

1 **The epigenomic landscape regulating organogenesis in human embryos linked to** 2 **developmental disorders**

3 Dave T. Gerrard^{1*}, Andrew A. Berry^{1*}, Rachel E. Jennings^{1,2}, Matthew J Birket¹, Sarah J
4 Withey¹, Patrick Short³, Sandra Jiménez-Gancedo⁴, Panos N Firbas⁴, Ian Donaldson¹, Andrew D.
5 Sharrocks¹, Karen Piper Hanley^{1,5}, Matthew E Hurles³, José Luis Gomez-Skarmeta⁴, Nicoletta
6 Bobola¹ and Neil A. Hanley^{1,2,^}

7 1. Faculty of Biology, Medicine & Health, Manchester Academic Health Sciences Centre,
8 University of Manchester, Oxford Road, Manchester M13 9PT, UK

9 2. Endocrinology Department, Manchester University NHS Foundation Trust, Grafton Street,
10 Manchester M13 9WU, UK

11 3. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

12 4. Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones
13 Científicas / Universidad Pablo de Olavide / Junta de Andalucía, Sevilla, Spain

14 5. Wellcome Centre for Cell-Matrix Research, University of Manchester, Oxford Road,
15 Manchester M13 9PT, UK

16

17 *These authors contributed equally to this work

18 ^Author for correspondence

19

20 **Key words:** Human, development, embryo, organogenesis, enhancer, promoter, gene regulation,
21 organ, tissue.

22

23 **How the genome activates or silences transcriptional programmes governs organ**
24 **formation. Little is known in human embryos undermining our ability to benchmark the**
25 **fidelity of in vitro stem cell differentiation or cell programming, or interpret the pathogenicity**
26 **of noncoding variation. Here, we studied histone modifications across thirteen tissues during**
27 **human organogenesis. We integrated the data with transcription to build the first overview of**
28 **how the human genome differentially regulates alternative organ fates including by**
29 **repression. Promoters from nearly 20,000 genes partitioned into discrete states without**
30 **showing bivalency. Key developmental gene sets were actively repressed outside of the**
31 **appropriate organ. Candidate enhancers, functional in zebrafish, allowed imputation of**
32 **tissue-specific and shared patterns of transcription factor binding. Overlaying more than 700**
33 **noncoding mutations from patients with developmental disorders allowed correlation to**
34 **unanticipated target genes. Taken together, the data provide a new, comprehensive genomic**
35 **framework for investigating normal and abnormal human development.**

36 **INTRODUCTION**

37 Organogenesis is the key phase when the body's tissues and organs are first assembled from
38 rudimentary progenitor cells. In human embryos this is the critical period during weeks five to eight
39 of gestation when disruption can lead to major developmental disorders. While approximately 35%

40 of developmental disorders are explained by damaging genetic variation within the exons of
41 protein-coding genes¹, de novo mutation (DNM) in the noncoding genome has been associated
42 increasingly with major developmental disorders². The noncoding genome also harbours over 80%
43 of single nucleotide polymorphisms (SNPs) implicated in genome wide association studies
44 (GWAS) for developmental disorders, or in GWAS of later onset disease, such as schizophrenia and
45 type 2 diabetes, where contribution is predicted from early development³. These genetic alterations
46 are presumed to lie in enhancers for developmental genes or in other regulatory elements such as
47 promoters for noncoding RNAs that may only be active in the relevant tissue at the appropriate
48 stage of organogenesis. Aside from rare examples⁴, this has remained unproven because of lack of
49 data in human embryos. While regulatory data are available in other species at comparable stages⁵,
50 extrapolation is of limited value because the precise genomic locations of enhancers are poorly
51 conserved^{6,7} even allowing for enriched sequence conservation around developmental genes^{8,9}.
52 Sequence conservation alone is also uninformative for when and in what tissue a putative enhancer
53 might function. Comprehensive regulatory information is available from later fetal development via
54 initiatives such as NIH Roadmap¹⁰ but these later stages largely reflect terminally differentiated,
55 albeit immature cells rather than progenitors responsible for organ formation. In contrast, a small
56 number of studies on a handful of isolated tissues, such as limb bud¹¹, craniofacial processes¹²,
57 pancreas¹³ or brain¹⁴, have demarcated regulatory elements directly during human organogenesis.
58 However, most organs remain unexplored. Moreover, nothing is known about patterns of regulation
59 deployed across tissues, which is an important factor because tissues are often co-affected in
60 developmental disorders. To address these gaps in our knowledge we set out to build maps of
61 genome regulation integrated with transcription during human organogenesis at comprehensiveness
62 currently unattainable from single cell analysis.

63 RESULTS

64 Organs and tissues from thirteen sites were microdissected and subjected to chromatin
65 immunoprecipitation followed by deep sequencing (ChIPseq) for three histone modifications
66 (Figure 1a): H3K4me3, enriched at promoters of transcribed genes; H3K27ac, at active enhancers
67 and some promoters; and H3K27me3 delineating regions of the genome under active repression by
68 Polycomb. Tiny tissue size and the scarcity of human embryonic tissue required some pooling and
69 precluded study of additional modifications. Biological replicates were undertaken for all but two
70 tissue sites (Supplementary Table 1). Tissues and stages were matched to polyadenylated RNAseq
71 datasets acquired at sufficient read depth to identify over 6,000 loci with previously unannotated
72 transcription (Supplementary table 2)¹⁵. Overlaying the data revealed characteristic tissue-specific
73 patterns of promoter and putative enhancer activity, and novel human embryonic transcripts. This
74 was particularly noticeable surrounding genes encoding key developmental transcription factors

75 (TFs), such as the example shown for *NKX2-5* in heart (Figure 1b). Tissues lacking expression of
76 the TF gene tended to carry active H3K27me3 modification (rather than simply lack marks).
77 Putative tissue-specific enhancer marks were characteristically distributed over several hundred
78 kilobases (heart-specific example to the far right of Figure 1b). These isolated H3K27ac marks were
79 often unpredicted by publically available data from cell lines or terminally differentiated lineages
80 and were not necessarily conserved across vertebrates (mean per-base phyloP score 0.175; range -
81 1.42 to +6.94 for n=51,559 regions)¹⁶. Unexpected H3K4me3 and H3K27ac peaks that failed to
82 map to the transcriptional start sites (TSSs) of annotated genes mapped to the TSS of novel human
83 embryo-enriched transcripts, such as the bidirectional *HE-TUCP-C5T408* and *HE-LINC-C5T409*
84 (Figure 1b; see Supplementary File 1H in reference¹⁵ for the complete catalogue).

85 By analysis based on a Hidden Markov Model the genome partitioned into different chromatin
86 states very similarly across tissues¹⁷. While three histone marks allowed for eight different
87 segmentations, aggregation into fewer states was possible (Figure 1c). On average across tissues,
88 3.3% of the genome was active promoter (States 1 & 2; H3K4me3 +/- H3K27ac) or putative
89 enhancer (State 3; H3K27ac) (range 1.7-6.1%; Figure 1c & Supplementary figure 1). 6.7% was
90 variably marked as actively repressed (States 6 & 7; range 3.3-13.0; H3K27me3), while on average
91 89.8% of the genome was effectively unmarked (States 4 & 5; range 81.7-94.0). ~0.2% seemingly
92 had both H3K4me3 and H3K27me3 marks with some detection of H3K27ac (State 8; range 0.16-
93 0.33). This latter state has been considered bivalent and characteristic of ‘poised’ genes whose
94 imminent expression then initiates cell differentiation pathways¹⁸⁻²⁰. Ascribing bivalency has been
95 reliant on setting an arbitrary threshold for whether a site is marked or not, and might simply reflect
96 mixed marks due to heterogeneity in a cell population. To avoid the need for thresholding we
97 clustered promoter profiles for each histone mark integrated with transcription over 3 kb either side
98 of 19,791 distinct protein-coding TSS in each tissue²¹ (Figure 2 and Supplementary figures 2-4).
99 Broader H3K4me3 and H3K27ac signals at the TSS correlated with higher levels of transcription
100 (Figure 2a-b; we termed this promoter state ‘Broad’ expressed versus ‘Narrow’ or ‘Bi-directional’
101 expressed). 25-30% of genes across tissues were unmarked and lacked appreciable transcription
102 (‘Inactive’). These promoters typically lacked CpG islands (<20% compared to 67.7% of the 19,791
103 genes). Conversely, 90-95% of TSS regions marked with H3K27me3 featured CpG islands with an
104 over-representation of TFs; 31.2% of TFs (n=1,659) were actively repressed in at least one tissue
105 compared to 20.0% of non-TF genes (odds ratio 1.82, confidence interval 1.63-2.04; p-value <2.2e-
106 16). H3K27me3 detection at the TSS was ~50% greater for genes encoding TFs (Supplementary
107 figure 5). H3K27me3 was only detected with minimal accompaniment of H3K4me3 or H3K27ac
108 and no transcription (Figure 2 a-b). The categorization for each of the 19,791 genes in all tissues is
109 listed in Supplementary table 3. This neat partitioning would not have been possible if the data were

110 overly confounded by cellular heterogeneity. The data argue against a major bivalent chromatin
111 state at gene promoters in progenitor cells during human organogenesis.

112 This classification allowed us to ask how promoter state changed across different tissues.
113 Tracking all states in all tissues was complex to visualise (Supplementary figure 6). Unifying
114 ‘Broad’, ‘Narrow’ and ‘Bi-directional’ expressed into a single category (‘Expressed’) displayed
115 how the majority of genes remained unaltered across the thirteen tissues (Figure 3a). In contrast,
116 29% of genes had a variable promoter state. Within this subset we predicted that genes responsible
117 for a specific organ’s assembly, such as developmental TFs, would need to be actively excluded or
118 ‘disallowed’ at inappropriate sites (as seen in Figure 1b for *NKX2-5*). We tested this in the
119 replicated datasets by comparing genes transcribed uniquely in one tissue for either inactivity (no
120 mark) or active repression elsewhere (H3K27me₃; disallowed). Gene ontology (GO) analysis of the
121 ‘uniquely expressed/disallowed elsewhere’ gene sets identified the appropriate developmental
122 programme in all instances (as shown for heart in Figure 3b; e.g. ‘heart development’). In contrast,
123 tissue-specific transcription initiated from genes that were simply inactive in other organs tended to
124 highlight differentiated cell function (Figure 3b; e.g. sarcomere organization). These observations
125 highlight the preferential use of H3K27me₃ at the promoters of genes controlling cell fate decisions
126 but not differentiated function. To study this over time, we included datasets from human
127 pluripotent stem cells (hPSCs) and adult tissue, to scrutinize regulatory changes temporally.
128 Different sets of repressed genes lost their H3K27me₃ mark to become expressed as cells
129 transitioned from pluripotency to embryonic pancreatic progenitors or from pancreatic progenitors
130 to mature pancreas (Figure 3c). Surprisingly, the same KEGG term relating to monogenic diabetes
131 emerged in both instances (Figure 3d). However, the genes underlying the first transition related to
132 early function in pancreatic organogenesis, hypoplasia or aplasia; while the genes in the second
133 transition specifically related to post-embryonic pancreatic islet cell differentiation and beta-cell
134 function²².

135 Having recognized the disallowed status of developmental TFs in inappropriate tissues, we
136 wanted to test whether our putative intergenic human embryo-enriched enhancers were capable of
137 driving appropriate reporter gene expression at the correct locations in developing zebrafish. We
138 identified H3K27ac marks that were enriched in the human embryo compared to 161 ENCODE or
139 NIH Roadmap datasets^{10,23} and not detected in the FANTOM5 project²⁴. We developed an
140 algorithm to test for embryonic tissue specificity and filtered for sequence conservation (not
141 necessarily in zebrafish; see Materials & Methods). We manually inspected the remainder for
142 proximity (<1 mb) to genes encoding TFs and, in particular, to increase clinical relevance, to those
143 associated with major developmental disorders. We ensured no H3K4me₃ or polyadenylated
144 transcription in the immediate vicinity (i.e. an unannotated promoter). We tested ten such enhancers

145 out of 44 within 1 mb of *TBX15*, *HEY2*, *ALX1*, *IRX4*, *PITX2*, *HOXD13*, *NKX2-5*, *WT1*, *SOX11* and
146 *SOX9* for their ability to direct appropriate GFP expression in stable lines of transgenic zebrafish
147 (Supplementary table 4). Two (h-003-kid near *WT1* and h-022-mix near *SOX11*) failed to generate
148 any GFP in any location. The remaining eight all yielded GFP at the predicted site in zebrafish
149 embryos (Figure 4 and Supplementary table 4), despite only one of the putative enhancer sequences
150 being conserved in zebrafish (Fig. 4a, h-027-lim near *TBX15*). These data imply that our H3K27ac
151 detection marks novel functional enhancers operating over considerable distance. Moreover, inter-
152 species sequence conservation is not needed for appropriate reporter gene expression (i.e. it is the
153 TFs which bind that are conserved).

154 We wanted to explore the link between these regulatory elements and surrounding gene
155 expression at genome-wide scale. Assured that ChIPseq marks were reproducible within biological
156 replicates without batch effect (Supplementary figure 7) we parsed the genome into 3,087,584 non-
157 overlapping 1 kb bins. Reads within each bin were counted for each mark. Phi correlation between
158 biological replicates indicated this approach to peak calling was very similar to using MACS
159 (Supplementary figure 8). Counts were downsampled and averaged within tissues and correlated
160 with RNAseq data from the same tissue over 1 mb in either direction (i.e. a 2 mb window). On
161 average this window included 44 annotated genes (range: 0-247 genes). For those H3K27ac marks
162 which functioned in zebrafish the strongest correlation was with the appropriate TF gene, for
163 instance *TBX15* over ~500 kb in limb (Figure 4a). Moreover, different H3K27ac marks could be
164 correlated to the same gene, potentially allowing previously unknown enhancers to be grouped, for
165 instance in the adrenal around the adrenal hypoplasia gene, *NR0B1*, located on the X chromosome
166 (Supplementary figure 9).

167 Parsing the ChIPseq data into bins allowed integration of information across tissues, which is
168 challenging when based on empirical modelling by MACS. Placing raw read counts per bin in rank
169 order produced near identical ‘elbow’ plots for all marks in all tissues. This allowed the point of
170 maximum flexure to be used quantitatively for calling marks in a binary ‘yes/no’ fashion (Figure
171 5a). This simplified calling facilitated exploration of regulatory patterns across tissues. Requiring a
172 bin to be marked in any two or more samples identified 48,570 different H3K27ac patterns genome-
173 wide. By genome coverage all tissue-specific patterns ranked within the top one percent (Figure 5b;
174 heart came first). H3K27ac showed far more tissue selective patterns than H3K4me3 or H3K27me3
175 (Figure 4b and Supplementary figures 10-11). Motif analysis on the tissue-specific H3K27ac
176 regions allowed imputation of master TFs for individual tissues, such as NR5A1 in 54.5% of
177 adrenal-specific bins (n=18,411) compared to 25% of the remaining 141,706 bins (Figure 5c).
178 Mutation of *NR5A1* causes adrenal agenesis in human and mouse (OMIM 184757). MEF, TBX and
179 bHLH family members emerged in the heart-specific bins (Figure 5c); all are associated with

180 congenital heart disease²⁵. Having integrated our data we could also uncover regulatory regions that
181 were shared precisely across two or more tissues to explore developmental disorders which
182 manifest in multiple organs. Novel enrichment for composite PITX1/bHLH motifs was found in
183 limb and palate (Figure 5c). GATA binding motifs were enriched in heart and pancreas. Shared
184 patterns could be explicitly instructed by requiring detection in four or more samples
185 (Supplementary figure 12). We hypothesized that patterns shared across many tissues ought to
186 contain elements regulating generic developmental functions. Scrutinising bins marked in over half
187 of all H3K27ac samples (n=30,226 bins versus remaining background of 80,352 bins) identified
188 enrichment for the ETS motif. ETS transcription factors are involved in cell cycle control and
189 proliferation²⁶.

