-cell analysis.

**Results**

**Assembly of datasets across 13 tissues.** Thirteen different types of organs and tissues were contributed by 61 human post-implantation embryos, microdissected and subjected to chromatin immunoprecipitation followed by deep sequencing (ChIPseq) for three histone modifications (Fig. 1a): H3K4me3, enriched at promoters of transcribed genes; H3K27ac, at active enhancers and some promoters; and H3K27me3 delineating regions of the genome under active repression by Polycomb. Tiny tissue size and the scarcity of human embryonic tissue required some pooling and precluded study of additional modifications. Biological replicates were undertaken for all but two tissue sites (details are in Supplementary Data 1 and 2). These all included both male and female tissue. Tissues and stages were matched to polyadenylated RNAseq datasets acquired at sufficient read depth to identify over 6000 loci with previously unannotated transcription (Supplementary Data 3). Overlaying the data revealed characteristic tissue-specific patterns of promoter and putative enhancer activity, and unannotated human embryonic transcripts. This was particularly noticeable surrounding genes encoding key developmental transcription factors (TFs), such as the example shown for NKX2-5 in the heart (Fig. 1b). Tissues lacking expression of the TF gene tended to carry active H3K27me3 modification rather than simply lack marking. Putative tissue-specific enhancer marks were characteristically distributed over several hundred kilobases (heart-specific peak (red) over 200 kb from NKX2-5 to the far right of Fig. 1b). These isolated H3K27ac marks were often unpredicted by publicly available data from cell lines or terminally differentiated lineages and did not necessarily show sequence conservation across vertebrates (mean per-base phyloP score 0.175; range −1.42 to +6.94 for \( n = 51,559 \) regions). Unexpected H3K4me3 and H3K27ac peaks that failed to map to the transcriptional start sites (TSSs) of annotated genes mapped to the TSS of previously unidentified human embryo-enriched transcripts, such as the example shown in Fig. 1b for the bidirectional HE-TUCP-C5T408 and HE-LINC-C5T409 (for the complete catalogue see Supplementary File 1H in ref. 13). Recognising the importance of features surrounding key developmental genes such as NKX2-5, all the individual data are available to browse as tracks on the UCSC Genome Browser. These preliminary observations encouraged us to undertake full integration of the different datasets to enable a series of genome-scale analyses.

**Major promoter states partition without obvious bivalency.** By analysis based on a Hidden Markov Model, the genome partitioned into different chromatin states very similarly across tissues. While three histone marks allowed for eight different segmentations, aggregation into fewer states was possible (Fig. 1c). On average across tissues, 3.3% of the genome was active promoter (States 1 & 2; H3K4me3 +/− H3K27ac) or putative enhancer (State 3; H3K27ac) (range 1.7–6.1%; Fig. 1c & Supplementary Fig. 1). In all, 6.7% was variably marked as actively repressed (States 6 & 7; range 3.3–13.0; H3K27me3), while on average 89.8% of the genome was effectively unmarked (States 4 & 5; range 81.7–94.0). Approximately 0.2% seemingly had both H3K4me3 and H3K27me3 marks (State 8; range 0.16–0.33). This latter state has been considered bivalent and characteristic of poised genes whose imminent expression then initiates cell differentiation pathways. Ascribing bivalency has been reliant on setting an arbitrary threshold for a binary decision of whether a site is marked or not. This risks the impression of equivalence when in fact one or other mark might be far more prominent. Moreover, apparent bivalency could simply reflect mixed marks due to heterogeneity in a cell population (the minor co-detection of H3K27ac in State 8 would not be expected in the presence of H3K27me3). Taken together, this reliance on an arbitrary threshold is suboptimal. Therefore, we used ngsplot to cluster promoter profiles for each histone mark integrated with transcription over 3 kb either side of 19,791 distinct protein-coding TSS in each tissue (Fig. 2; Supplementary Figs. 2–4). We started under default settings for three histone modifications and RNA-seq, which generated the five most prominent clusters. Broader H3K4me3 and H3K27ac signals at the TSS correlated with higher levels of transcription (Fig. 2a–b; we termed this promoter state broad expressed versus narrow or bidirectionally expressed). In total, 25–30% of genes across tissues were unmarked and lacked appreciable transcription (inactive). These promoters typically lacked CpG islands (<20% compared with 67.7% of the 19,791 genes). Conversely, 90–95% of TSS regions marked with H3K27me3 featured CpG islands with an over-representation of TFs; 31.2% of TFs (\( n = 1659 \)) were actively repressed in at least one tissue compared with 20.0% of non-TF genes (odds ratio 1.82, confidence interval 1.63–2.04; \( P < 2.2e-16 \)). H3K27me3 detection at the TSS was ~50% greater for over-representation of TFs (Fig. 2a–b; we termed this promoter state broad expressed versus narrow or bidirectionally expressed). In total, 25–30% of genes across tissues were unmarked and lacked appreciable transcription (inactive). These promoters typically lacked CpG islands (<20% compared with 67.7% of the 19,791 genes). Conversely, 90–95% of TSS regions marked with H3K27me3 featured CpG islands with an over-representation of TFs; 31.2% of TFs (\( n = 1659 \)) were actively repressed in at least one tissue compared with 20.0% of non-TF genes (odds ratio 1.82, confidence interval 1.63–2.04; \( P < 2.2e-16 \)). H3K27me3 detection at the TSS was ~50% greater for genes encoding TFs (Supplementary Fig. 5). The categorisation for each of the 19,791 genes in all tissues is listed in Supplementary Data 4. When prioritised under default settings, patterns across tissues were strikingly similar with only minor variation in...
bidirectional transcription in RPE (Supplementary Fig. 2) and a technical factor limiting the detection of expression at the TSS for particularly long transcripts in the liver, lung and brain (Supplementary Figs. 3 and 4). Consistently, neat partitioning would not have been possible if the data were overly confounded by cellular heterogeneity. While a major bivalent chromatin state at gene promoters was not detected under these parameters, it was noticed that H3K4me3 levels were slightly higher in the active repression group than for inactive genes (Fig. 2b, H3K4me3, red and grey lines, respectively). H3K27ac signal for active repression and inactive virtually overlapped, very close to the background signal. Therefore, we extended the ngsplot parameters to allow more subcategorization in the search for bivalency. Explicitly, we wanted to see if we could split the H3K27me3 signal into a subset of characteristic bivalent genes with clear cut H3K4me3. Seven (kidney), eleven (liver), or ten categories (all other replicated tissues) split the H3K27me3 signal. For each tissue, the smaller H3K27me3 sub-category now clustered with weak detection of H3K4me3 (and some low-level transcription). We assessed enrichment of these genes in each tissue for putative bivalent characteristics, such as imprinting or TFs characteristic of the particular tissue fate. However, in all instances the same generic categories, such as extracellular matrix organisation, receptor tyrosine kinases and a variety of neural terms emerged. Supplementary Figs. 6–11 show examples for the adrenal, kidney and pancreas. These categories and the genes underlying them are well-recognised to reflect mesenchyme and neural development as has been detailed for the pancreas22 with characteristic WNT and FGF family members23. In summary, promoter regions partitioned neatly into major chromatin states allied to

