Published online 7 March 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqac017

Generation and network analysis of an RNA-seq transcriptional atlas for the rat

Kim M. Summers[1,*,†], Stephen J. Bush[2,†], Chunlei Wu[3] and David A. Hume[1,*,†]

[1]Mater Research Institute—University of Queensland, Translational Research Institute, 37 Kent St, Woolloongabba, QLD 4102, Australia, [2]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, OX3 9DS, UK and [3]Department of Integrative and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Received November 14, 2021; Revised January 13, 2022; Editorial Decision February 14, 2022; Accepted February 15, 2022

ABSTRACT

The laboratory rat is an important model for biomedical research. To generate a comprehensive rat transcriptomic atlas, we curated and downloaded 7700 rat RNA-seq datasets from public repositories, downsampled them to a common depth and quantified expression. Data from 585 rat tissues and cells, averaged from each BioProject, can be visualized and queried at http://biogps.org/ratatlas. Gene co-expression network (GCN) analysis revealed clusters of transcripts that were tissue or cell type restricted and contained transcription factors implicated in lineage determination. Other clusters were enriched for transcripts associated with biological processes. Many of these clusters overlap with previous data from analysis of other species, while some (e.g. expressed specifically in immune cells, retina/pineal gland, pituitary and germ cells) are unique to these data. GCN analysis on large subsets of the data related specifically to liver, nervous system, kidney, musculoskeletal system and cardiovascular system enabled deconvolution of cell type-specific signatures. The approach is extensible and the dataset can be used as a point of reference from which to analyse the transcriptomes of cell types and tissues that have not yet been sampled. Sets of strictly co-expressed transcripts provide a resource for critical interpretation of single-cell RNA-seq data.

INTRODUCTION

In the year of the rat (2020), the Rat Genome Database (RGD) celebrated 20 years of development (1). Those 20 years saw completion of the draft genome (2). Around 90% of protein-coding genes had an inferred 1:1 ortholog in humans. Subsequent technology advances allowed the sequencing of multiple inbred strains, including several with disease-associated alleles (3). Szpirer (4) catalogued >350 rat genes where rat lines with natural or introduced variants provide models for human disease.

Analysis of transcriptional regulation in human and mouse has been driven by large consortium projects such as GTEx (5) and FANTOM (6), and there are many online resources for these species. Multi-tissue transcriptional atlas projects have also been published for other species, including chicken, sheep, buffalo, pig and goat (7–11). Although it was once suggested that guilt by association is the exception rather than the rule in gene regulatory networks (12), the principle is now very well established. Genes associated with specific organs, cell types, organelles and pathways (e.g. the cell cycle, protein synthesis, oxidative phosphorylation/mitochondria) are co-expressed along with transcription factors that regulate them (5,6,8,13–18). An extension of the principle of co-regulated expression is that it is possible to extract signatures of specific cell types, for example the stromal component of tumours (19) or resident tissue macrophages (20), based upon analysis of a large number of samples in which their relative abundance is variable.

The functional annotation of the rat genome is still a work in progress. Many rat genes in Ensembl are described as 'novel rat gene' and annotated solely by a gene number. Transcriptional regulation has evolved rapidly among mammalian species (21,22). Even where there is 1:1 orthology at the level of protein-coding sequence and conservation of synteny with other mammals, the expression may not be conserved. Two substantial studies have contributed to annotation of the rat transcriptome through RNA-seq analysis of a partly overlapping set of major rat organs (23,24). Long-read RNA sequencing has also contributed to refinement of rat transcriptome annotation (25). Because of the extensive use of the rat as a model in biomedical research, there are thousands of RNA-seq datasets in the public domain from isolated cells and tissues in various states of activation that could provide an additional resource for functional annotation. By combining random library downsiz-

*To whom correspondence should be addressed. Tel: +61 7 34437625; Fax: +61 7 34437779; Email: Kim.Summers@mater.uq.edu.au
Correspondence may also be addressed to David A. Hume. Tel: +61 7 34437315; Fax: +61 7 34437779; Email: David.Hume@uq.edu.au
†These authors contributed equally to the paper as first authors.

ing to reduce sampling bias and the high-speed 'pseudo-aligner' Kallisto (26) to quantify expression, we previously established a pipeline (7,11) to enable meta-analysis of published RNA-seq data. Here, we have used this pipeline to produce an extended expression atlas for the laboratory rat. To demonstrate the robustness of the integrated data, we have carried out network analysis to identify sets of co-expressed transcripts. The dataset is downloadable and the pipeline is extensible to allow inclusion of additional data and regeneration of the network as new RNA-seq data become available.

MATERIALS AND METHODS

Selecting samples for an expression atlas of the rat

To create a comprehensive expression atlas for the rat, we first downloaded the daily updated NCBI BioProject summary file from ftp://ftp.ncbi.nlm.nih.gov/bioproject/summary.txt (obtained 19 July 2021) and parsed it to obtain all BioProjects with taxonomy ID 10116 (*Rattus norvegicus*) and a data type of 'transcriptome or gene expression', supplementing this list by manually searching NCBI Geo and NCBI PubMed for the keywords 'RNA-seq AND rat'. BioProjects were selected to extend the diversity of tissues, cells and states from two existing rat transcriptomic atlases that analyse gene expression in a subset of major rat tissues (23,24). For each BioProject, we automatically extracted the associated metadata using pysradb v1.0.1 (27) with parameter '–detailed' or by manual review. Metadata for each BioProject, indicating (where available) the breed/strain, sex, age, tissue/cell type extracted and experimental condition (e.g. treatment or control), are detailed in Supplementary Table S1, which includes both the data downloaded via the pipeline and additional information retrieved manually from the European Nucleotide Archive record, NCBI BioProject record and cited publications. For incorporation into the expression atlas, we required that all samples have, at minimum, tissue/cell type recorded. Overall, the input to the atlas comprised 7682 samples from 363 BioProjects.