190 Noncoding mutations in promoters or enhancers have been linked increasingly to major
191 developmental disorders^{4,27}. Previously, as part of the Deciphering Developmental Disorders
192 (DDD) study, we studied 7,930 individuals and their parents²⁸. 87% of patients had
193 neurodevelopmental disorders. 10% had congenital heart defects. 68% of patients lacked disease-
194 associated DNMs within exomes ('exome-negative') pointing to the likely importance of the
195 noncoding genome². We sequenced 6,139 non-coding regions (4.2 Mb) selected as ultra-conserved
196 regions (UCRs: n=4,307), experimentally validated enhancers (EVEs: n=595) or as putative heart
197 enhancers (PHE: n=1,237) and found 739 non-coding DNMs². 78% of the 6,139 regions were
198 marked by H3K27ac or H3K4me3 in our embryonic tissues, with a higher percentage overlap for
199 the EVEs (87%) and near perfect overlap for the PHEs (99%) (Figure 6a). An additional 9% were
200 marked by H3K27me3 suggesting non-coding regulation in a currently unsampled tissue. The
201 distribution of DNMs was very similar (Figure 6b). Nearly half of the regions containing DNMs
202 were marked by H3K27ac and/or H3K4me3 that was replicated in at least one tissue. Most
203 commonly, this included the heart or brain, in keeping with the predominance of
204 neurodevelopmental and cardiac phenotypes in the DDD cohort and the PHEs selected for
205 sequencing (Figure 6c). 75% of the PHEs with DNMs mapped to replicated H3K27ac and/or
206 H3K4me3 in our heart dataset. This rose to 100% if the need for replication was removed. We did
207 not observe enrichment for DNMs in patients with heart or limb phenotypes in elements marked by
208 H3K27ac but the power of this test was most likely hampered by low patient numbers (Figure 6d).
209 Enrichment for DNMs in elements marked by H3K27ac was detected for the greater number of
210 cases with neurodevelopmental disorders (1.45-fold, 95% confidence interval 1.09-1.90; p=0.0056)
211 (Figure 6d). This was similar to our previous report using NIH Roadmap H3K27ac and/or DNaseI
212 hypersensitivity data derived from second trimester fetal brain². Our results support a role for non-
213 coding mutations in severe neurodevelopmental disorders and that regulatory marks active during
214 human organogenesis could help stratify disease-relevant non-coding regions.

215 Prioritising DNMs for potential pathogenicity and how they might disrupt surrounding gene
216 function is very challenging. For developmental disorders our mapping allowed focus on enhancers
217 and promoters in the relevant tissue at an appropriate embryonic stage. Our comprehensive tissue-
218 by-tissue catalogue of transcription also allowed more detailed consideration of DNMs in close
219 proximity to previously unappreciated human embryonic noncoding RNAs. Correlating histone
220 modifications with gene expression across all tissues offered a means of prioritising target gene(s).
221 As an example, we identified a G-to-T DNM in a patient with a neurodevelopmental disorder in a
222 UCR on chromosome 16 (Figure 7; chr16:72,427,838). The mutation is situated in the middle of the
223 annotated LINC RNA, *LINC01572*, expressed in testis. Our data illustrated numerous surrounding
224 human embryonic noncoding transcripts. In fact, the DNM was located at the TSS of *HE-OT-*
225 *AC004158.3*, expressed at 19.5-fold higher levels in human embryonic brain than any other tissue
226 (mean read count of quantile normalized transcripts in brain, 1317.2; mean in other tissues, 32.2), in
227 a 4 kb region of brain-specific H3K27ac (and to a lesser extent, H3K4me3). Amongst at least 18
228 protein-coding genes in the surrounding region, the H3K27ac signal was most highly correlated to
229 expression of *ZNF821* ($r=0.92$) located approximately 550 kb away and anticorrelated to expression
230 of the adjacent gene, *ATXNIL* ($r=-0.65$; Figure 7). Taken together, these data and correlations,
231 available to browse as tracks on the UCSC Genome Browser, build a human embryonic atlas of
232 developmental regulatory information linked to gene expression for the overlay of variants
233 identified by clinical sequencing and GWAS.

234 **DISCUSSION**

235 Previous studies of enhancer usage in human embryos have tended to focus on individual tissues
236 inferring, amongst other findings, aspects of genome regulation responsible for human-specific
237 attributes¹¹⁻¹⁴. Here, we incorporated epigenomic data with transcription across thirteen sites during
238 human organogenesis to build tissue-by-tissue maps of enhancers and promoters linked to gene
239 expression. While similar to prior work in mouse⁵ and building on our previous transcriptomic
240 atlas¹⁵, the integrated approach here offers new opportunities to understand how human organ
241 formation is regulated in health and disease.

242 As cost has declined, whole genome sequencing (WGS) has become an important tool in main
243 stream clinical investigation, opening up potential genetic diagnoses in the 98.5% of the human
244 genome that lies outside of coding sequences. However, assessing the non-coding genome is very
245 challenging: millions of rare variants are returned in each individual, while only one might be
246 pathogenic²⁹. Functional analysis, even of a handful of variants, is clinically impractical. For non-
247 coding mutations to affect organogenesis (either in developmental disorders or in later life disease
248 such as type 2 diabetes where there is an embryonic contribution) it is logical that mutations are
249 located in regulatory regions of the genome that are active in post-implantation human embryos. As

250 evident from Figure 1c, our identification of this landscape offers a timely, new pipeline for
251 stratifying 98.5% of the genome down to 3% on average per tissue (States 1-3). Enrichment of
252 tissue-specific TF binding in these enhancers and promoters reinforced our previous findings based
253 solely on computational analysis of 5' flanking regions for the importance of NR5A1 in adrenal and
254 HNF4A in liver¹⁵. However, the integrated sampling of numerous sites uncovered far more complex
255 patterns of regulation operating across tissues. The enrichment of *PITX1* binding motifs in active
256 regulatory regions uniquely shared across limb bud and palate fits with mutations in *PITX1* causing
257 limb defects and cleft palate³⁰. Similarly, GATA4 and GATA6, inferred from regulatory regions
258 shared uniquely between heart and pancreas, are the only two TFs linked to the dual phenotype of
259 cardiac malformation and monogenic diabetes^{31,32}. Overlaying GWAS data with chromosomal
260 conformation studies from older human fetal brain has prioritized target genes for risk of
261 schizophrenia³³. While these techniques are yet to be applied at scale in much smaller human
262 embryonic tissues, because we integrated data from many tissues we could correlate enhancer
263 activity to target genes over megabase distances (Figure 7). Where correlations are linked to
264 expression of the same gene, it then becomes possible to group enhancers. Alongside the need for
265 larger patient cohorts, grouping individual enhancers into larger clusters should increase statistical
266 power, which can be otherwise limiting when causally linking non-coding elements to
267 developmental disorders.

268 Deciphering profiles of H3K27me3 alongside other regulatory marks and expression profiles
269 was also informative. We did not observe bivalent marking of developmental promoters 'poised' for
270 gene expression. Instead, we discovered that organ-specific developmental programmes were
271 disallowed in other human embryonic tissues by active repression at a series of gene promoters. The
272 ontology of these gene sets, including many encoding TFs, inferred they are an important aspect of
273 ensuring correct cell fate decision. This realization opens up a new opportunity for more rigorous
274 benchmarking of differentiated hPSCs, including organoids, both for proximity to the intended
275 lineage in how appropriate gene expression is activated but also against a clearly defined set of
276 epigenomic features for how undesired cell fates are avoided.

277 In summary, we present an integrated atlas of epigenomic regulation and transcription
278 responsible for human organogenesis and make all datasets freely available. The uncovering of
279 novel regulatory regions and patterns of regulation across organs arose because of direct study of
280 human embryonic tissue. The data complement current international projects such as the Human
281 Cell Atlas³⁴, by providing greater resolution of regulatory information and depth of sequence
282 information. Moreover, our integrated analyses establish a new framework for prioritising and
283 interpreting disease-associated variants discovered by WGS³⁵ and provide clear routes towards
284 understanding the underlying mechanisms.

285 **MATERIALS & METHODS**

286 **Sample dissection**

287 Human embryonic material was collected under ethical approval, informed consent and
288 according to the Codes of Practice of the Human Tissue Authority as described previously¹⁵. Tissue
289 collection took place on our co-located clinical academic campus overseen by our research team
290 ensuring immediate transfer to the laboratory. Material was staged by the Carnegie classification
291 and individual tissues and organs were immediately dissected (Supplementary table 1). The material
292 collected here for epigenomic analysis was matched to material isolated for a previous
293 transcriptomic study¹⁵ and the dissection process was identical. In brief, the pancreas, adrenal gland,
294 whole brain, heart, kidney, liver, limb buds, lung, stomach, and anterior two-thirds of the tongue
295 were visible as discrete organs and tissues. All visible adherent mesenchyme, including capsular
296 material (adrenal), was removed under a dissecting microscope. The ureter was removed from the
297 renal pelvis. A window of tissue was removed from the lateral wall of the left ventricle of the heart.
298 The dissected segment of liver avoided the developing gall bladder. The trachea was removed
299 where it entered the lung parenchyma. The stomach was isolated between the gastro-oesophageal
300 and pyloric junctions. The palatal shelves were dissected on either side of the midline. The eye was
301 dissected and the RPE peeled off mechanically from its posterior surface (facilitated by the dark
302 pigmentation of the RPE allowing straightforward visualisation).

303 Tissues were gently teased apart before cross-linking in 1% formaldehyde for 10 min at room
304 temperature. Fixation was quenched with 125 mM glycine for 5 min at room temperature before
305 centrifugation, removal of the supernatant and washing twice with 1 ml PBS. The final PBS
306 supernatant was discarded and samples stored at -80 °C until use (Supplementary table 1).

307 **Chromatin immunoprecipitation (ChIP), RNA isolation and sequencing**

308 All ChIPseq datasets were in biological replicate except for stomach and tongue (Supplementary
309 table 1). Each sample was placed in lysis buffer [10 mM HEPES, 0.5 mM EGTA, 10 mM EDTA,
310 0.25% Triton X-100 and protease inhibitor cocktail (Roche)] on ice for 5 min and nuclei released
311 with 10 strokes in a Dounce homogeniser. Nuclei were pelleted by centrifugation at 700 rcf for 10
312 min at 4 °C and the supernatant discarded. Nuclei were resuspended in ice cold wash buffer (10
313 mM HEPES, 0.5mM EGTA, 1 mM EDTA, 20 mM NaCl and protease inhibitor cocktail) then
314 pelleted by centrifugation at 700 rcf for 10 min at 4 °C and the supernatant discarded. Nuclei were
315 lysed (50 mM Tris HCl, 10 mM EDTA, 1% SDS and protease inhibitor cocktail) and sonicated
316 under prior optimised conditions (Diagenode Bioruptor). Sufficient sample was prepared to allow in
317 parallel immunoprecipitation for H3K4me3, H3K27ac and H3K27me3 to minimise technical
318 variation. 1 µg DNA equivalent was used for each pulldown. Samples were diluted with 9 volumes
319 of dilution buffer (16.7 mM Tris-HCL, 1.2 mM EDTA, 167 mM NaCl, 0.01% SDS and 1.1% Triton

320 X-100). 20 μ l ChIP grade magnetic beads were washed twice in dilution buffer and incubated with
321 each sample for 3 h on a tube rotator at 4 °C to preclear the sample. The beads were separated and
322 the pre-cleared lysate transferred to a separate tube. The magnetic bead pellet was discarded. For
323 each histone modification 3 μ g of antibody (Supplementary table 1) were added to each sample
324 followed by incubation on a tube rotator at 4 °C overnight. 30 μ l magnetic beads were washed twice
325 in immunoprecipitation dilution buffer and incubated with samples for 3 h at 4 °C. Beads were
326 collected and washed twice with wash buffer A (20 mM Tris-HCl, 2 mM EDTA, 50 mM NaCl,
327 0.1% SDS and 1% Triton X-100), once with wash buffer B (10 mM Tris-HCl, 1 mM EDTA, 250
328 mM LiCl, 1% NP40 and 1% Deoxycholate) and twice with TE buffer (10 mM Tris-HCl and 1 mM
329 EDTA). Beads were then incubated in elution buffer (1% SDS and 100 mM NaHCO₃) for 30 min
330 at 65 °C and the beads discarded. The resulting samples were incubated with 167 mM NaCl for 5 h
331 at 65 °C to remove crosslinks followed by 1 h incubation with 14 μ g proteinase K. The resulting
332 chromatin was then purified (MinElute, QIAGEN).

333 DNA libraries were constructed according to the TruSeq® ChIP Sample Preparation Guide
334 (Illumina, Inc.). Briefly, sample DNA (5-10 ng) was blunt-ended and phosphorylated, and a single
335 'A' nucleotide added to the 3' ends of the fragments in preparation for ligation to an adapter with a
336 single base 'T' overhang. Omitting the size selection step, the ligation products were then PCR-
337 amplified to enrich for fragments with adapters on both ends. The final purified product was then
338 quantitated prior to cluster generation on a cBot instrument (Illumina). The loaded flow-cell was
339 sequenced (paired-end) on a HiSeq2500 (Illumina). In total, ChIPseq was carried out in three
340 batches with hierarchical clustering analysis to examine for batch effect (Supplementary figure 7).

341 RNAseq for this study has been described previously¹⁵; using identical methodology, we added
342 single datasets for pancreas and tongue and two datasets for lung to create biological transcriptomic
343 replicates for all tissues (Supplementary table 2).

344 **Mapping of ChIPseq and RNAseq**

345 The first batch of ChIPseq was mapped originally to hg19 using Bowtie 1.0.0 (parameters -m1 -
346 n2 -l28, uniquely mapped reads only)³⁶ and peaks called using MACS2 (2.0.10.20131216)³⁷ against
347 a common input sample (derived from all tissues). To prioritise candidate enhancers for transgenic
348 testing, H3K27ac data from ENCODE (7 cell lines) and NIH Roadmap (154 samples)^{10,23} were
349 mapped similarly. Subsequently, all data, including the external H1 hPSC and adult pancreas data
350 (Figure 3c), were mapped to hg38 using STAR (2.4.2a)³⁸. ChIPseq reads were trimmed to 50 bp for
351 consistency and only uniquely mapped reads were retained. GENCODE 25 gene annotations were
352 used for RNAseq mapping and read counting³⁹.

353

354 **Chromatin and promoter state analysis**

355 Genomic segmentation was performed using chromHMM (version 1.11)¹⁷ labelling samples by
356 tissue and histone modification. The three histone marks allowed for eight segment states.

357 Clustered promoter states were identified for an annotated set of 19,791 protein-coding genes in
358 each tissue using ngs.plot on unnormalized reads for the combined dataset of replicated RNAseq
359 and ChIPseq for H3K4me3, H3K27ac and H3K27me3²¹. Default settings allowed for five clusters
360 based on rank profiles of read counts 3 kb either side of the TSS. The returned clusters were then
361 classified according to characteristics detected in both replicates: ‘Actively repressed’ (H3K27me3
362 signal >50% of maximum and mean transcript counts <10% of maximum); ‘Narrow expressed’
363 [H3K4me3 signal >25% of maximum with > 90% of reads downstream of the TSS and skew >0.65
364 (measured across 100 equidistant percentiles from TSS to +3 kb); and mean transcript counts >10%
365 of maximum]; ‘Broad expressed’ (as for ‘Narrow expressed’ but with skew <0.65); ‘Bi-directional
366 expressed’ (H3K4me3 signal >25% of maximum with <90% of reads downstream of the TSS; and
367 mean transcript counts >10% of maximum); ‘Bi-dir2’ (as for ‘Bi-directional expressed’ but without
368 the H3K4me3 signal); ‘Expressed2’ (H3K4me3 signal >25% of maximum with mean transcript
369 counts <10% of maximum); and ‘Inactive’ (<25 of maximum for H3K4me3 and H3K27me3 and
370 mean transcript counts <10% of maximum). This approach left each gene uniquely assigned to one
371 cluster in any tissue. ‘Bi-dir2’ was only identified in RPE (Supplementary figure 2). ‘Expressed2’
372 was detected in lung, liver and brain (Supplementary figure 3). While superficially this category
373 lacked significant transcription, in fact, total gene-level read counts were very similar to ‘Broad
374 expressed’. However, longer mRNA and longer first introns limited transcript detection at the TSS
375 (Supplementary figure 4). The full listings are in Supplementary table 3.

376 The over-representation of TFs in the TSS regions marked with H3K27me3 and featuring CpG
377 islands was assessed on the dataset of 1,659 genes encoding all the TFs compared against the
378 remaining 18,132 non-TF genes using Fisher’s exact test (two-sided).

379 Alluvial plots were created using the R package Alluvial Diagrams version 0.2-0⁴⁰ with
380 modification of the R code to reorder the horizontal splines (alluvia) within each tissue to keep
381 similar colours together.