**Fig. 1 Epigenomic landscape across 13 human embryonic tissues.** a Thirteen different human embryonic sites were sampled for RNAseq and ChIPseq, as described in the “Methods” and in Supplementary Data 1 and 2. The same colour coding for each tissue is applied throughout the paper in overlaid ChIPseq tracks. The heart (left ventricle) dataset is summarised as Heart/LV from hereon. b 300 kb locus around the NKX2-5 gene, the most discriminatory TF gene for human embryonic heart15. The locus contains five unannotated human embryonic (HE) transcripts enriched in heart [three LINC RNAs and two transcripts of uncertain coding potential (TUCP)]. Heart/LV-specific (red) H3K4me3 and H3K27ac marks were detected at the NKX2-5 TSS and adjacent transcripts (HE-TUCP-CST408 and HE-LINC-CST409). Embryonic heart-specific H3K27ac marks were visible up to 200 kb away (e.g., at the extreme right of panel). H3K27me3 marked the region from NKX2-5 to HE-LINC-CST409 in all non-heart tissues (the track appears black from the superimposition of all the different colours other than red). ENCODE data are from seven cell lines26. c Genome coverage by ChromHMM for the different histone modifications was similar across all tissues (Supplementary Fig. 1) with an average 89.8% of the genome unmarked (range: 81.7–94.0; States 4 & 5), and 3.3% consistent with being an active promoter and/or enhancer (range: 1.7–6.1; States 1–3).
transcription. Sets of bivalent genes were not observed before gene sets indicative of cell-types common across all tissues limited the extent of subcategorization.

Mapping major promoter states discovers disallowed gene sets. Our classification allowed us to ask how the major promoter states changed across different tissues. Tracking all states in all tissues was complex to visualise (Supplementary Fig. 12). Unifying broad, narrow and bidirectional expressed into a single category (expressed) displayed how the majority of genes remained unaltered across tissues (Fig. 3a). In contrast, 29% of genes had a variable promoter state. Within this subset, we predicted that genes responsible for a specific organ’s assembly, such as those encoding developmental TFs, would need to be actively excluded or disallowed at inappropriate sites (as seen in Fig. 1b for NKX2-5). We tested this in the replicated datasets by comparing genes transcribed uniquely in one tissue for either inactivity (no mark) or active repression elsewhere (H3K27me3; disallowed). Gene ontology (GO) analysis of the uniquely expressed/disallowed elsewhere gene sets identified the appropriate developmental programme in all instances (as shown for the heart in Fig. 3b; e.g., heart development). These gene sets are listed for each tissue in Supplementary Data 5a. In contrast, tissue-specific transcription initiated from genes that were simply inactive in other organs tended to highlight differentiated cell function (Fig. 3b; e.g., sarcomere organisation). These observations highlight the preferential use of H3K27me3 at the

**Fig. 2 Classification of major promoter states.**

- **a** Clustered heatmaps surrounding the transcriptional start sites (TSS +/− 3 kb) of 19,791 annotated genes. The example shown is for adrenal. One replicate is shown for each data type for simplicity. Replicates across all tissues were near identical. Two minor variations on this pattern were detected in RPE (Supplementary Fig. 2) and the liver, lung and brain (Supplementary Fig. 3).

- **b** Mean signal levels for the genes clustered in (a). Traces are coloured according to the text colour in (a). Broad expressed genes show approximately double the level of transcription and twice the width of H3K4me3 and H3K27ac marks compared with narrow expressed genes.
promoters of genes controlling cell fate decisions, but not differentiated function. Moreover, the fact that these cell fate genes did not emerge for any tissue in the sub-categorisation of H3K27me3 signal with low-level H3K4me3 (Supplementary Figs. 6–11) further supports lack of bivalency at their gene promoters.