382 **Transgenic analysis in zebrafish**

383 A systematic approach identified candidate enhancers that were human embryo-enriched and
384 tissue-specific. We identified marks from the first batch of H3K27ac with RPKM ≥ 25 and ≥ 2.5 -
385 fold enrichment in the human embryo compared to ENCODE (7 cell lines)²³ or NIH Roadmap
386 datasets (154 tissues, including fetal datasets from the second trimester)¹⁰; and that were undetected
387 in the FANTOM5 project²⁴. To filter these embryonic marks for tissue specificity an initial dataset
388 was selected at random and peaks called that were > 200bp. The H3K27ac datasets from other

389 embryonic tissues were then overlaid sequentially in random order. Only called peaks > 200 bp
390 were included. After each addition, any peaks with < 50% overlap between the new and existing
391 dataset were retained. For those retained regions overlapping sequence was filtered out. Once
392 completed, the final set of human embryo-enriched, tissue-specific sequences were again filtered for
393 regions >200 bp. Re-running the tissue specificity algorithm for random addition of datasets
394 resulted in a 99.6% match to the first analysis. These candidate enhancer regions were filtered for
395 sequence conservation (PhastCons LOD score >50)⁴¹ and correlated with surrounding transcription
396 (≤ 1 mb in either direction). We manually inspected the remainder for proximity (<1 mb) to genes
397 encoding TFs associated with major developmental disorders and ensured no H3K4me3 or
398 polyadenylated transcription in the immediate vicinity (i.e. an unannotated promoter). This resulted
399 in 44 candidate enhancers from which we tested ten. The candidate sequences were first cloned in
400 TOPO vector using pCR8/GW/TOPO TA cloning kit (Catalogue number K252020, Invitrogen
401 Thermo Fisher Scientific) and then recombined to the reporter vector Minitol2-GwB-zgata2-GFP-
402 48⁴² using the Gateway LR clonase II Enzyme mix (Cat. No. 11791020, Invitrogen Thermo Fisher
403 Scientific). The reporter vector contains a robust midbrain enhancer as an internal control for
404 transgenesis.

405 Transgenic fish were generated with the Tol2 transposon/transposase method of transgenesis⁴³.
406 *Danio rerio* embryos were collected from natural spawning and injected in the yolk at the one-cell
407 stage. The injection mixture contained 50 ng/ μ l Tol2 transposase mRNA, purified enhancer test
408 vector and 0.05% phenol red. The concentration of the enhancer test vector was between 15 and 30
409 ng/ μ l. Injected embryos were visualized from 24 hpf to 48 hpf in an Olympus stereomicroscope
410 coupled to a fluorescence excitation light source in order to detect the pattern of GFP.

411 Embryos and adults zebrafish were maintained under standard laboratory conditions. They were
412 manipulated according to Spanish and European regulation. All protocols used have been approved
413 by the Ethics Committee of the Andalusian Government (license numbers 450-1839 and 182-41106
414 for CABD-CSIC-UPO).

415 **Genome binning, normalisation and thresholding**

416 The genome was parsed into 3,087,584 non-overlapping contiguous 1 kb bins to compare
417 ChIPseq profiles across tissues and replicates. Reads were counted into bins according to their
418 mapped start position using csaw⁴⁴. Reads from mitochondrial and unplaced chromosome
419 annotations were removed. A further 697 bins were filtered out for possessing >10,000 reads in all
420 samples or if the mean read count from input controls was $\geq 50\%$ of the mean read count of all
421 samples or for being situated in pericentromeric regions (using table ideogram from UCSC; listed in
422 Supplementary table 4). For correlations with surrounding transcription binned read counts were

423 down-sampled statistically using subSeq⁴⁵ weighting each sample by the value of the 99th
424 percentile.

425 Downsampling of read counts to the 99th percentile was used to generate the custom ‘elbow’
426 threshold that called bins as marked or not for subsequent downstream analyses. When read counts
427 were ordered and plotted by rank the resulting graph was typically exponential with most bins
428 having zero or very few reads (below the elbow threshold) and a small number of bins with very
429 high read counts (above the elbow threshold). The elbow was defined as the point on the line with
430 the shortest Euclidian distance to the maximum rank intercept with the x-axis. Our code
431 (ariseFromElbow) to determine these thresholds from a vector of counts is available on github⁴⁶. Phi
432 correlation was used to measure the agreement between tissue replicates called by the 1 kb binning
433 method compared to MACS³⁷. Hierarchical clustering of datasets was undertaken to assess potential
434 batch effect (displayed by heatmap) based on the combined set of the 10,000 most highly ranked
435 bins from each sample. Sets of tissue-specific (replicated in exactly one tissue) and tissue-selective
436 bins (replicated in a given tissue and up to a half of all samples) were produced for each embryonic
437 tissue. EulerGrids showing pattern frequencies of bins across samples were produced using the
438 function plotEuler⁴⁷ as an adaptation of a proposal from Reynolds and colleagues⁴⁸ on
439 Biostars.org⁴⁹.

440 **Annotation set enrichment for genes and genomic regions**

441 Lists of genes and genomic regions (e.g. 1 kb bins) were tested for enrichment of annotations
442 using the R package XGR version 1.1.1 under default parameters⁵⁰. For TSS clusters (e.g. Alluvial
443 plots) only the remaining annotations used in the ngs.plots were included as background.

444 **Motif analysis**

445 HOMER v4.9 was used to search for enriched motifs in selected sets of bins⁵¹. For selected 1 kb
446 bins marked with H3K27Ac, the background set was the remainder of bins with replicated
447 H3K27Ac across all tissues [n=160,043].

448

449

450 REFERENCES

- 451 1. Deciphering Developmental Disorders Study: Prevalence and architecture of de novo
452 mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
- 453 2. Short, P.J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders.
454 *Nature* **555**, 611-616 (2018).
- 455 3. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association
456 studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901 (2017).
- 457 4. Weedon, M.N. et al. Recessive mutations in a distal PTF1A enhancer cause isolated
458 pancreatic agenesis. *Nat Genet* **46**, 61-64 (2014).
- 459 5. van Arensbergen, J. et al. Derepression of Polycomb targets during pancreatic organogenesis
460 allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res* **20**,
461 722-732 (2010).
- 462 6. Nord, A.S. et al. Rapid and pervasive changes in genome-wide enhancer usage during
463 mammalian development. *Cell* **155**, 1521-1531 (2013).
- 464 7. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of
465 transcription factor binding. *Science (New York, N.Y.)* **328**, 1036-1040 (2010).
- 466 8. Dickel, D.E. et al. Ultraconserved Enhancers Are Required for Normal Development. *Cell*
467 **172**, 491-499.e415 (2018).
- 468 9. Woolfe, A. et al. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate
469 Development. *PLOS Biology* **3**, e7 (2004).
- 470 10. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes (Roadmap
471 Epigenomics Consortium). *Nature* **518**, 317-330 (2015).
- 472 11. Cotney, J. et al. The Evolution of Lineage-Specific Regulatory Activities in the Human
473 Embryonic Limb. *Cell* **154**, 185-196 (2013).
- 474 12. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J.P. & Cotney, J. High-Resolution
475 Epigenomic Atlas of Human Embryonic Craniofacial Development. *Cell Rep* **23**, 1581-1597
476 (2018).
- 477 13. Cebola, I. et al. TEAD and YAP regulate the enhancer network of human embryonic
478 pancreatic progenitors. *Nat Cell Biol* **17**, 615-626 (2015).
- 479 14. Reilly, S.K. et al. Evolutionary genomics. Evolutionary changes in promoter and enhancer
480 activity during human corticogenesis. *Science* **347**, 1155-1159 (2015).
- 481 15. Gerrard, D.T. et al. An integrative transcriptomic atlas of organogenesis in human embryos.
482 *eLife* **5**, e15657+ (2016).
- 483 16. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral
484 substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2010).
- 485 17. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and
486 characterization. *Nature Methods* **9**, 215-216 (2012).
- 487 18. Li, F. et al. Bivalent Histone Modifications and Development. *Curr Stem Cell Res Ther* **13**,
488 83-90 (2018).
- 489 19. Harikumar, A. & Meshorer, E. Chromatin remodeling and bivalent histone modifications in
490 embryonic stem cells. *EMBO Rep* **16**, 1609-1619 (2015).
- 491 20. Lesch, B.J. & Page, D.C. Poised chromatin in the mammalian germ line. *Development* **141**,
492 3619-3626 (2014).
- 493 21. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-
494 generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284
495 (2014).
- 496 22. Jennings, R.E., Berry, A.A., Strutt, J.P., Gerrard, D.T. & Hanley, N.A. Human pancreas
497 development. *Development* **142**, 3126-3137 (2015).
- 498 23. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types.
499 *Nature* **473**, 43-49 (2011).
- 500 24. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature*
501 **507**, 455-461 (2014).

- 502 25. Rana, M.S., Christoffels, V.M. & Moorman, A.F. A molecular and genetic outline of cardiac
503 morphogenesis. *Acta Physiol (Oxf)* **207**, 588-615 (2013).
- 504 26. Sizemore, G.M., Pitarresi, J.R., Balakrishnan, S. & Ostrowski, M.C. The ETS family of
505 oncogenic transcription factors in solid tumours. *Nat Rev Cancer* **17**, 337-351 (2017).
- 506 27. Martin, H.C. et al. Quantifying the contribution of recessive coding variation to
507 developmental disorders. *Science* **362**, 1161-1164 (2018).
- 508 28. Firth, H.V. & Wright, C.F. The Deciphering Developmental Disorders (DDD) study. *Dev*
509 *Med Child Neurol* **53**, 702-703 (2011).
- 510 29. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
- 511 30. Klopocki, E. et al. Deletions in PITX1 cause a spectrum of lower-limb malformations
512 including mirror-image polydactyly. *Eur J Hum Genet* **20**, 705-708 (2012).
- 513 31. Allen, H.L. et al. GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nat*
514 *Genet* **44**, 20-22 (2011).
- 515 32. Shaw-Smith, C. et al. GATA4 mutations are a cause of neonatal and childhood-onset
516 diabetes. *Diabetes* **63**, 2888-2894 (2014).
- 517 33. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing
518 human brain. *Nature* **538**, 523-527 (2016).
- 519 34. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
- 520 35. Turnbull, C. et al. The 100 000 Genomes Project: bringing whole genome sequencing to the
521 NHS. *BMJ* **361**, k1687 (2018).
- 522 36. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient
523 alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
- 524 37. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137+
525 (2008).
- 526 38. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford,*
527 *England)* **29**, 15-21 (2013).
- 528 39. Harrow, J. et al. GENCODE: The reference human genome annotation for The ENCODE
529 Project. *Genome Research* **22**, 1760-1774 (2012).
- 530 40. Bojanowski, M. & Edwards, R. alluvial: Alluvial Diagrams. [https://cran.r-](https://cran.r-project.org/web/packages/alluvial/index.html)
531 [project.org/web/packages/alluvial/index.html](https://cran.r-project.org/web/packages/alluvial/index.html) (2016).
- 532 41. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
533 genomes. *Genome Research* **15**, 1034-1050 (2005).
- 534 42. Gehrke, A.R. et al. Deep conservation of wrist and digit enhancers in fish. *Proc Natl Acad*
535 *Sci U S A* **112**, 803-808 (2015).
- 536 43. Kawakami, K. et al. A transposon-mediated gene trap approach identifies developmentally
537 regulated genes in zebrafish. *Dev Cell* **7**, 133-144 (2004).
- 538 44. Lun, A.T.L. & Smyth, G.K. csaw: a Bioconductor package for differential binding analysis
539 of ChIP-seq data using sliding windows. *Nucleic Acids Research* **44**, e45 (2016).
- 540 45. Robinson, D.G. & Storey, J.D. subSeq: Determining Appropriate Sequencing Depth
541 Through Efficient Read Subsampling. *Bioinformatics* **30**, 3424-3426 (2014).
- 542 46. Gerrard, D.T. arseFromElbow. `utilsGerrardDT/arseFromElbow.R` (2019).
- 543 47. Gerrard, D.T. plotEuler. <https://github.com/davetgerrard/utilsGerrardDT> (2019).
- 544 48. Reynolds, A. Venn/Euler Diagram Of Four Or More Sets (BioStars.org).
545 <https://www.biostars.org/p/77362/#77377> (2013).
- 546 49. Parnell, L.D. et al. BioStar: An Online Question & Answer Resource for the Bioinformatics
547 Community. *PLOS Computational Biology* **7**, e1002216 (2011).
- 548 50. Fang, H., Knezevic, B., Burnham, K.L. & Knight, J.C. XGR software for enhanced
549 interpretation of genomic summary data, illustrated by application to immunological traits.
550 *Genome Med* **8**, 129 (2016).
- 551 51. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-
552 regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589
553 (2010).
- 554 52. <http://ms.mcmaster.ca/peter/s743/poissonalpha.html>.

555 53. Erwin, G.D. et al. Integrating diverse datasets improves developmental enhancer prediction.
556 *PLoS Comput Biol* **10**, e1003677 (2014).
557

558 **ACKNOWLEDGEMENTS**

559 We are very grateful to all women who consented to take part in our research programme and for
560 the assistance of research nurses and clinical colleagues at the Manchester University NHS
561 Foundation Trust. We thank Peter Briggs and Andy Hayes of the Bioinformatics and Genomic
562 Technologies Core Facilities at the University of Manchester. The work was supported by
563 Wellcome grants 088566, 097820 and 105610, with additional support from MRC project grants
564 MR/L009986/1 to NB and NAH, MR/J003352/1 to KPH, and MR/000638/1 and MR/S036121/1 to
565 NAH. REJ was an MRC clinical research training fellow and SJW was an MRC doctoral account
566 PhD student. JLGS was supported by the Marató TV3 Foundation (Grant 201611).

567

568 **AUTHOR CONTRIBUTIONS**

569 DTG, AAB and NAH devised the study and planned experiments. KPH, MB, SJW, REJ, ADS
570 and NB were involved in study design and oversight of human embryonic material collection (REJ,
571 KPH). AAB processed the human embryonic material and prepared samples for all sequencing
572 analyses. DTG and ID conducted the bioinformatics analyses. JLGS, SJG, PNF undertook the
573 transgenic analyses. NAH, DTG, PS and MEH conducted the analysis of developmental disorders.
574 DTG and NAH wrote the manuscript with input from AAB and editing from KPH, NB, ADS and
575 JLGS. NAH is the guarantor.

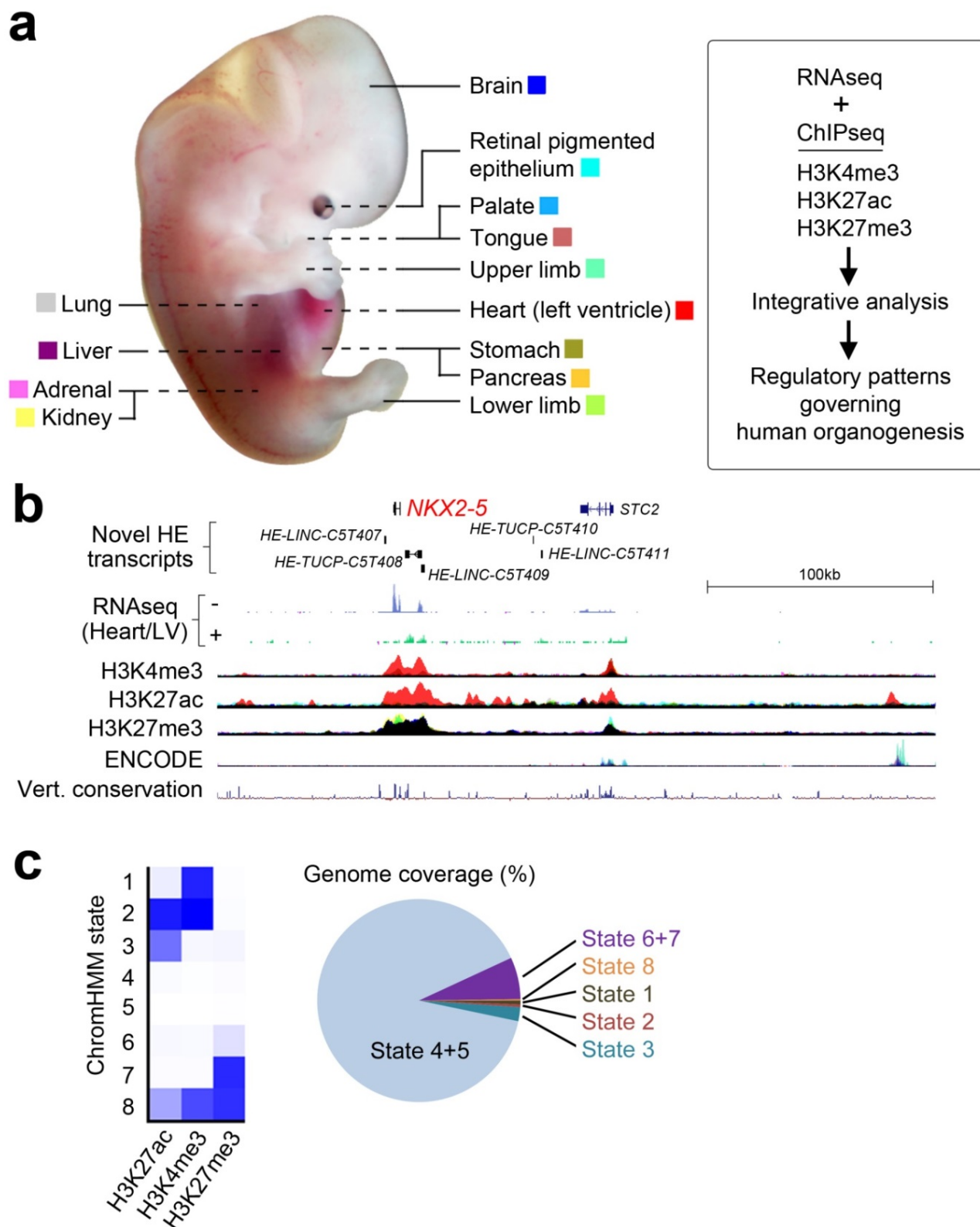
576

577 **AUTHOR INFORMATION**

578 Novel ChIPseq and RNAseq reads have been deposited in the European Genome Phenome
579 repository under accessions: EGAS00001003738 and EGAS00001003163.