To scrutinise regulatory changes temporally, we focussed on pancreas and included datasets from human pluripotent stem cells (hPSCs) and adult tissue. Different sets of repressed genes lost their H3K27me3 mark to become expressed as cells transitioned from pluripotency to embryonic pancreatic progenitors or from pancreatic progenitors to mature pancreas (Fig. 3c;
genes are listed in Supplementary Data 5b). Surprisingly, the same KEGG term relating to monogenic diabetes emerged in both instances (Fig. 3d). However, the genes underlying the first transition related to early function in pancreatic organogenesis, hypoplasia, or aplasia (e.g., GATA6, SOX9 and PDX1); while the genes in the second transition specifically related to post-embryonic pancreatic islet cell differentiation and beta-cell function (e.g., INSM1, MAFA and NKX2-2)24,25.

**Human regulatory sequences function in zebrafish embryos.** Having recognised the disallowed status of developmental TFs in inappropriate tissues, we wanted to test whether our putative intergenic human embryo-enriched enhancers were capable of driving appropriate reporter gene expression at the correct locations in developing zebrafish. We identified H3K27ac marks that were enriched in the human embryo compared with 161 ENCODE or NIH Roadmap datasets10,26 and not detected in the FANTOM5 project27. We developed an algorithm to test for embryonic tissue specificity and filtered for sequence conservation (not necessarily in zebrafish; see “Methods”). We manually inspected the remainder for proximity (<1 mb) to genes encoding TFs and, in particular, to increase clinical relevance, to those associated with major developmental disorders. We ensured no H3K4me3 or polyadenylated transcription in the immediate vicinity (i.e., an unannotated promoter). We tested 10 such enhancers out of 44 within 1 mb of TBX15, HEY2, ALX1, IRX4, PITX2, HOXD13, NKX2-5, WT1, SOX11 and SOX9 for their ability to direct appropriate GFP expression in stable lines of transgenic zebrafish (Supplementary Data 6). Two (h-003-kid near WT1 and h-022-mix near SOX11) failed to generate any GFP in any location. The remaining eight all yielded GFP at the predicted site in zebrafish embryos (Fig. 4; Supplementary Data 6), despite only one of the putative enhancer sequences being conserved in zebrafish (Fig. 4a, h-027-lim near TBX15). Taken together, these data imply that our H3K27ac detection marks previously unannotated human enhancers, which function over considerable evolutionarily distance, despite the absence of detectable sequence conservation. This is most likely associated with overall conservation of TF activity between homologous organs across species28.

**Patterns of H3K27ac modification are unique across tissues.** We wanted to explore the link between these regulatory elements and surrounding gene expression at genome-wide scale. Assured that ChIPseq marks were reproducible within biological replicates without batch effect (Supplementary Fig. 13), we parsed the genome into 3,087,584 non-overlapping 1 kb bins. Reads within each bin were counted for each mark. Phi correlation between biological replicates indicated this approach to peak calling was very similar to using MACS (Supplementary Fig. 14). Counts were downsampled and averaged within tissues and correlated with the corresponding RNAseq data over 1 mb in either direction (i.e., a 2 mb window). On average, this window included 44 annotated genes (range: 0–247 genes). For those H3K27ac marks which functioned in zebrafish, the strongest correlation was with the appropriate TF gene, for instance TBX15 over ~500 kb in limb (Fig. 4a). Moreover, within the 2 mb window, different H3K27ac marks could be correlated to the same gene, potentially allowing previously unknown enhancers to be grouped, for instance in the adrenal around the adrenal hypoplasia gene, NR0B1, located on the X chromosome (Supplementary Fig. 15).

Parsing the ChIPseq data into bins allowed integration of information across tissues, which is challenging when based on empirical modelling by MACS. Placing raw read counts per bin in rank order produced near identical elbow plots for all marks in all tissues. This avoided arbitrary assignment and allowed the point of maximum flexure to be used quantitatively for calling marks in a binary yes/no fashion (Fig. 5a). The simplified calling facilitated exploration of regulatory patterns across tissues. Requiring a bin to be marked in any two or more samples identified 48,570 different H3K27ac patterns genome wide. The top 40 are shown in Fig. 5b. While tissue specificity for the heart/left ventricle was the most common pattern, all replicated organs ranked within the top 0.6% of patterns. Nine out of the 11 replicated tissues ranked in the top 0.2% (Supplementary Data 7). These data indicate high reproducibility and consistency across tissue replicates. H3K27ac showed far more tissue-selective patterns than H3K4me3 or H3K27me3 (Fig. 4b; Supplementary Figs. 16 and 17). Motif analysis on the tissue-specific H3K27ac regions allowed imputation of master TFs for individual tissues, such as NR5A1 in 54.5% of adrenal-specific bins (n = 18,411) compared to 25% of the remaining 141,706 bins (Fig. 5c). Mutation of NR5A1 causes adrenal agenesis in human and mouse (OMIM 184757). MEF, TBX and bHLH family members emerged in the heart-specific bins (Fig. 5c); all are associated with congenital heart disease29. This emergence of TFs with known, critical tissue-specific functions underscores the validity of parsing the ChIPseq data into bins as well as the consistency of data between tissue replicates.