580 To view data in the UCSC genome browser, a trackhub is available at
581 <http://www.humandevolutionalbiology.manchester.ac.uk/>. The authors declare no competing
582 financial interests. Correspondence and requests for further information should be addressed to
583 NAH (neil.hanley@manchester.ac.uk).

584 **FIGURES**

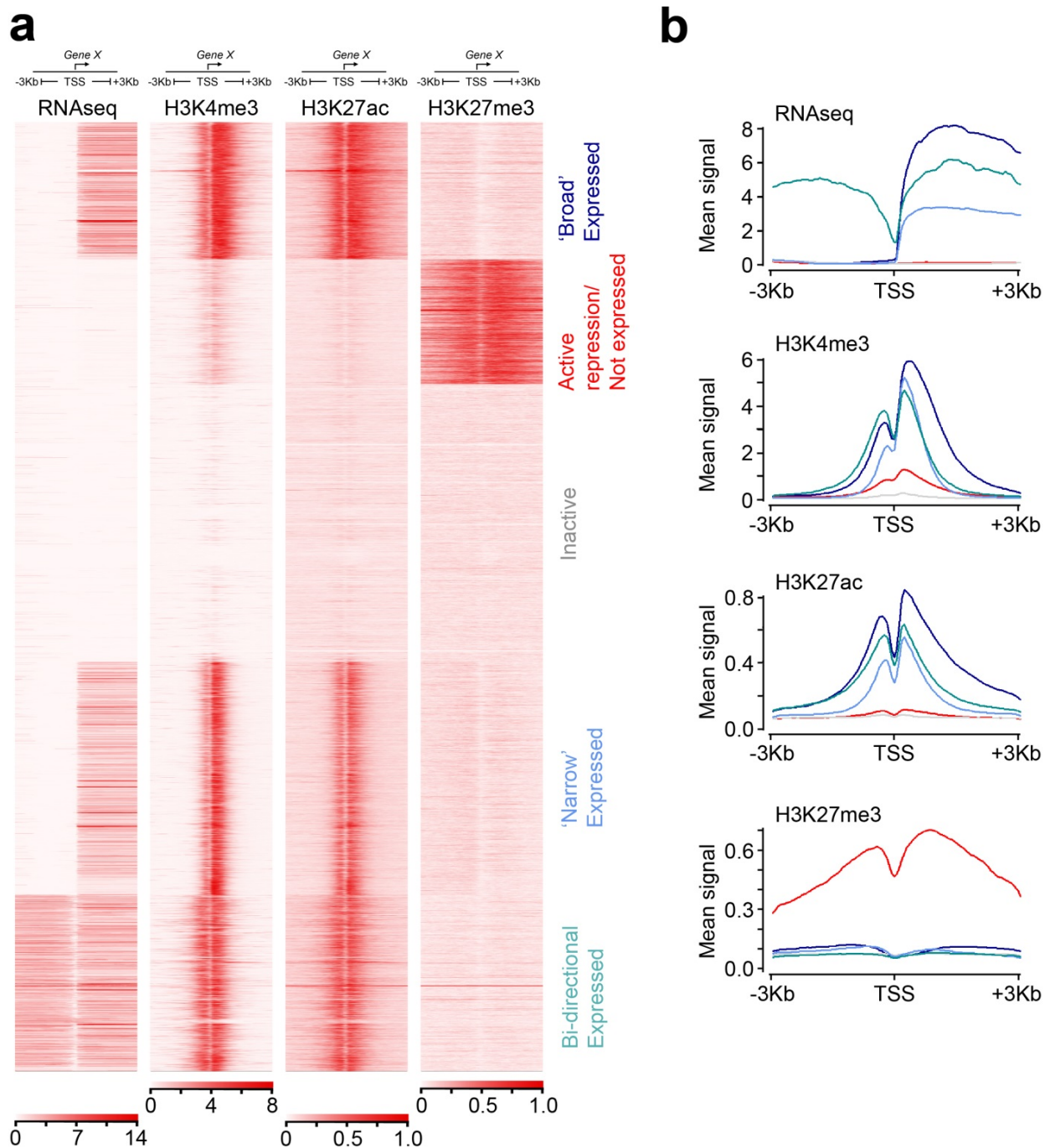


585

586 **Figure 1. Epigenomic landscape across thirteen human embryonic tissues.**

587 a) Thirteen different human embryonic sites were sampled for RNAseq¹⁵ and ChIPseq as described
 588 in the Materials and Methods and in Supplementary tables 1 and 2. The same colour coding for each
 589 tissue is applied throughout the manuscript in overlaid ChIPseq tracks. The heart (left ventricle)

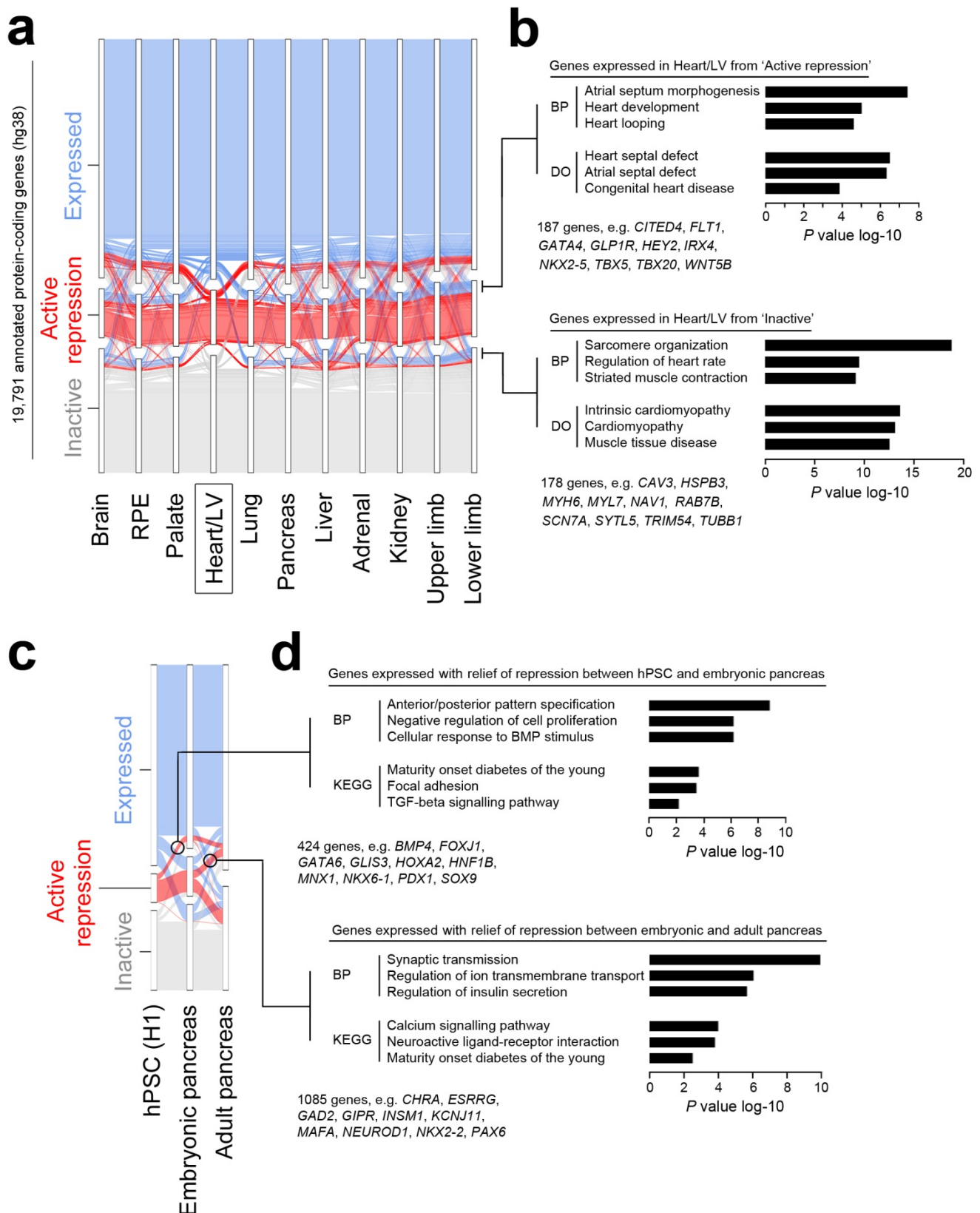
590 dataset is summarised as ‘Heart/LV’ from hereon. b) 300 kb locus around the *NKX2-5* gene, the
591 most discriminatory TF gene for human embryonic heart¹⁵. The locus contains five novel human
592 embryonic (*HE*) transcripts enriched in heart [three *LINC* RNAs and two transcripts of uncertain
593 coding potential (*TUCP*)]. Heart/LV-specific (red) H3K4me3 and H3K27ac marks were detected at
594 the *NKX2-5* TSS and adjacent novel transcripts (*HE-TUCP-C5T408* and *HE-LINC-C5T409*). Novel
595 heart-specific H3K27ac marks were visible up to 200 kb away (e.g. at the extreme right of panel).
596 H3K27me3 marked the region from *NKX2-5* to *HE-LINC-C5T409* in all non-heart tissues (the track
597 appears black from the superimposition of all the different colours other than red). ENCODE data
598 are from seven cell lines²³. c) Genome coverage by ChromHMM for the different histone
599 modifications was similar across all tissues (Supplementary figure 1) with an average 89.8% of the
600 genome unmarked (range: 81.7-94.0; States 4 & 5) and 3.3% consistent with being an active
601 promoter and/or enhancer (range: 1.7-6.1; States 1-3).



602

603 **Figure 2. Classification of genes into discrete states according to characteristics at the**
 604 **promoter associated with transcription.**

605 a) Clustered heatmaps surrounding the transcriptional start sites (TSS +/- 3 kb) of 19,791 annotated
 606 genes. The example shown is for adrenal. One replicate is shown for each data-type for simplicity.
 607 Replicates across all tissues were near identical. Two minor variations on this pattern were detected
 608 in RPE (Supplementary figure 2) and liver, lung and brain (Supplementary figure 3). b) Mean signal
 609 levels for the genes clustered in a). Traces are coloured according to the text colour in a). 'Broad
 610 expressed' genes show approximately double the level of transcription and twice the width of
 611 H3K4me3 and H3K27ac marks compared to 'Narrow expressed' genes.

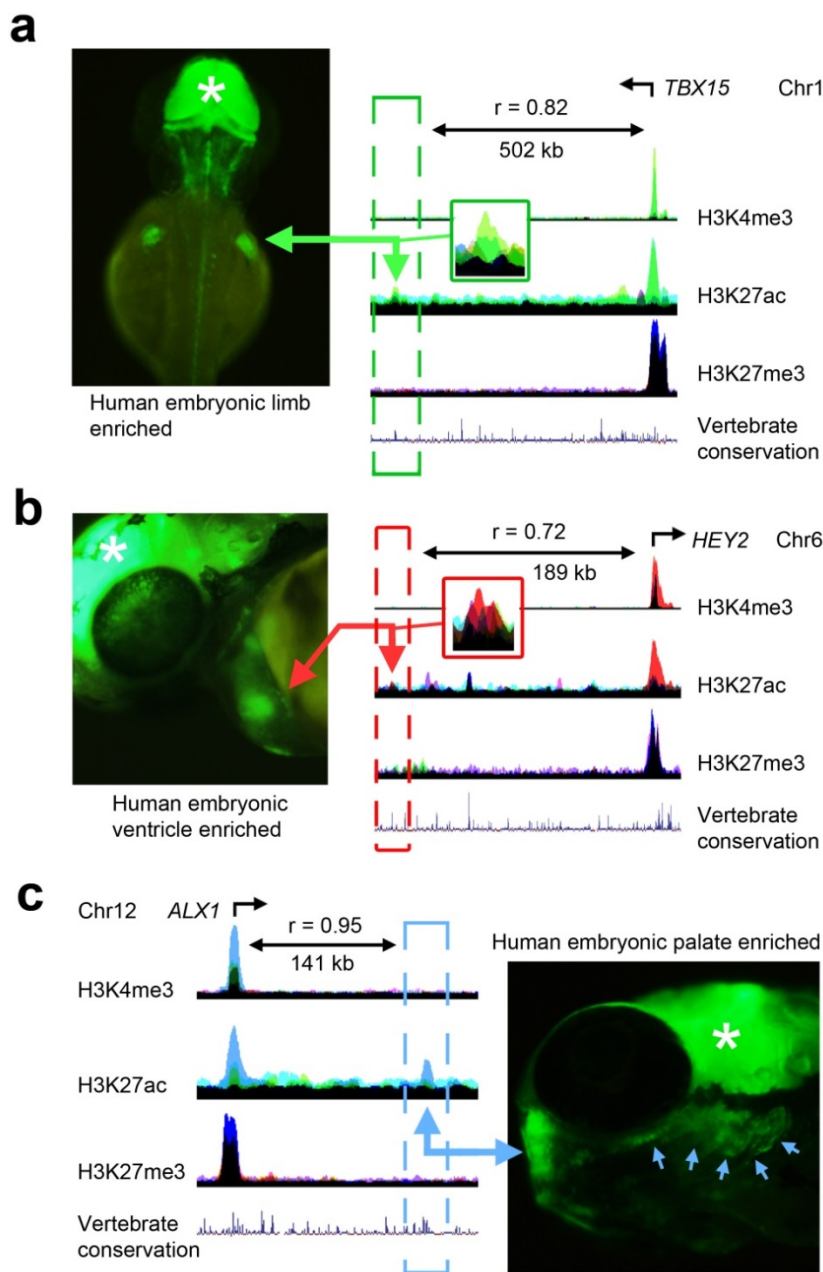


612

613 **Figure 3. Integration of promoter states across tissues and over time to decipher**
 614 **developmental programmes and associated disease in individual organs.**

615 a) Alluvial plot showing promoter state for 19,791 annotated genes across all tissues with replicated
 616 datasets. To aid visualisation all the different transcribed states are amalgamated into a single
 617 'Expressed' category (the alluvial plot for all individual states is shown in Supplementary figure 6).

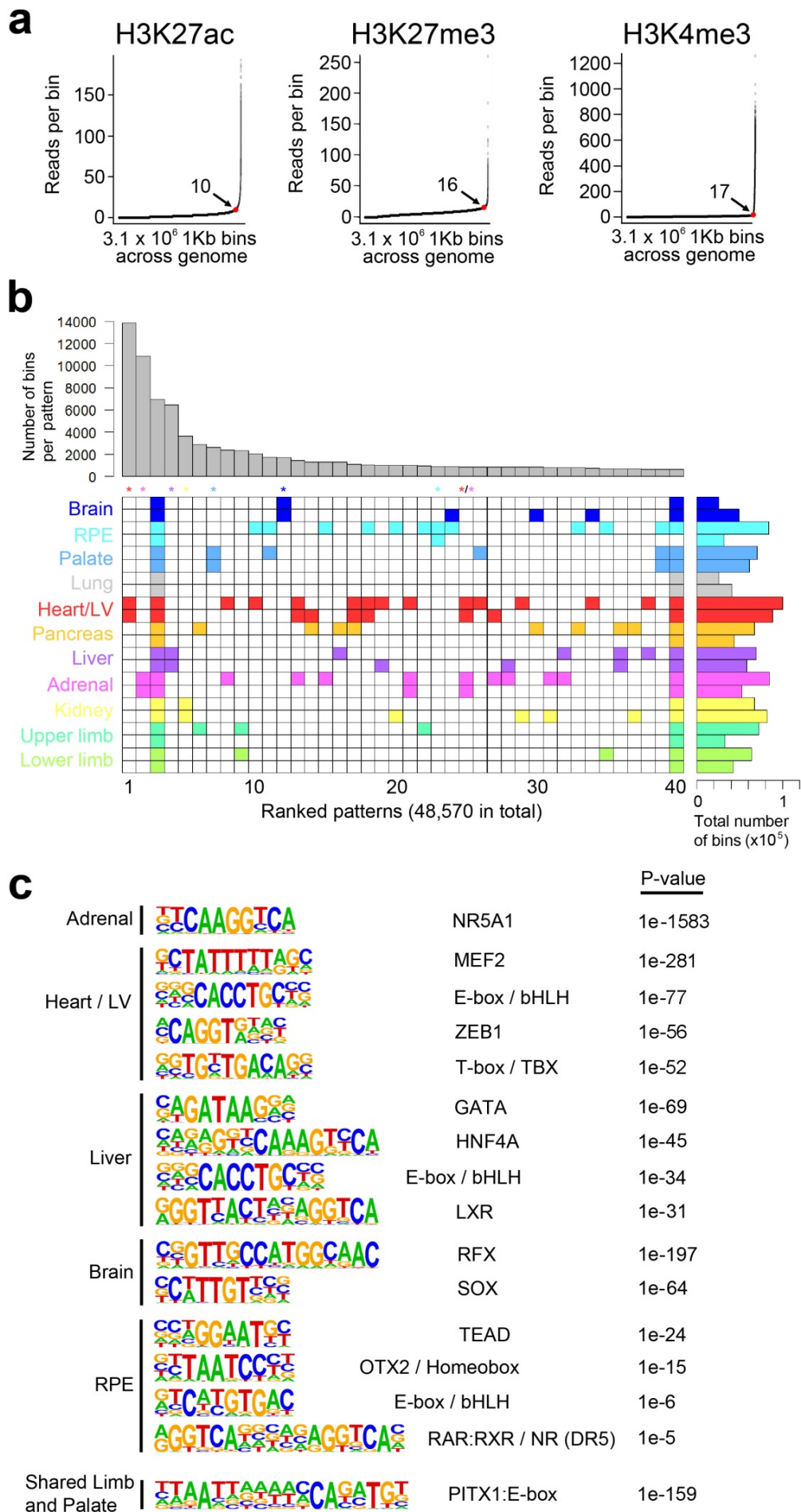
618 The example shown is centred on the promoter state in the Heart/LV dataset. Those genes with an
619 ‘Expressed’ promoter state in heart and either ‘Active repression’ or ‘Inactive’ elsewhere are
620 indicated to the right of the panel and subject to gene enrichment analyses in b). b) Gene enrichment
621 analysis of genes with an ‘Expressed’ promoter state in heart and either ‘Active repression’ or
622 ‘Inactive’ in all remaining tissues. Examples of the genes underlying the biological process (BP) or
623 disease ontology (DO) terms and their total number are listed beneath the bar charts. c) Alluvial plot
624 showing the variance in promoter state between H1 human pluripotent stem cells (hPSCs), the
625 embryonic pancreas (prior to endocrine differentiation²²) and the adult pancreas. Circles capture
626 those genes that shift from ‘Active repression’ to ‘Expressed’ at the stage of either embryonic or
627 adult pancreas. d) Gene enrichment analyses of encircled genes from c). Examples of the genes
628 underlying the BP and KEGG terms and their total number are listed beneath the bar charts. While
629 maturity onset diabetes of the young emerged in both analyses, the underlying genes were different
630 reflecting developmental roles prior to or after pancreatic endocrine differentiation²².
631
632



633

634 **Figure 4. Stable transgenic analysis in developing zebrafish of regions marked by H3K27ac**
 635 **in human embryonic tissues.**

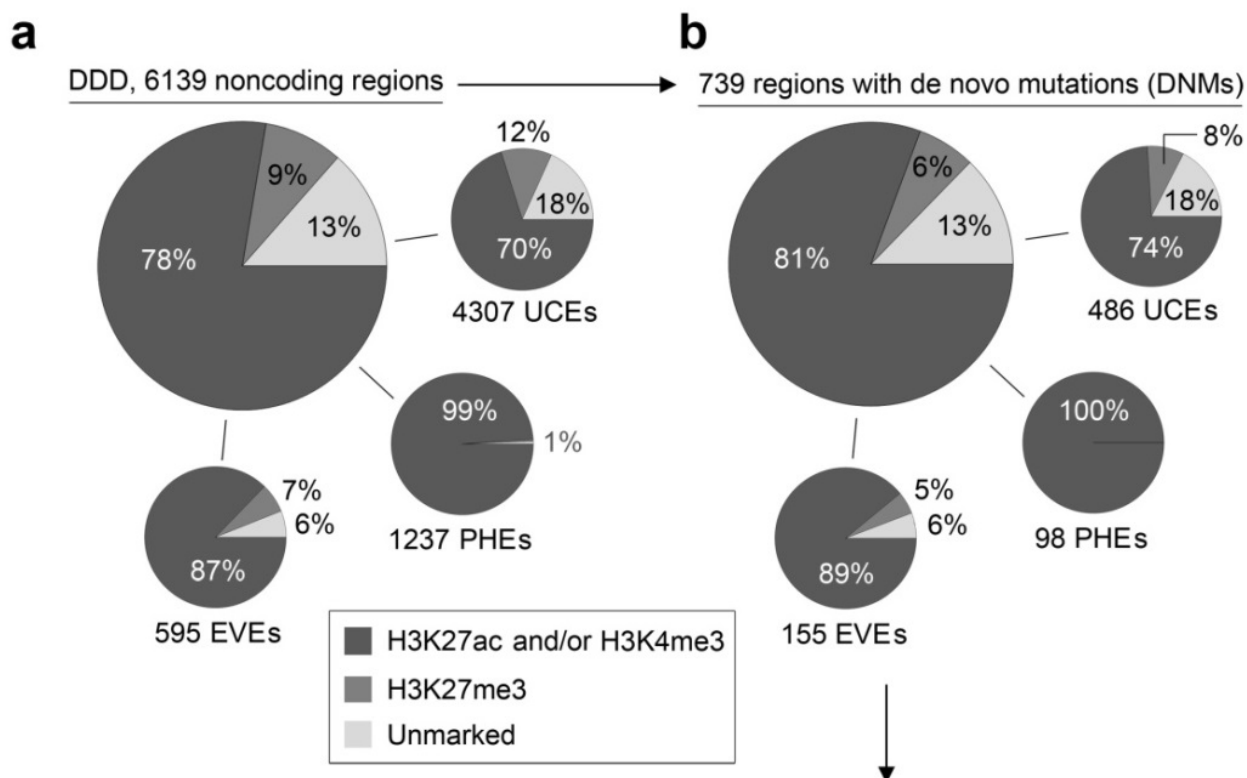
636 H3K27ac marked regions were tested in multiple lines of stable transgenic zebrafish (details in
 637 Supplementary table 4; same colour coding of tracks as in Figure 1). a) 231 bp limb enhancer, 502
 638 kb downstream of *TBX15*, with corresponding green fluorescent protein (GFP) detection in fin bud
 639 at 48 hours post-fertilisation (hpf). b) 355 bp heart/LV enhancer, 189 kb upstream of *HEY2*, with
 640 corresponding ventricular GFP detection at 48 hpf. c) 1.5 kb palate enhancer, 141 kb downstream of
 641 *ALX1*, with GFP in the developing trabecula and mandible (blue arrows) at 48 hpf. Correlations
 642 between the enhancer and transcription of the transcription factor are shown for each example. Note
 643 the H3K27me3 marks over the gene in each instance in other tissues. *, midbrain GFP expression
 644 from the integral enhancer in the reporter vector used as positive control for transgenesis.



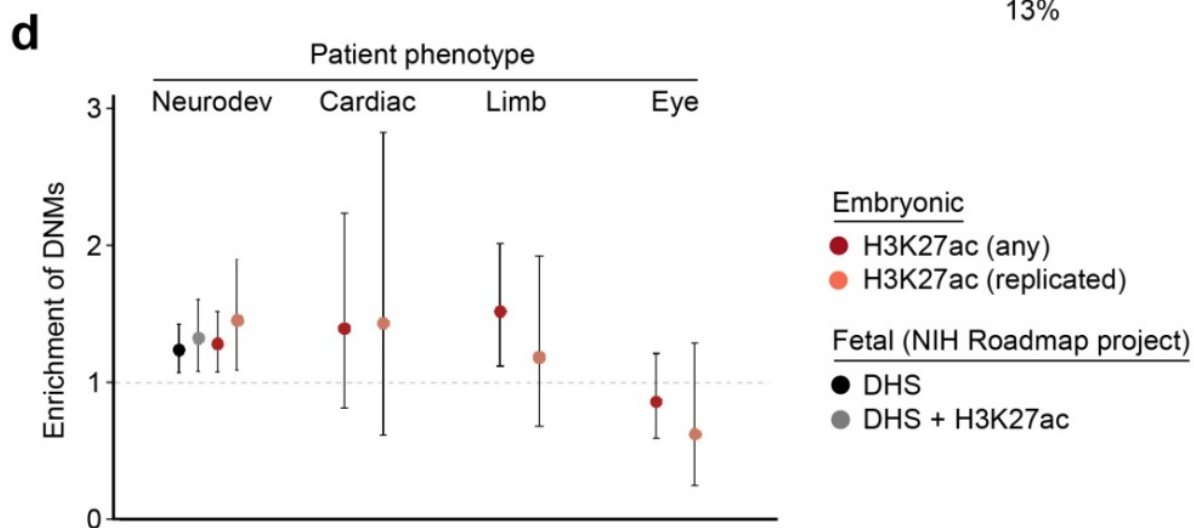
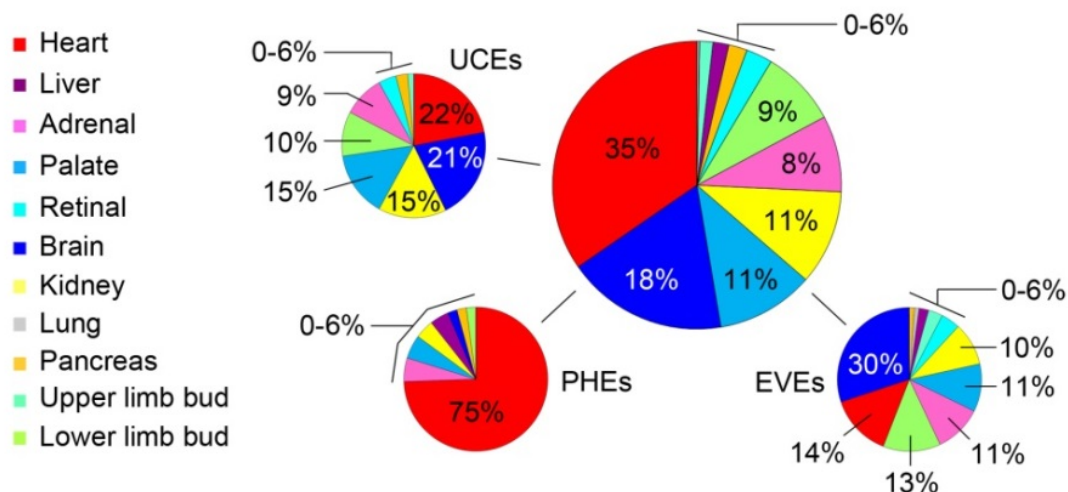
646 **Figure 5 (previous page). Imputing patterns of enhancer activity and regulation by**
647 **developmental transcription factors across human embryonic tissues.**

648 a) Elbow plots for each histone modification following allocation of the genome into 3.1 million
649 consecutive bins of 1 kb. The example shown is for adrenal providing the number of reads per bin
650 at the point of maximum gradient change (the ‘elbow point’, red dot) and a quantitative measure of
651 whether a bin was marked or not (e.g. >10 or <10 respectively for H3K27ac). Converting marks
652 into a binary ‘yes/no’ call at any point in the genome facilitated data integration across the different
653 tissues. While the number of reads per bin at the elbow point was different for each mark across the
654 tissues the shape of the curve remained the same. b) Euler grid for bins marked by H3K27ac
655 (defined by elbow plots) in replicated tissues (i.e. two rows/replicates per tissue). Total number of
656 marked bins per individual dataset is shown to the right. The example in b) required a bin to be
657 called in any two or more samples and is ordered by decreasing bin count per pattern (bar chart
658 above the grid). A total of 48,570 different patterns were identified of which the top 40 are shown.
659 Tissue-specificity for all sites emerged in the top 265 (0.5%) patterns; colour-coded asterisks above
660 columns). For example, nearly 14,000 bins marked only in the two Heart/LV H3K27ac datasets
661 ranked first as the most frequent pattern. The seventh most frequent pattern in nearly 3,000 bins was
662 palate-specific. Tissue-specific patterns were far less apparent at promoters (H3K4me3, n=18,432;
663 Supplementary figure 10) or for H3K27me3 (n=26,339; Supplementary figure 11). While patterns
664 across multiple tissues were permitted by stipulating marks in ≥ 2 samples (e.g. heart and adrenal in
665 column 24), they could be enforced by stipulating marks in at least four samples (Supplementary
666 figure 12). c) Enrichment of known TF-binding motifs in the tissue-specific patterns of H3K27ac
667 identified in b). Five individual tissues are shown as examples alongside analysis of the shared
668 regulatory pattern identified for limb and palate identifying marked enrichment of a compound
669 PITX1:E-box motif.

670



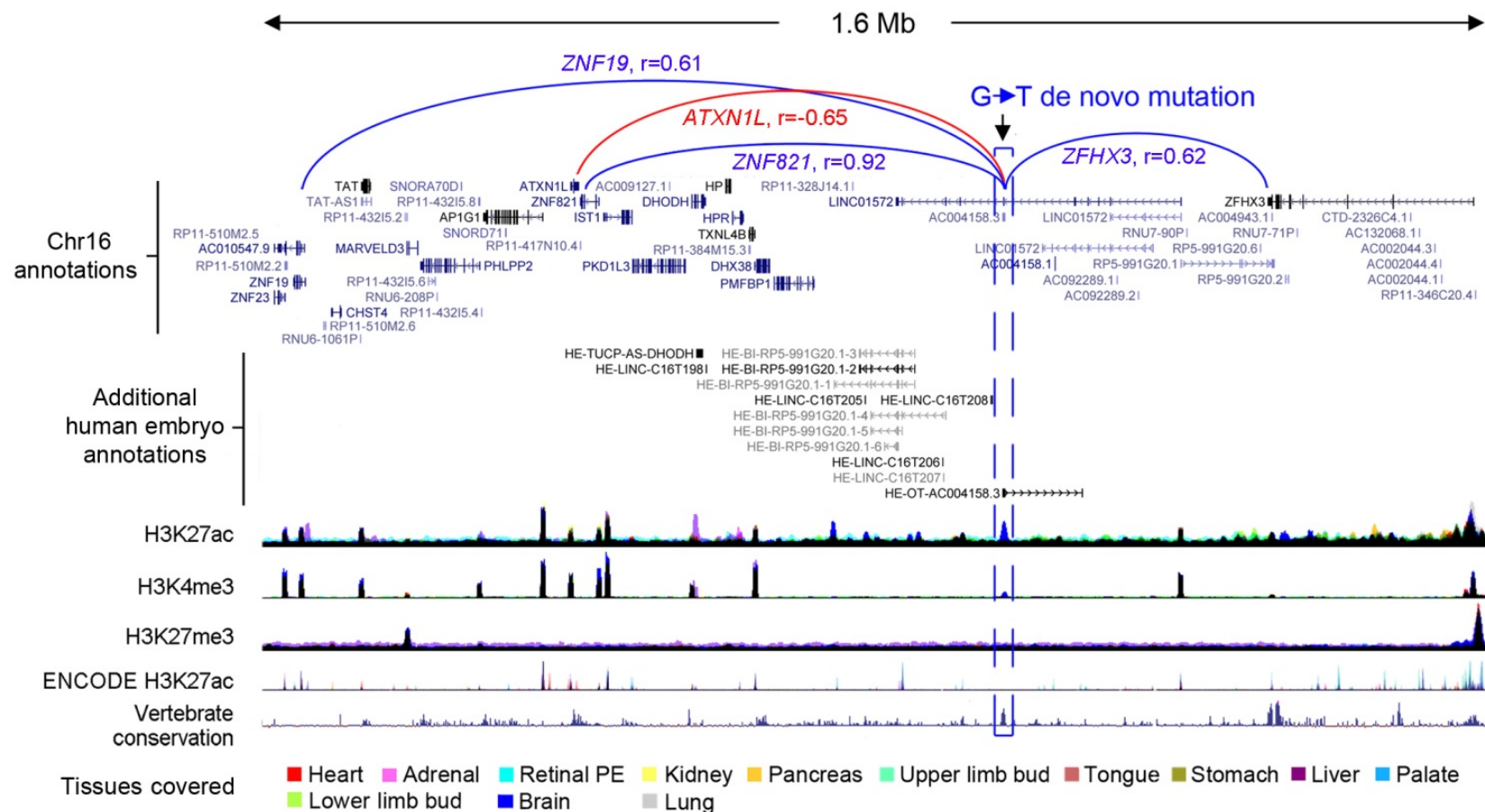
c 338/739 (46%) overlay H3K27ac and/or H3K4me3 replicated in ≥ 1 tissue