Having integrated our data, we could also uncover regulatory regions that were shared precisely across two or more tissues relevant to developmental disorders which manifest in multiple organs. Enrichment for composite PITX1/bHLH motifs was found in the limb and palate (Fig. 5c). GATA-binding motifs were enriched in the heart and pancreas. Shared patterns could be explicitly instructed by requiring detection in four or more samples (Supplementary Fig. 18). We hypothesised that patterns shared across many tissues ought to contain elements regulating generic developmental functions. Scrutinising bins marked in over half of all H3K27ac samples (n = 30,226 bins versus remaining background of 80,352 bins) identified enrichment for the ETS motif. ETS transcription factors are involved in cell cycle control and proliferation30.

**Overlying the epigenomes with developmental disorders.** Non-coding mutations in promoters or enhancers have been linked increasingly to major developmental disorders3,31. Previously, as part of the Deciphering Developmental Disorders (DDD) study, we studied 7930 individuals and their parents32. In all, 87% of patients had neurodevelopmental disorders. 10% had congenital heart defects. 68% of patients lacked disease-associated DNMs within exomes (exome-negative) pointing to the likely importance of the non-coding genome2. We sequenced 6139 non-coding regions (4.2 mb) selected as ultra-conserved regions (UCRs: n = 4307), experimentally validated enhancers (EVEs: n = 595) or as putative heart enhancers (PHE: n = 1237) and found 739 non-coding DNMs3. In total, 78% of the 6139 regions were marked by H3K27ac or H3K4me3 in our embryonic tissues, with a higher percentage overlap for the EVEs (87%) and near-perfect overlap for the PHEs (99%) (Fig. 6a). An additional 9% were marked by H3K27me3, suggesting non-coding regulation in a currently unsampled tissue. The distribution of DNMs was very similar (Fig. 6b). Nearly half of the regions containing DNMs were marked by H3K27ac and/or H3K4me3 that was replicated in at least one tissue. Most commonly, this included the heart or brain, in keeping with the predominance of neurodevelopmental and cardiac phenotypes in the DDD cohort and the PHEs selected for sequencing (Fig. 6c). In total, 75% of the PHEs with DNMs mapped to replicated H3K27ac and/or H3K4me3 in our heart dataset. This rose to 100% if the need for replication was
removed. We did not observe enrichment for DNMs in patients with heart, eye or limb phenotypes in elements marked by H3K27ac. Similarly, organ-level association with clinical phenotype was not possible. In both circumstances, the power of testing was markedly curtailed by limitations in patient numbers (Fig. 6d). Enrichment for DNMs in elements marked by H3K27ac was detected with neurodevelopmental disorders (87% of the DDD cohort; 1.45-fold, 95% confidence interval 1.09–1.90; \( P = 0.0056 \)) (Fig. 6d). This was similar to our previous report using NIH Roadmap H3K27ac and/or DNaseI hypersensitivity data.
Our results support a role for non-coding mutations in severe neurodevelopmental disorders.

Prioritising individual DNMs for potential pathogenicity and how they might disrupt surrounding gene function is very challenging. We anticipate an atlas of human embryonic regulatory marks comprising different tissues will help stratify disease-relevant non-coding regions; and that correlation to surrounding transcription offers a preliminary means for prioritising putative target gene(s). We have made all correlations freely available as tracks on the UCSC Genome Browser and present two examples here as preliminary exemplars. A DNM linked to neurodevelopmental disorders in a UCR on chromosome 16 is in the middle of the annotated testicular LINC RNA,
Fig. 5 Patterns of enhancer activity and transcription factor binding across tissues. a Elbow plots for each histone modification following allocation of the genome into 3.1 million consecutive bins of 1kb. The example shown is for adenral providing the number of reads per bin at the point of maximum gradient change (the elbow point, red dot) and a quantitative measure of whether a bin was marked or not (e.g., >10 or <10, respectively, for H3K27ac). Converting marks into a binary yes/no call at any point in the genome facilitated the data integration across the different tissues. While the number of reads per bin at the elbow point was different for each mark across the tissues, the shape of the curve remained the same. b Euler grid for bins marked by H3K27ac (defined by elbow plots) in replicated tissues (i.e., two rows/replicates per tissue). Total number of marked bins per individual dataset is shown to the right. The example in (b) required a bin to be called in any two or more samples and is ordered by decreasing bin count per pattern (bar chart above the grid). A total of 48,570 different patterns were identified, of which the top 40 are shown. Tissue specificity for all sites emerged in the top 265 (0.5%) patterns; colour-coded asterisks above columns). For example, nearly 14,000 bins marked only in the two Heart/LV H3K27ac datasets ranked first as the most frequent pattern. The most frequent pattern in ~3000 bins was palate-specific. Tissue-specific patterns were far less apparent at promoters (H3K4me3, n = 18,432; Supplementary Fig. 10) or for H3K27me3 (n = 26,339; Supplementary Fig. 11). While patterns across multiple tissues were permitted by stipulating marks in ≥2 samples (e.g., heart and adrenal in column 24), they could be enforced by stipulating marks in at least four samples (Supplementary Fig. 12).

c Enrichment of known TF-binding motifs in the tissue-specific patterns of H3K27ac identified in (b). Five individual tissues are shown as examples alongside analysis of the shared regulatory pattern identified for the limb and palate identifying marked enrichment of a compound PITX1:E-box motif. Motif-enrichment was conducted using a one-sided Binomial test implemented in findMotifsGenome.pl of the Homer package.

LINC01572 (Fig. 7; chr16:72427838). Our data illustrate that the DNM is also located at the TSS of HE-OT-AC0041583, expressed at 19.5-fold higher levels in human embryonic brain than in the brain, 1317.2; mean in other tissues, 32.2), in a 4 kb region surrounding this putative enhancer, expression of the gene, RBM24, located ~1 mb away, was markedly enriched in heart compared to other organs (Fig. 8b) and the gene most correlated to the enhancer (r = 0.88). RBM24 is required for cardiac development with knockout mice dying at E12.5-14.5 with ventricular septal defects and compromised cardiac muscle assembly. We used CRISPR-Cas9 to delete the enhancer in our established NKX2-5-GFP reporter hPSC model (Fig. 8c). Cells were viable, but loss of the enhancer markedly curtailed the generation of NKX2-5-positive cardiomyocyte progenitors (Fig. 8d–e). This was associated with markedly reduced expression of RBM24, but not of other genes in the locus (Fig. 8f).

Taken together, we have assembled datasets of regulatory activity linked to transcription for a range of human embryonic tissues during the period of organogenesis. All data are available to browse as tracks on the UCSC Genome Browser ready for overlay with genetic variants identified by clinical sequencing and GWAS.

Discussion

Previous studies of enhancer usage in human embryos have tended to focus on individual tissues suffering, amongst other findings, aspects of genome regulation responsible for human-specific attributes. Here, we incorporated epigenomic data with transcription across 13 sites during human organogenesis to build tissue-by-tissue maps of enhancers and promoters linked to gene expression. While similar to prior work in mouse and building on our previous transcriptomic atlas, the integrated approach here offers opportunities to understand how human organ formation is regulated in health and disease.

It is important to draw out certain features of the study, including limitations. The datasets comprised pooled tissue samples across several embryos including both sexes. This removed the opportunity for analysis of male or female tissue in isolation; however, it mitigated against the risk of misinterpreting organ-level differences that might be due to sex, for instance related to H3K27me3 and X chromosome inactivation. Moreover, the aggregated data across many embryos reduced the risk of other misleading conclusions due to one-off biological or technical factors, such as deterioration during transport. Our tissue collection and molecular analyses occur on a single campus with immediate processing. For instance, in quality control, we have found overnight delay in preservation media can lead to a 16-fold decline in the levels of some key developmental TFs, despite tissue remaining viable and suitable for immunohistochemistry and subsequent tissue culture. It is also obvious that the bulk analyses contain aspects of tissue heterogeneity, as we reached in all tissues with our deeper analyses of promoter state (Supplementary Figs. 6–11). However, the emergent features of our data such as the gene ontology for major promoter states (Fig. 3b, d) and the binding motifs within tissue-specific enhancers (Fig. 5c) indicate the majority contribution from organ-specific progenitor cells. Our past analysis based on variance also minimises any contribution from cell-types common to all tissues. While single-cell technology allows deconvolution of organ-level heterogeneity, at present the techniques do not permit the extent of analysis for histone modification or transcription as we have undertaken here. We envisage our integrated atlas will provide a valuable complement to current and future single-cell analyses of human organogenesis.

Assessing the non-coding genome is very challenging: millions of rare variants are returned from whole-genome sequencing (WGS) in each individual, while only one might be pathogenic. Although we provided proof-of-principle here for a cardiac enhancer, functional analysis, even of a handful of variants, is clinically impractical. For non-coding mutations to affect organogenesis (either in developmental disorders or in later life disease such as type 2 diabetes where there is an embryonic contribution), it is logical that mutations are located in regulatory elements for the limb and palate identifying marked enrichment of a compound PITX1:E-box motif. Motif-enrichment was conducted using a one-sided Binomial test implemented in findMotifsGenome.pl of the Homer package.

Fig. 8d–e)
Fig. 6 Overlay of non-coding de novo mutations linked to developmental disorders. 

a The Deciphering Developmental Disorders (DDD) study included 6139 non-coding regions in its sequence analysis of trios comprising affected individuals and unaffected parents. These non-coding regions were selected on the basis of high sequence conservation (ultra-conserved elements, UCEs, n = 4307), experimental validation (experimentally validated enhancers, EVEs, n = 595) or identification as a putative heart enhancer (PHE, n = 1237). Overlap with any H3K27ac, H3K4me3 or H3K27me3 1 kb bins is shown as an aggregate and for each individual category (UCE, EVE or PHE).

b Equivalent overlap is shown for the 739 regions in which disease-associated de novo mutations (DNMs) were identified. In total, 46% of DNM-positive regions were situated (+/− 1 kb) in at least one tissue-replicated H3K27ac and/or H3K4me3 bin. Over half of the disease-associated overlap was covered by the heart/LV (35%) and brain (18%). 75% of the disease-associated PHE regions were situated within 1 kb of a heart/LV-specific histone mark.

c Enrichment in the number of DNMs overlapping (+/− 1 kb) H3K27ac marks during human organogenesis for individuals with neurodevelopmental (n = 671 cases), cardiac (n = 124 cases), limb (n = 312 cases) and eye (n = 288 cases) phenotypes. The circles represent the observed/expected ratio with asymmetrical error bars showing the 95% confidence limits calculated for a Poisson distribution (http://ms.mcmaster.ca/peter/s743/poissonalpha.html). For the neurodevelopmental phenotypes, this included analysis against DNAse hypersensitivity data and H3K27ac data from second trimester fetal brain.
promoters reinforced our previous findings based solely on computational analysis of 5′ flanking regions for the importance of NR5A1 in the adrenal and HNF4A in the liver. However, the integrated sampling of numerous sites uncovered far more complex patterns of regulation operating across tissues. The enrichment of PITX1-binding motifs in active regulatory regions uniquely shared across the limb bud and palate fits with mutations in PITX1 causing limb defects and cleft palate. Similarly, GATA4 and GATA6, inferred from regulatory regions shared uniquely between the heart and pancreas, are the only two TFs linked to the dual phenotype of cardiac malformation and monogenic diabetes. Overlaying GWAS data with chromosomal compensation studies from older human fetal brain has prioritised target genes for risk of schizophrenia. These techniques are yet to be applied at scale in much smaller human embryonic tissues. However, we argue because the data we have collected from many tissues we can begin to make correlations of enhancer activity to target genes over megabase distances (Fig. 7). While these preliminary correlations are somewhat rudimentary, we have made them available as tracks on UCSC Genome Browser because where linked to expression of the same gene, it might become possible to group individual enhancers into larger clusters to increase statistical power. The latter can be otherwise limited when causally linking non-coding elements to developmental disorders.