671
672

673 **Figure 6 (previous page). Intersection of the epigenomic landscape during human**
674 **organogenesis with noncoding de novo mutations linked to developmental disorders in**
675 **patients.**

676 a) The Deciphering Developmental Disorders (DDD) study included 6,139 noncoding regions in its
677 sequence analysis of trios comprising affected individuals and unaffected parents^{2,28}. These
678 noncoding regions were selected on the basis of high sequence conservation (ultra-conserved
679 elements, UCEs, n=4,307), experimental validation (experimentally validated enhancers, EVEs,
680 n=595) or identification as a putative heart enhancer (PHE, n=1,237). Overlap with any H3K27ac,
681 H3K4me3 or H3K27me3 1 kb bins is shown as an aggregate and for each individual category
682 (UCE, EVE or PHE). b) Equivalent overlap is shown for the 739 regions in which disease-
683 associated de novo mutations (DNMs) were identified. c) 46% of DNM-positive regions were
684 situated (+/- 1 kb) in at least one tissue-replicated H3K27ac and / or H3K4me3 bin. Over half of the
685 disease-associated overlap was covered by heart/LV (35%) and brain (18%). 75% of the disease-
686 associated PHE regions were situated within 1kb of a heart/LV-specific histone mark. d)
687 Enrichment (observed/expected) in the number of DNMs overlapping (+/- 1 kb) H3K27ac marks
688 during human organogenesis. For neurodevelopmental phenotypes this included analysis against
689 DNase hypersensitivity data and H3K27ac data from second trimester fetal brain¹⁰. Error bars show
690 the 95% confidence limits for the mean calculated for a Poisson distribution⁵².



691

692 **Figure 7. Overlap of individual disease-associated de novo mutations with the human embryonic epigenome correlated to surrounding**
 693 **gene expression.**

694 An intergenic G-to-T de novo mutation (DNM; hg38, chr16:72427838) is shown for a patient with a neurodevelopmental phenotype. Tracks are shown
 695 demonstrating novel human embryonic noncoding transcription (enriched in human embryonic brain), the three epigenomic marks, ENCODE data²³
 696 and conservation amongst vertebrates. The DNM overlaps a brain-specific (dark blue) H3K27ac and small H3K4me3 mark. The highest correlations
 697 are shown, notably to the promoters of *ZNF821* (r=0.92) (dark blue) with anticorrelation (r=-0.65, red) to the adjacent gene, *ATXN1L*.

698 **SUPPLEMENTARY TABLES**

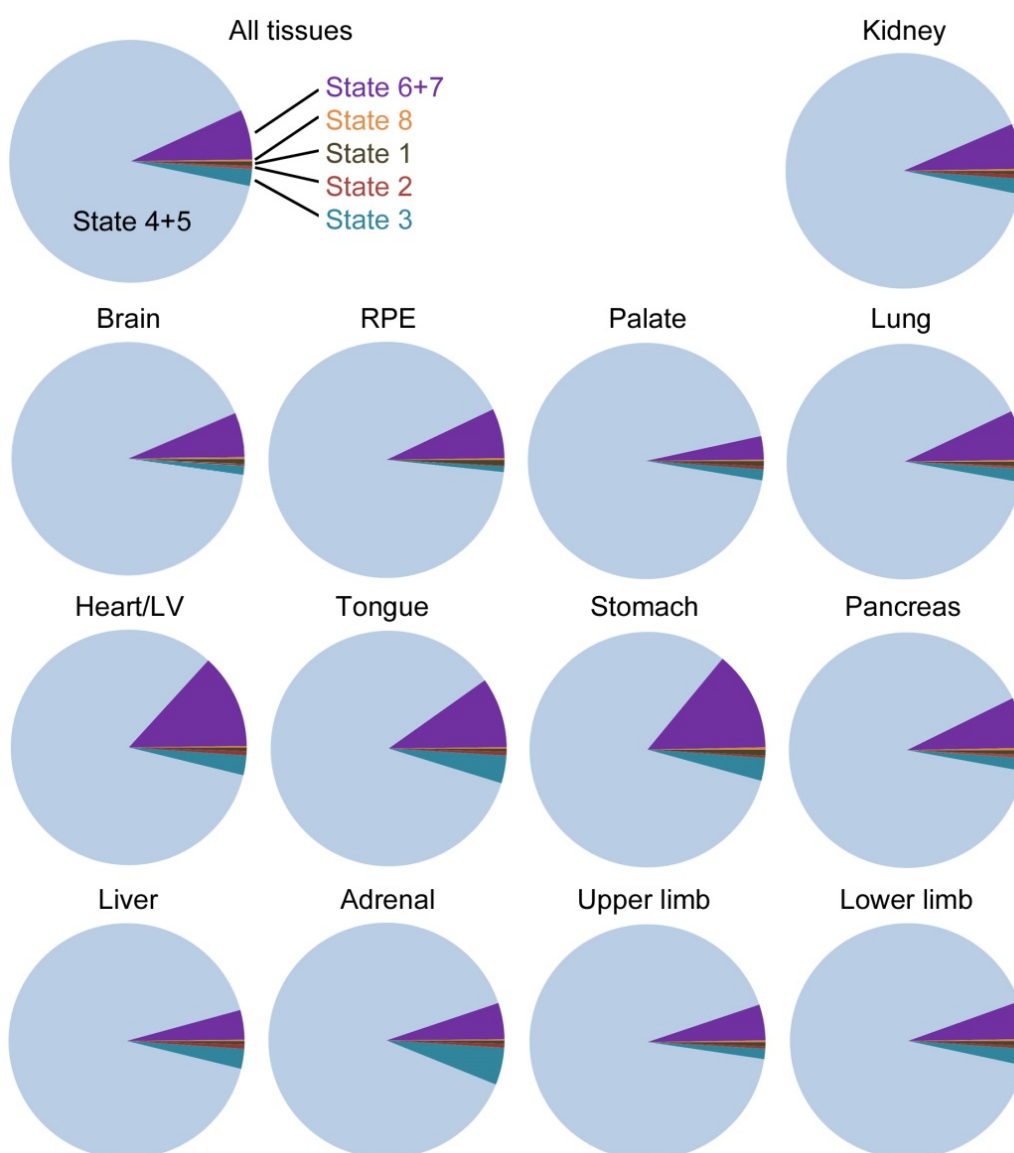
699 Five supplementary tables are available in the appended file (.xls). The legends for individual
700 supplementary tables are contained within the file worksheets.

701

702

703 **SUPPLEMENTARY FIGURES**

704

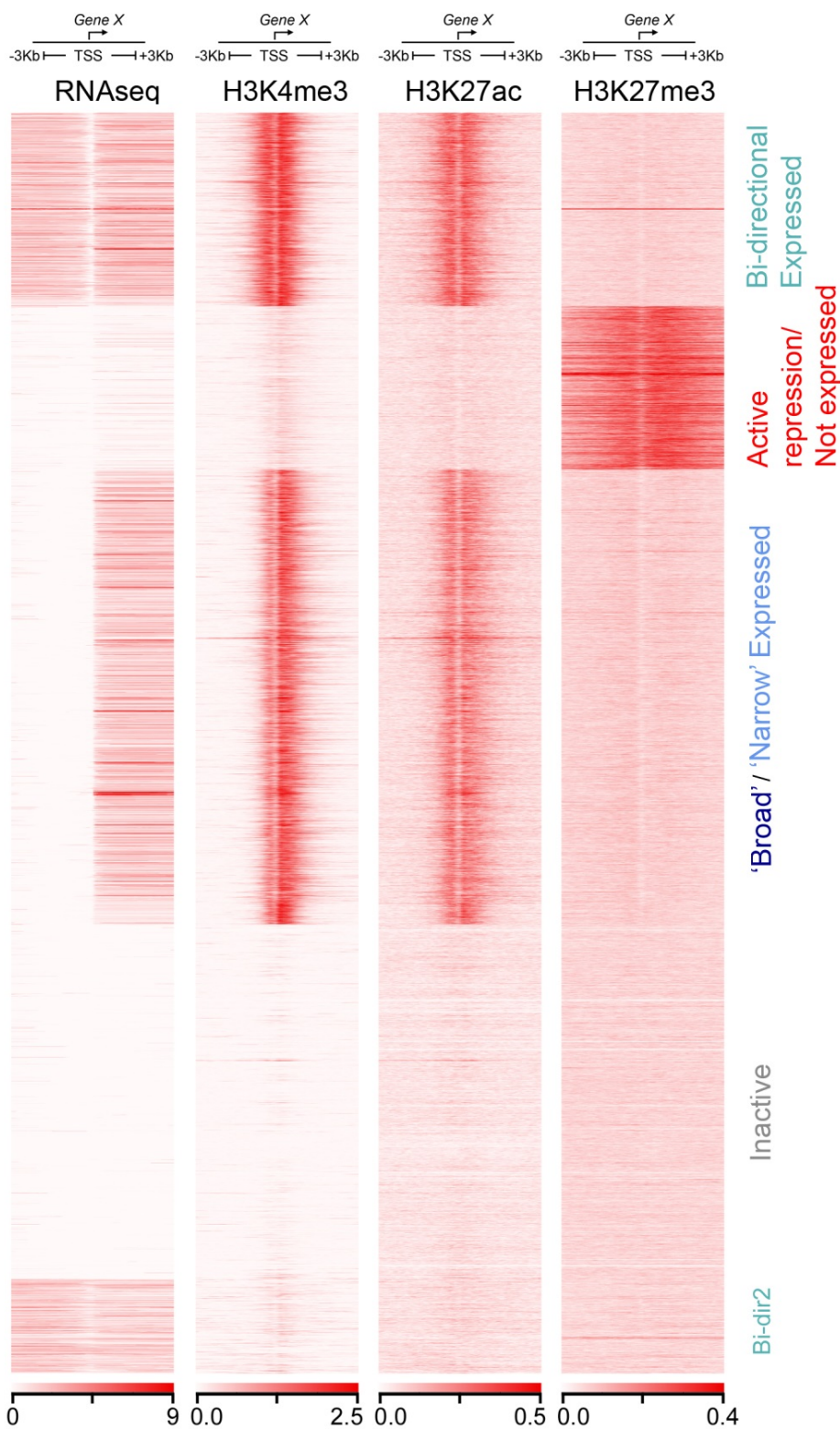


705

706 **Supplementary figure 1 (relates to Figure 1). Genome coverage for the different histone**
707 **modifications in all tissues.**

708 Pie charts for individual tissues of genome coverage according to chromatin state by ChromHMM.

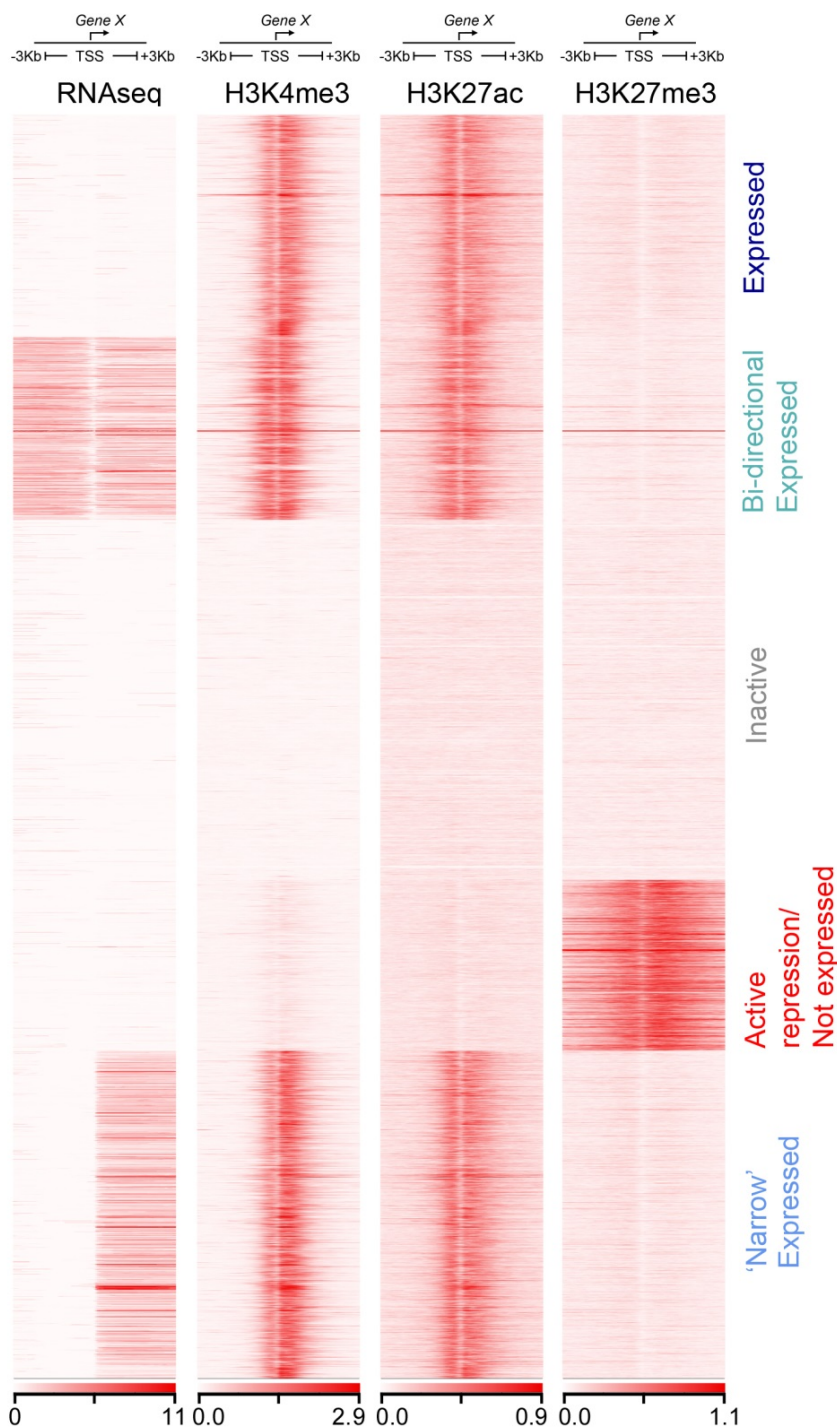
709 The average for all tissues is shown in Figure 1c.