Deciphering profiles of H3K27me3 alongside other regulatory marks and expression profiles was informative. We did not observe bivalent marking of developmental promoters poised for gene expression before reaching the limits of resolution, at which gene sets for structures such as mesenchyme and nerves were common to all tissues. Instead, we discovered that organ-specific developmental programmes were disallowed in older human embryonic tissues by active repression at a series of gene promoters. The ontology of these gene sets, including many encoding TFs, inferred they are an important aspect of ensuring correct cell fate decisions. This realisation opens up an opportunity for more rigorous benchmarking of differentiated hPSCs, including organoids, both for proximity to the intended lineage in how appropriate gene expression is activated but also against a clearly defined set of epigenomic features for how undesired cell fates are avoided.

In summary, we present an integrated atlas of epigenomic regulation and transcription responsible for human organogenesis. We make all datasets freely available alongside analysis tracks on the UCSC Genome Browser. The uncovering of cryptic regulatory regions and patterns of regulation across organs arose because of direct study of human embryonic tissue. The data complement current international projects such as the Human Cell Atlas, by providing greater resolution of regulatory information and depth of sequence information. Moreover, our integrated analyses establish a framework for prioritising and interpreting disease-associated variants discovered by WGS and provide clear routes towards understanding the underlying mechanisms.

Methods

Sample dissection. Human embryonic material was collected under ethical approval from the North West Research Ethics Committee (18/NW/0096), informed consent from all participants and according to the Codes of Practice of the Human Tissue Authority. Tissue collection took place on our co-located clinical academic campus overseen by our research team ensuring immediate transfer to the laboratory. Material was staged by the Carnegie classification, and individual tissues and organs were immediately dissected (Supplementary Data 1 and 2). The material collected here for epigenomic analysis was matched to material isolated for a previous transcriptomic study, and the dissection process was identical. In brief, the pancreas, adrenal gland, whole brain, heart, kidney, liver, limb buds, lung, stomach and anterior two-thirds of the tongue were visible as discrete organs and tissues. All visible adherent mesenchymes, including capsular material (adrenal), were removed under a dissecting microscope. The ureter was removed from the renal pelvis. A window of tissue was removed from the lateral wall of the left ventricle of the heart. The dissected segment of liver avoided the developing gall bladder. The trachea was removed where it entered the lung parenchyma. The stomach was isolated between the gastro-oesophageal and pyloric junctions. The palatal shelves were dissected on either side of the midline. The eye was disected, and the RPE peeled off mechanically from its posterior surface (facilitated by the dark pigmentation of the RPE allowing straightforward visualisation).
Tissues were gently teased apart before cross-linking in 1% formaldehyde for 10 min at room temperature. Fixation was quenched with 125 mM glycine for 5 min at room temperature before centrifugation, removal of the supernatant and washing twice with 1 ml of PBS. The final PBS supernatant was discarded and samples stored at −80 °C until use (Supplementary Data 1). The sex of tissue from embryos was determined by PCR for X and Y chromosome-specific primers (Supplementary Data 9).

Chromatin immunoprecipitation (ChIP), RNA isolation and sequencing. All ChIPseq datasets were in biological replicate, except for the stomach and tongue (Supplementary Data 1). Each sample was placed in lysis buffer [10 mM HEPES, 0.5 mM EGTA, 10 mM EDTA, 0.25% Triton X-100 and protease inhibitor cocktail (Roche)] on ice for 5 min, and nuclei released with ten strokes in a Dounce homogeniser. Nuclei were pelleted by centrifugation at 700 rcf for 10 min at 4 °C, and the supernatant discarded. Nuclei were resuspended in ice cold wash buffer.
Eight protein-coding genes were assayed on a cardiac-specific H3K27ac peak (red, broken line box) within the last intron of ATXN1. This enhancer harbours a DNM from the DDD cohort associated with congenital heart disease. Fold enrichment in the heart/LV dataset compared with the average across all other tissues for RNAseq read counts of the genes shown in (a). RBM24 and CAP2 are considerably enriched in the human embryonic heart. Magnified schematic of the enhancer shown in (a) showing the location of the DNM and the CRISPR-Cas9 approach for deletion. EBs from wild-type and enhancer deletion (mutant) hPSCs containing the N2KX2-5-GFP reporter. The images showing nine EBs for wild-type and mutant are after 14 days of the cardiomyocyte differentiation protocol. Size bar, 500 μm. Box and whisker plot (box showing 25th—75th percentile and median line with min—max as whiskers) quantifying GFP across all wild-type (n = 29) and mutant EBs (n = 30). RT-qPCR for expression of all the protein-coding genes across the 2.4 m locus depicted in (a). Ten different clones were used in three independent experiments for mutant EBs with ten, ten and nine clones for wild-type control. Error bars represent S.E.M. from the three independent differentiation experiments (the individual dots). Points for RT-qPCR were day 0 (undifferentiated hPSCs) and day 14 (cardiomyocyte progenitors). While CAP2 expression appeared reduced, RBM24 was the only gene with significantly lowered expression following deletion of the cardiac-specific enhancer shown in (a). Significance was assessed using a two-tailed Student’s t-test (ns, not significant).

Clustered promoter states were identified for an annotated set of 19,791 protein-coding genes in each tissue using ngsp.plot—k-means clustering (version 2.61) on unnormalized reads for the combined dataset of replicated RNA-seq and ChiPseq for H3K4me3, H3K27ac and H3K27me3. Default settings allowed for five clusters based on rank profiles of read counts 3 kb either side of the TSS. The returned clusters were then classified according to characteristics detected in both replicates into five major promoter states (Fig. 2): actively repressed (H3K27me3 signal >50% of maximum and mean transcript counts <10% of maximum); Narrow expressed [H3K4me3 signal >0% and mean transcript counts >10% of maximum]; Broad expressed (as for narrow expressed, but with skew >0.65); bidirectional expressed (H3K4me3 signal >25% of maximum with <90% of reads downstream of the TSS; and mean transcript counts >10% of maximum); Broad expressed (as for narrow expressed, but with skew >0.65); bidirectional expressed (H3K4me3 signal >25% of maximum with <90% of reads downstream of the TSS; and mean transcript counts >10% of maximum); and inactive (<25% of maximum for H3K4me3 and H3K27me3 and mean transcript counts <10% of maximum). This approach left each gene uniquely assigned to one cluster in any tissue. Bi-dar2 was only identified in the heart, liver, and brain (Supplementary Fig. 3). While superficially this category lacked significant transcription, in fact, total gene-level read counts were very similar to Broad expressed. However, longer mRNA and longer first introns limited transcription detection at the TSS (Supplementary Fig. 4). The full lists are in Supplementary Data 4. The over-representation of TSSs in the TSS regions marked with H3K27me3 and featuring CpG islands was assessed on the dataset of 1659 genes encoding all the TFs compared against the remaining 18,132 non-TF genes using Fisher’s exact test (two-sided). To search explicitly for bivalency of H3K4me3 and H3K27me3 at gene promoters, the default parameters of ngsp.plot were extended to allow more clusters (7, 10 or 11 as described in the “Results”). This generated a sub-category of H3K27me3 for ‘bivalency that also contains H3K4me3’ (Supplementary Fig. 10).

Allotrope plots were created using the R package Alluvial Diagrams version 0.2-0β with modification of the R code to reorder the horizontal splines (alluvia) within each tissue to keep similar colours together.

Annotation set enrichment for genes and genomic regions. Lists of genes from the up- and down-regulated promoters were used for enrichment of Gene Ontology (GO) terms using the xEnricherGen function from the R package XGR version 1.1.1 under default parameters44. Background comprised all remaining annotations used in the ngspplots either for embryonic tissues (Fig. 3a, b) or from hPSCs to adult through pancreas (Fig. 3c, d). Pathway annotation and gene set enrichment analysis for the subset of H3K27me3/H3K4me3 dual-marked promoters in each tissue was undertaken using ReactomePA (release 3.10) under default parameters, which associates genes to their known functions based on the REACTOME pathway database89.

Transgenic analysis in zebrasfish. A systematic approach identified candidate enhancers that were human embryo-enriched and tissue-specific. We identified markers from the first batch of H3K27ac with RPMK ≥ 25 and ≥2.5-fold enrichment in the human embryo compared to ENCODE (7 cell lines)66 or NIH Roadmap datasets (154 samples)105 were mapped similarly. Subsequently, all data, including the external H1 hPSC and adult pancreas data (Fig. 3c), were mapped to hg38 using STAR (2.4.2a)45. ChiPseq reads were trimmed to 50 bp for consistency, and only uniquely mapped reads were retained. For ChiPseq, spliced mappings were suppressed by setting the parameter “alignIntronMax” to 1. The full STAR parameters for ChiPseq are as follows: “alignIntronMax=1;readSearchStartMax=30;—outSAMAttribute All,—outSAMtype BAM SortedByCoordinate”. GENCODE 25 gene annotations were used for RNAseq mapping and read counting70. The full STAR parameters for RNA-seq were as follows: “—outSAMAttribute All,—quantMode GeneCounts—out, SAMtype BAM SortedByCoordinate”.

Chromatin and promoter state analysis. Genomic segmentation was performed using chromHMM (version 1.11) under default parameters labelling samples by tissue and histone modification. The three histone marks allowed for eight segment states.
Genome binning, normalisation and thresholding. The genome was parsed into 3,087,584 non-overlapping contiguous 1 kb bins to compare ChIPseq profiles across tissues and replicates. Reads were counted into bins according to their mapped start position using ciasaw (version 3.11)17. Reads from mitochondrial and unplaced scaffolds were removed. This resulted in 44 candidate enhancers from which we tested ten. The candidate sequences were upregulated in a given tissue and up to a half of all samples) were produced for each embryonic tissue. The injection mixture contained 50 ng/μl Toø2 transposase mRNA, purified enhancer test vector and 0.05% phenol red. The concentration of the enhancer test vector was between 15 and 30 ng/μl. Injected embryos were visualised from 24 hpf to 48 hpf in an Olympus stereomicroscope coupled to a fluorescence excitation light source in order to detect the pattern of GFP. Embryos and adults zebrafish were maintained under standard laboratory conditions. They were regularly scored for the presence of Spanish and European eels. All protocols used have been approved by the Ethics Committee of the Andalusian Government (license numbers 450-1839 and 182-41106 for CABS-CSIC-UP).