710

711 **Supplementary figure 2 (relates to Figure 2). Variant promoter state observed for retinal**
712 **pigmented epithelium (RPE).**

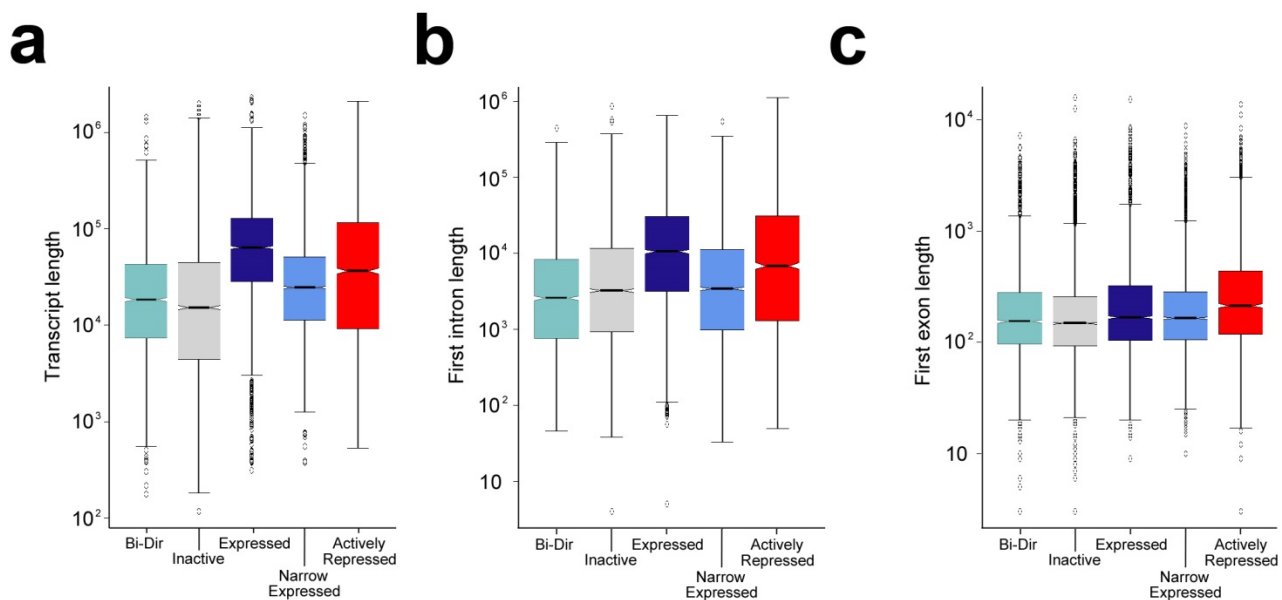
713 In RPE the 'broad' and 'narrow' expressed categories identified in most tissues and shown in
714 Figure 2a were aggregated by ngsplot and termed 'broad/narrow' expressed. This occurred in
715 favour and as a consequence of clustering a subset of genes with less robust bidirectional
716 transcription and barely any marking for H3K4me3 or H3K27ac (termed 'Bi-dir2').



717

718 **Supplementary figure 3 (relates to Figure 2). Variant promoter states observed for brain,**
719 **lung and liver.**

720 In brain, lung and liver the 'Broad expressed' category, evident in Figure 2a, is marked as
721 'Expressed' (example shown for lung). While the H3K4me3 and H3K27ac marks were
722 indistinguishable from the 'Broad expressed' category of Figure 2a, there was no accompanying
723 RNaseq detection at the TSS. These genes encoded longer transcripts with characteristically long
724 first introns (Supplementary figure 4) leading to an under-representation of RNaseq reads at the
725 TSS. Total transcript count across the entire gene was equivalent for 'Broad expressed' (Figure 2a)
726 and 'Expressed' genes.



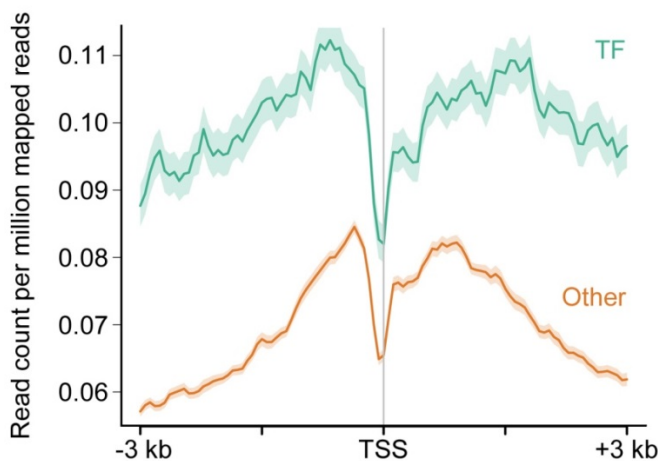
727

728 **Supplementary figure 4 (relates to Figure 2 and Supplementary figure 3). Characteristics of**
 729 **the 'Expressed' promoter state (brain, lung and liver).**

730 a) Transcript length b) first intron length and c) first exon length for the categories of genes detected
 731 in brain, lung and liver (Supplementary figure 3). Although overall transcript counts for 'Expressed'
 732 were similar to 'Broad Expressed' (Figure 2a), RNAseq was not detected at the TSS due to longer
 733 first introns and overall transcript length. The first exon was similar to other categories. The
 734 example shown is for lung with the same findings observed for brain and liver.

735

736

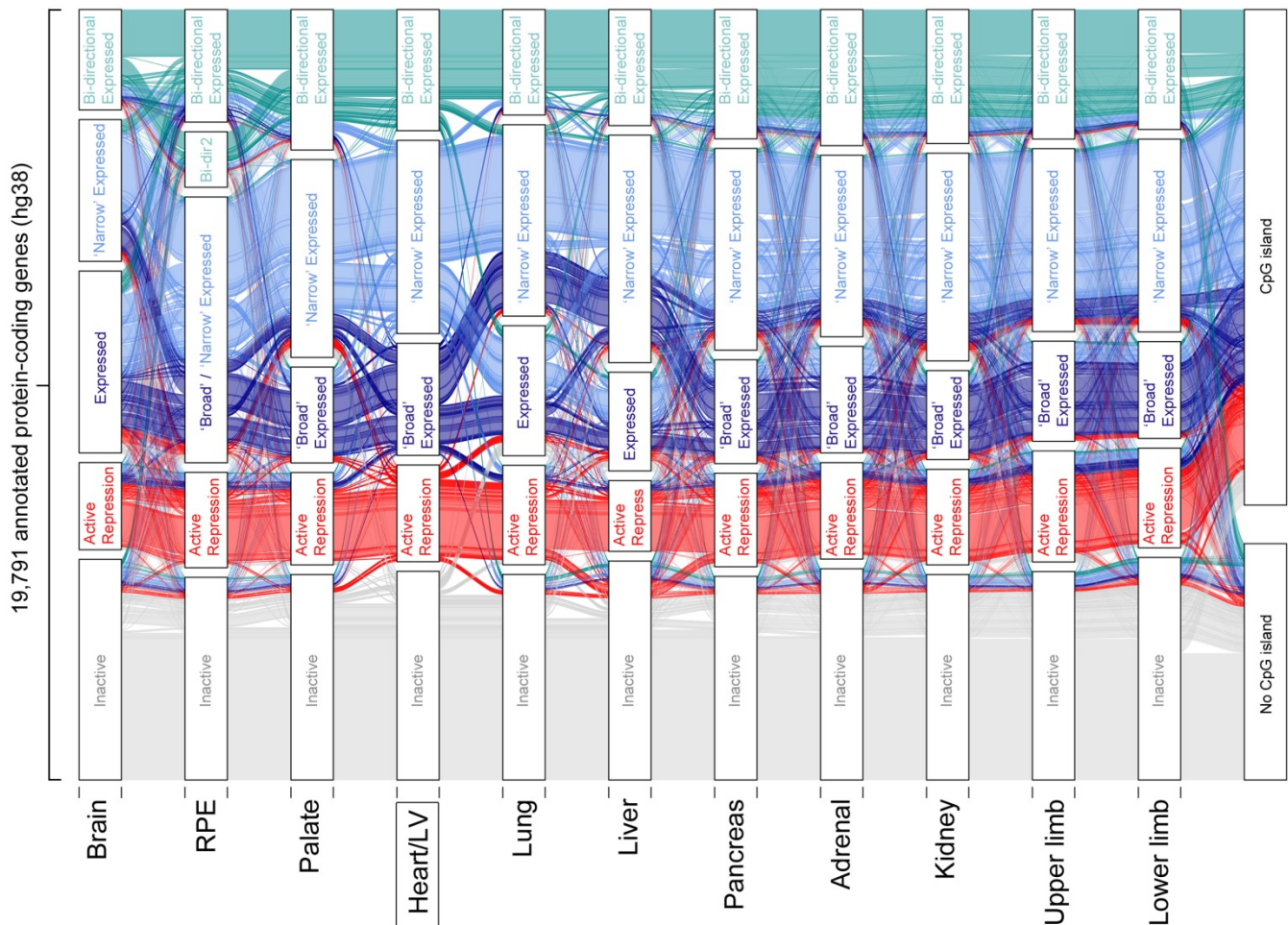


737

738 **Supplementary figure 5 (relates to Figure 2). Levels of H3K27me3 at the TSS of actively**
 739 **repressed genes which encoded transcription factors (TF) compared to those encoding all**
 740 **other proteins (Other).**

741 Within the 'Active Repression' category across all tissues genes encoding transcription factors
 742 (TFs) possessed appreciably greater marking with H3K27me3 at their transcriptional start site
 743 (TSS) compared to those genes encoding other proteins.

744

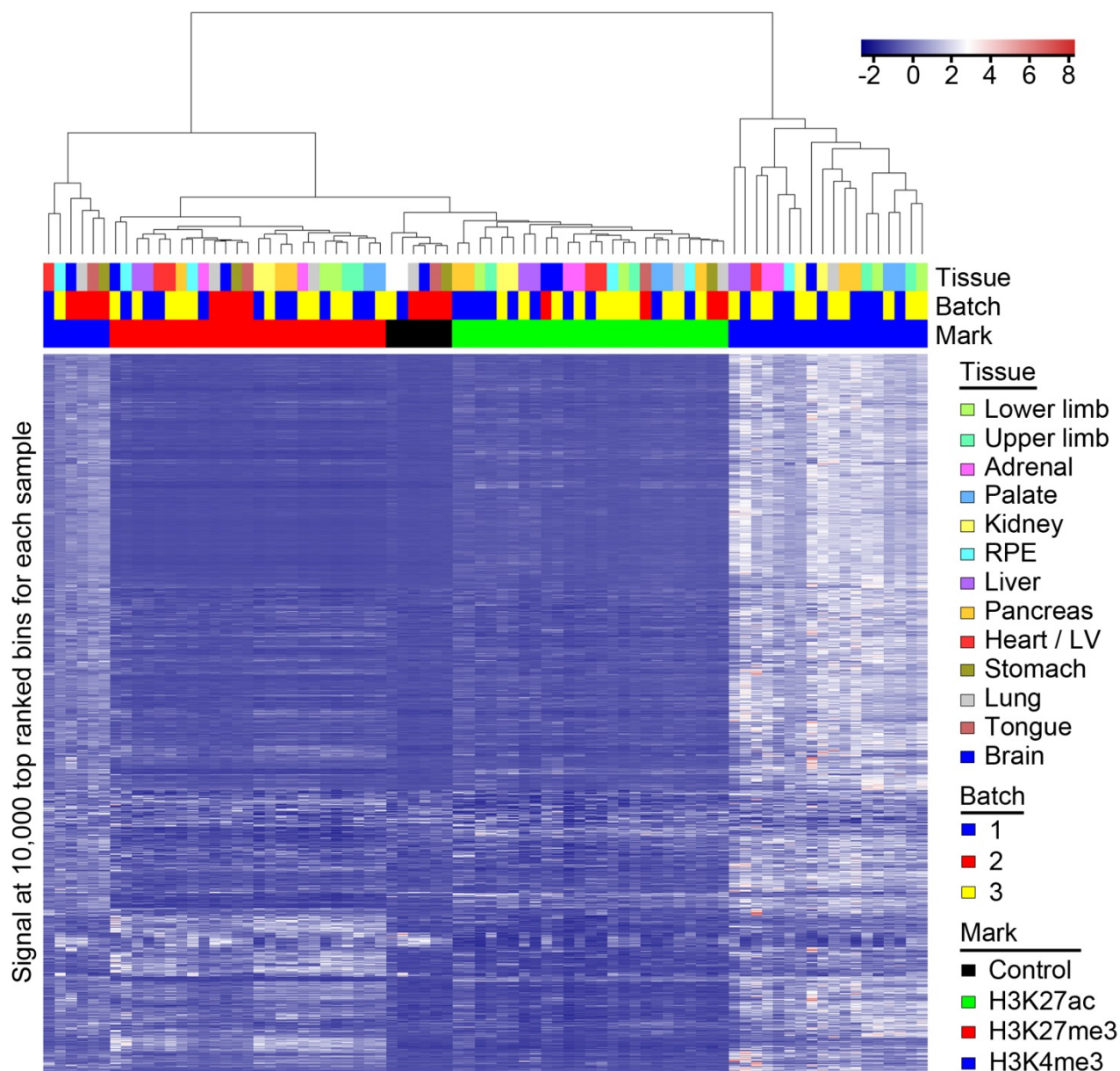


745

746 **Supplementary figure 6 (relates to Figure 3). Integration of all identified promoter states**
747 **across tissues.**

748 Alluvial plot showing promoter state for 19,791 annotated genes across all tissues with replicated
749 datasets. All the amalgamated promoter states associated with gene transcription in Figure 3 are
750 shown individually here: 'Broad expressed', 'Narrow expressed', 'Expressed', 'Bidir' and 'Bidir2'.
751 The plot shown is centred on (and with variance from) the promoter state in the Heart/LV dataset.
752 The plot also categorises genes according to the presence or absence of a CpG island at the
753 promoter. Genes with an 'Inactive' promoter state characteristically lacked a CpG island. Promoters
754 either actively transcribed or repressed tended to possess a CpG island.

755



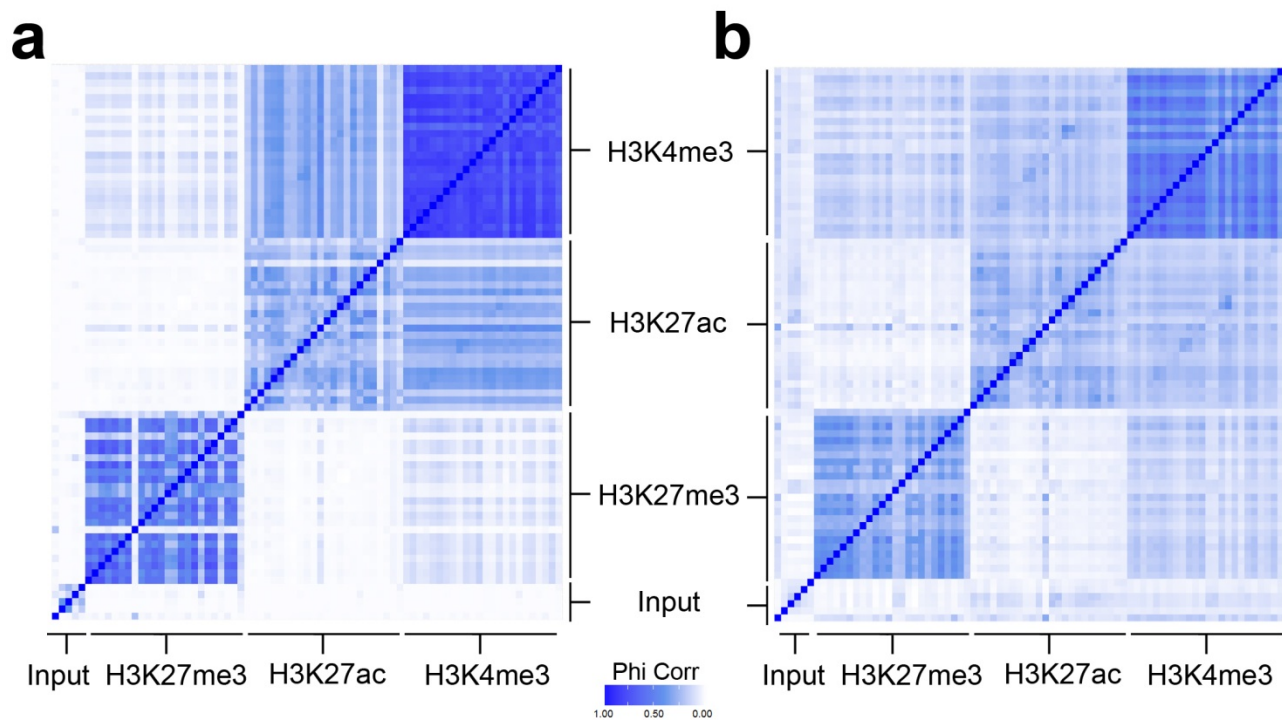
756

757

758 **Supplementary figure 7.**

759 Heatmap showing hierarchical clustering of ChIPseq datasets based on the combined set of 10,000
760 most highly ranked bins from each sample. Samples clustered according to mark and did not cluster
761 according to sequencing batch.