Motif analysis. HOMER v4.9 was used to search for enriched motifs in selected sets of bins16. For selected 1 kb bins marked with H3K27Ac, the background set was the remainder of bins with replicated H3K27Ac across all tissues (n = 160,943). hPSC culture, cardiomyocyte differentiation and enhancer inactivation. hPSCs (hPSCs (NXK2-5GFP) and hPSCs (NXK2-5GFP reporter) were maintained on mouse embryonal fibroblasts and passaged using TrypLE Select (Thermo Fisher Scientific) with cardiomyocyte differentiation performed via embryoid body (EB) formation34,60,61; and then recombined to the reporter vector Minitol2-GwB (D896–D901 (2017)).

References
1. Definitions and scope. Developmental Disorders Study, Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438 (2017).
2. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature 555, 611–616 (2018).
3. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901 (2017).
4. Weldon, M. N. et al. Recurrent mutations in a distal PTEA enhancer cause isolated pancreatic agenesis. Nat. Genet. 46, 61–64 (2014).
5. van Arensberg, J. et al. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. Genome Res. 20, 722–732 (2010).
6. Nord, A. S. et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell 155, 1521–1531 (2013).
7. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328, 1036–1040 (2010).
21. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs-plot: quick mining and visualization of next-generation sequencing data by including genomic databases. BMC Genomics 15, 294 (2014).
22. Yung, T. et al. Sufu- and Spop-mediated downregulation of Hedgehog signaling promotes beta cell differentiation through organ-specific niche signals. Nat. Commun. 10, 4647 (2019).
23. Heller, R. S. et al. Expression patterns of Wnts, Frizzleds, sFRPs, and misexpression in transgenic mice suggesting a role for Wnts in pancreas and foregut pattern formation. Dev. Dyn. 225, 260–270 (2002).
24. Jennings, R. E., Berry, A. A., Strutt, J. P., Garrard, D. T. & Hanley, N. A. Human pancreas development. Development 142, 3126–3137 (2015).
25. Jennings, R. E. et al. Development of the human pancreas from foregut to endocrine commitment. Diabetes 62, 3514–3522 (2013).
26. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).
27. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
28. Stergachis, A. B. et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature 515, 365–370 (2014).
29. Rana, M. S., Christoffels, V. M. & Moorman, A. F. A molecular and genetic outline of cardiac morphogenesis. Acta Physiol. 207, 588–615 (2013).
30. Sizemore, G. M., Pitarresi, J. R., Balakrishnan, S. & Ostrowski, M. C. The ETS family of oncogenic transcription factors in solid tumours. Nat. Rev. Cancer 17, 337–351 (2017).
31. Martin, J. C. et al. Quantifying the contribution of recessive coding variation to developmental disorders. Science 362, 1161–1164 (2018).
32. Firth, H. V. & Wright, C. F. The deciphering developmental disorders (DDD) study. Dev. Med. Child Neurol. 53, 702–703 (2011).
33. Yang, J. et al. RBM24 is a major regulator of muscle-specific alternative splicing. Dev. Cell 31, 87–99 (2014).
34. Birket, M. J. et al. Expansion and patterning of cardiovascular progenitors derived from human pluripotent stem cells. Nat. Biotechnol. 33, 970–979 (2015).
35. Elliott, D. A. et al. NKKx2-5(eGFP/w) hESCs for isolation of human cardiac embryonic material and prepared samples for all sequencing analyses. D.T.G., I.D. and N.A.H. were involved in study design and oversight of human embryonic material collection (N.A.H., R.E.J. and K.P.H.). A.A.B. processed the human embryonic material and clinical research training fellow, and S.J.W. was an MRC doctoral account PhD student. J.L.G.S. was supported by the Marató TV3 Fundacion (Grant No. 201611).
36. The authors declare no competing interests.
37. Birket, M. J. et al. Human stem cell model of fabry disease implicates LIMP-2 accumulation in cardiomyocyte pathology. Stem Cell Rep. 13, 380–393 (2019).
38. March, A. et al. Expansion of the global blood circulation. Dev. Cell 40, 24–36 (2018).
39. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).
40. Birket, M. J. et al. PGC-1alpha and reactive oxygen species regulate human embryonic stem cell-derived cardiomyocyte function. Stem Cell Rep. 1, 560–574 (2013).
41. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and cnidarian genomes. Science 328, 578–583 (2010).
42. Mukhopadhyay, D. et al. Human Na+ channel isoform Nav1.5 is required for early and mid-gestation. Mol. Cell. Neurosci. 47, 164–180 (2011).
43. Allen, H. L. et al. GATA6 haploinsufficiency causes pancreatic agenesis in humans. Nat. Genet. 44, 20–22 (2011).
44. Wang, S.-N. et al. GATA4 mutations are a cause of neonatal and childhood-onset diabetes. Diabetes 63, 2888–2894 (2014).
45. J.L.G.S. N.A.H. is the guarantor. Peer review information Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
46. Peer reviewer correspondence Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
47. Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-17305-2.
48. Reprints and permission information Reprints and permission information is available at http://www.nature.com/reprints.
49. Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
50. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
© The Author(s) 2020.