762



763

764

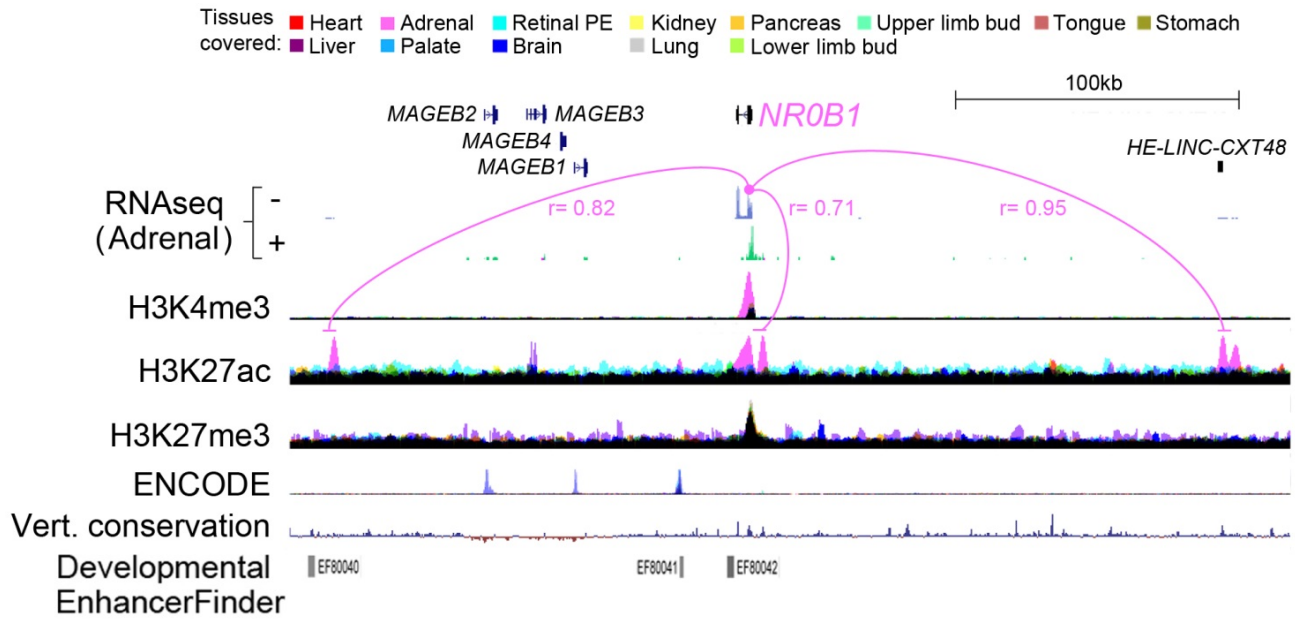
765 **Supplementary figure 8. Heatmap showing Phi correlation between samples when peaks are**
766 **called by MACs or allocated according to read count into 1 kb bins.**

767 a) Peak calling by MACs³⁷. b) Allocation according to read counts into 1 kb bins. The two
768 approaches produced similar heatmaps thus benchmarking the 1 kb bin approach. Each histone
769 modification is comprised of 26 individual rows and columns for the 12 tissue replicates plus single
770 datasets for tongue and stomach. The input is comprised of 5 control samples. The dark blue
771 diagonal line is the perfect correlation from assessing each sample against itself.

772

773

774



775

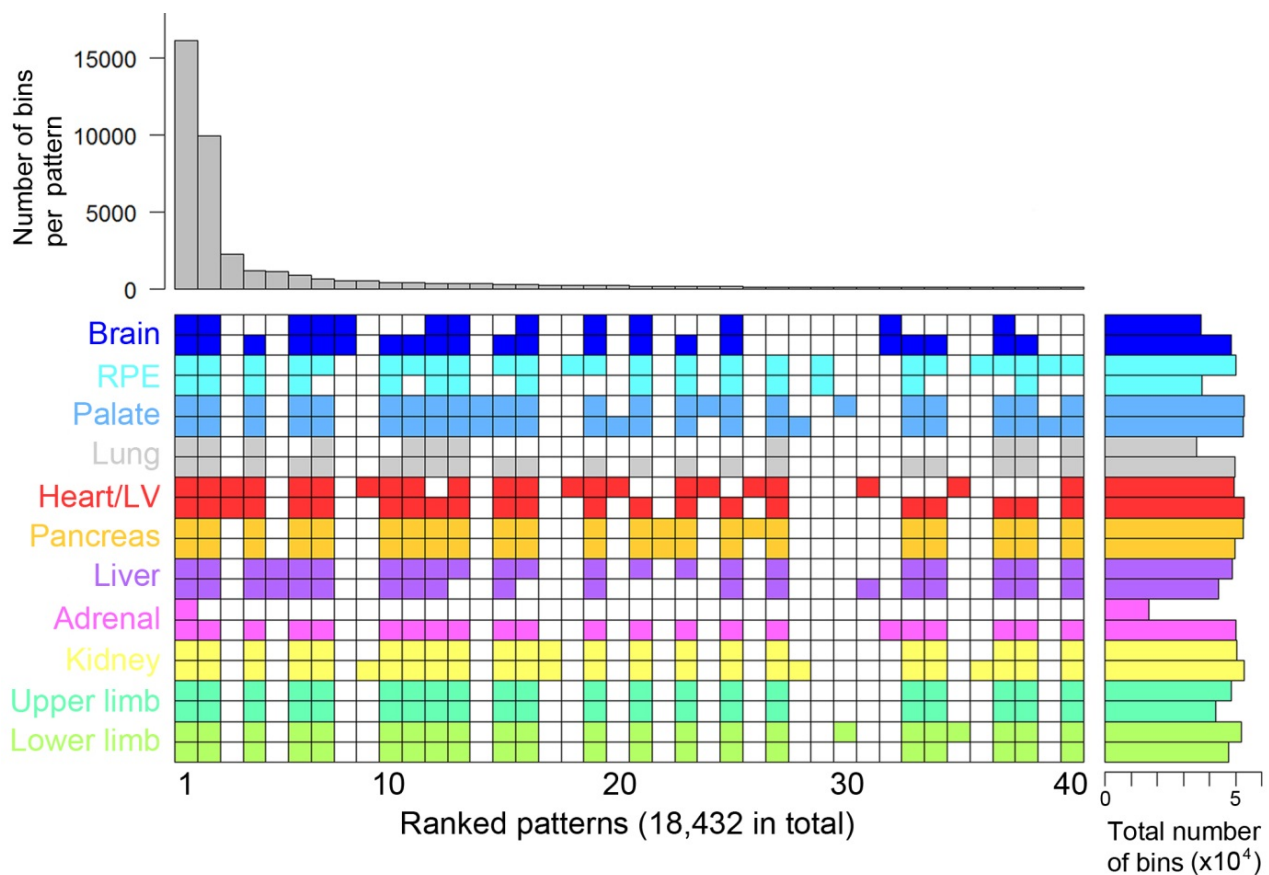
776

777 **Supplementary figure 9. Adrenal-specific epigenomic landscape over 300 kb at the *NR0B1***
778 **locus on the X chromosome.**

779 Assembled tracks show RNAseq for the adrenal and multi-layered data for all tissues for each
780 histone modification. One replicate of each track is shown for simplicity. *NR0B1* was only
781 expressed in the adrenal and has an adrenal-specific H3K4me3 mark. Multiple adrenal-specific
782 H3K27ac enhancer peaks were visible across 300 kb, all highly correlated with the expression of
783 *NR0B1*. Some of the enhancers are poorly conserved, unpredicted by in silico tools (Developmental
784 Enhancer Finder⁵³), and absent from ENCODE datasets²³. Identifying multiple enhancers facilitates
785 their grouping to assist with statistical power when assessing the potential pathogenicity of patient
786 variants in whole genome sequencing data.

787

788

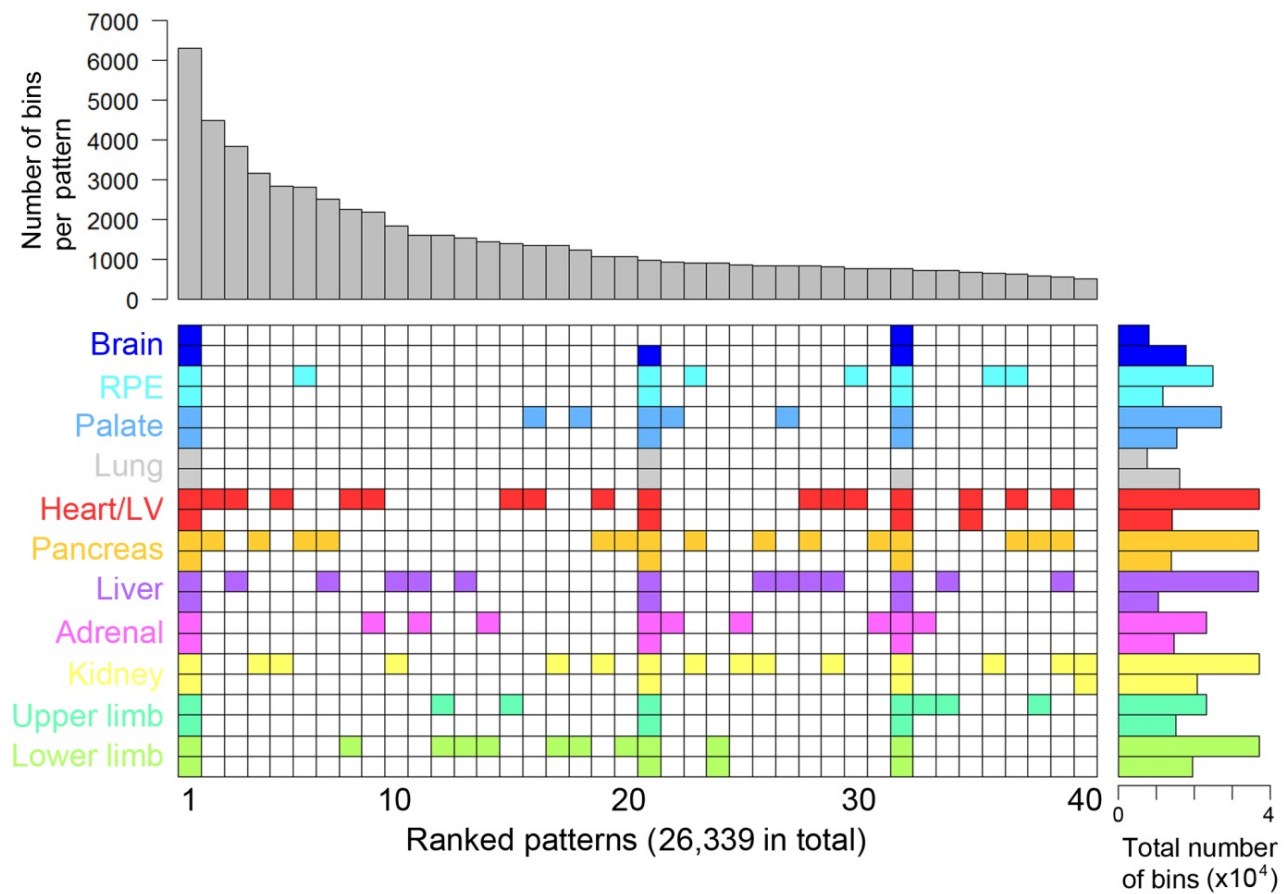


789

790 **Supplementary figure 10 (relates to Figure 5). Patterns of H3K4me3 across human**
 791 **embryonic tissues with the requirement of detection in at least 2 samples.**

792 Euler grid for bins marked by H3K4me3 (defined by elbow plots) in replicated tissues (i.e. two
 793 rows/replicates per tissue). Total number of marked bins per individual dataset is shown to the right.
 794 The grid shown required a bin to be called in any two or more samples and is ordered by decreasing
 795 bin count per pattern (bar chart above the grid). A total of 18,432 different patterns were identified
 796 (far fewer than the 48,570 found for the corresponding analysis of H3K27ac). The top 40 are shown.
 797 All tissue-specific patterns emerged in the top 2,267 (within the top 12.3% of patterns).

798

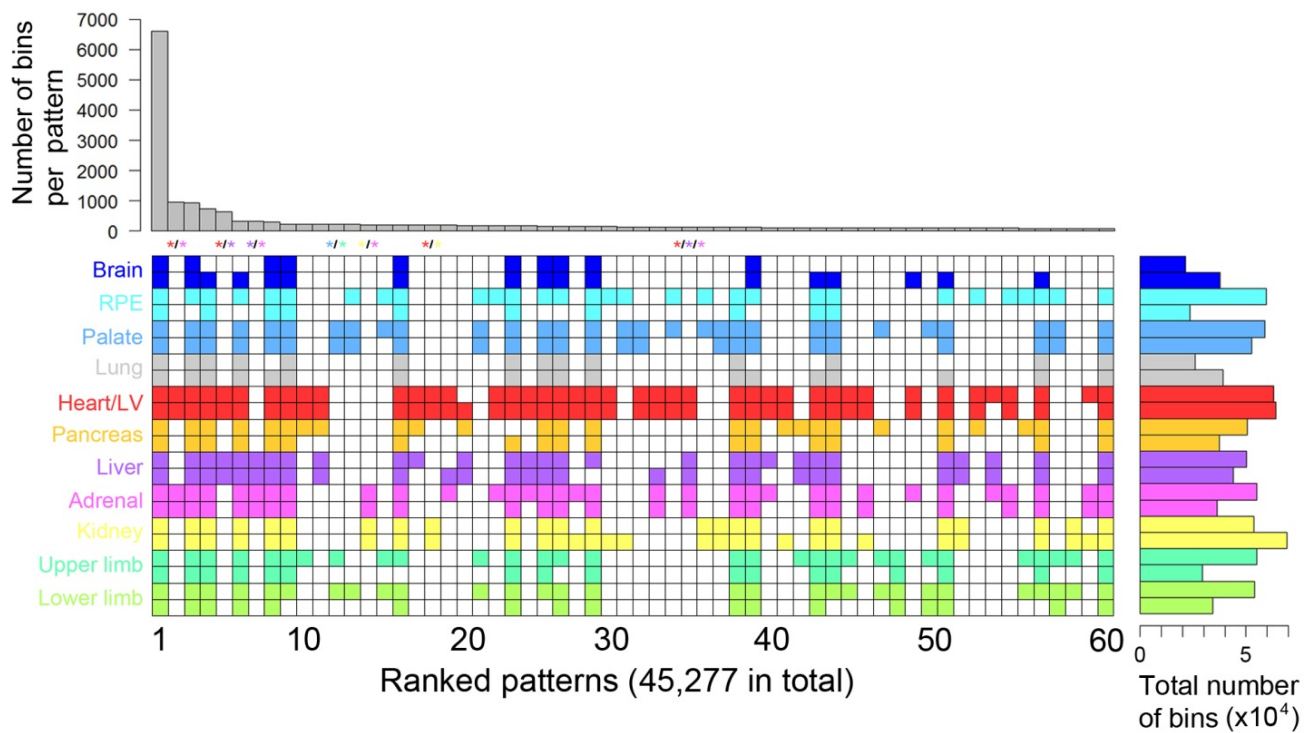


799

800

801 **Supplementary figure 11 (relates to Figure 5). Patterns of H3K27me3 across human**
 802 **embryonic tissues with the requirement of detection in at least 2 samples.**

803 Euler grid for bins marked by H3K27me3 (defined by elbow plots) in replicated tissues (i.e. two
 804 rows/replicates per tissue). Total number of marked bins per individual dataset is shown to the right.
 805 The grid shown required a bin to be called in any two or more samples and is ordered by decreasing
 806 bin count per pattern (bar chart above the grid). A total of 26,339 different patterns were identified
 807 (far fewer than the 48,570 found for the corresponding analysis of H3K27ac). The top 40 are shown.
 808 All tissue-specific patterns emerged in the top 836 (within the top 3.2% of patterns).



809

810

811 **Supplementary figure 12 (relates to Figure 5). Patterns of H3K27ac across tissues with the**
 812 **requirement of detection in at least 4 samples.**

813 Euler grid is shown for bins marked by H3K27ac (defined by elbow plots). Detecting patterns
 814 shared across tissues was enforced by detection in at least four samples (which must include at least
 815 two tissues). The grid includes replicated tissues (i.e. two rows/replicates per tissue). Total number
 816 of marked bins per individual dataset is shown to the right. The grid is ordered by decreasing bin
 817 count per pattern (bar chart above the grid). A total of 45,277 different patterns were identified. The
 818 top 60 are shown. Colour-coded asterisks above the columns indicate patterns shared uniquely
 819 across two (heart-adrenal, heart-liver, palate-limb, kidney-adrenal and heart-kidney) or three (heart-
 820 liver-adrenal) tissues.

821