Quantifying gene expression for the atlas

For each library, expression was quantified using Kallisto v0.44.0 (26) as described in detail in previous studies on other species (7–9,20). Kallisto quantifies expression at the transcript level, as transcripts per million (TPM), by building an index of *k*-mers from a set of reference transcripts and then 'pseudo-aligning' reads to it, matching *k*-mers in the reads to *k*-mers in the index. Transcript-level TPM estimates were then summed to give gene-level TPM.

To create the reference transcriptomic index, we performed a non-redundant integration of the set of Ensembl v98 Rnor6.0 protein-coding cDNAs (http://ftp.ensembl.org/pub/release-98/fasta/rattus_norvegicus/cdna/Rattus_norvegicus.Rnor_6.0.cdna.all.fa.gz, accessed 24 November 2019; $n = 31\,715$ transcripts) and the set of 69 440 NCBI mRNA RefSeqs (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Rattus_norvegicus/all_assembly_versions/suppressed/GCF_000001895.5_Rnor_6.0/GCF_000001895.5_Rnor_6.0_rna.fna.gz, accessed 24 November 2019), as previously described (7). The purpose of the integration was to include transcripts that had not already been assigned Ensembl transcript IDs and whose sequence was not already present in the Ensembl release (under any identifier). RefSeq mRNAs incorporate untranslated regions (UTRs) and so could encapsulate an Ensembl CDS. The trimmed UTRs from each mRNA were generated excluding all sequence outside the longest open reading frame. In total, the reference transcriptome comprised 71 074 transcripts, representing 25 013 genes. Using this reference, expression was quantified for 7682 publicly archived paired-end Illumina RNA-seq libraries. The BioProjects are summarized in Supplementary Table S1. Prior to expression quantification, and for the purpose of minimizing variation between samples, we randomly downsampled all libraries to 10 million reads, five times each, using seqtk v1.2 (https://github.com/lh3/seqtk, downloaded 4 June 2018). Expression level was then taken to be the median TPM across the five downsampled replicates.

Within individual BioProjects, median TPM for replicate samples of the same tissue, age or condition was averaged. The final expression atlas is based on the averaged median downsampled TPM per gene for each distinct set of replicate samples. As in previous projects for other species (7–11), the full dataset of 585 averaged expression data from cells and tissues is displayed on BioGPS (28,29) at biogps.org/ratatlas to enable comparative analysis across species. The full processed primary dataset and the averaged data are available for download at an Institutional Repository (https://doi.org/10.5287/bodleian:Am9akye72). The latter is a comma-separated text file, which can be directly loaded into the network analysis software used herein or alternatives such as Gephi (https://gephi.org) or Cytoscape (https://cytoscape.org). This file can be easily supplemented by addition of further RNA-seq data processed in the same way. All scripts for generating the atlas are available at https://github.com/sjbush/expr_atlas.

Network analysis and functional clustering of atlas samples

To examine the expression of genes across this wide range of tissues and cell types, the expression data were analysed using the network analysis tool BioLayout [derived from BioLayout *Express*^{3D} (30)], downloaded from http://biolayout.org. The same files can be uploaded into the recently developed open source package, Graphia (https://graphia.app), which supports alternative clustering approaches and dynamic modification of parameters.

The initial analysis used the values averaged by age and BioProject for each tissue. Subsequent analyses used individual values for samples of liver, musculoskeletal system, cardiovascular system, kidney and central nervous system. For each analysis, a sample-to-sample correlation matrix was initially constructed at the Pearson correlation coefficient (r) threshold necessary to include all samples in the analysis (shown in the 'Results' section and figure legends). Pearson correlations were then calculated between all pairs of genes to produce a gene-to-gene correlation matrix of all genes correlated at $r \geq 0.75$.

Gene co-expression networks (GCNs) were generated from the matrices, where nodes represent either samples or genes and edges represent correlations between nodes

above the selected correlation threshold. For the sample-to-sample analyses (essentially analogous to a principal component analysis), an initial screen at the *r* value that entered all samples was performed, followed by subsequent analyses with a higher *r* value that removed outliers and revealed more substructure in the networks. For each gene-to-gene analysis, an *r* value threshold of 0.75 was used for all analyses (Supplementary Figure S1).

For the gene-to-gene networks, further analysis was performed to identify groups of highly connected genes within the overall topology of the network, using the Markov clustering algorithm (MCL) (31). The MCL is an algebraic bootstrapping process in which the number of clusters is not specified. A parameter called inflation effectively controls granularity. The choice of inflation value is empirical and is based in some measure on the predicted complexity of the dataset (31). The inflation value was 1.7 or 2.2 as indicated and only genes expressed at ≥10 TPM in at least one sample were included. Gene Ontology (GO) terms and Reactome pathways were derived from the Gene Ontology Resource (http://geneontology.org, release of 18 August 2021) using PANTHER overrepresentation test (PANTHER release of 24 February 2021). The reference list used was *R. norvegicus* (all genes in database), the Gene Ontology database was the release of 2 July 2021 (DOI: 10.5281/zenodo.5080993) and the Reactome pathway analysis used Reactome version 65, released 17 November 2020. These resources are all available at the Gene Ontology Resource (http://geneontology.org).

RESULTS

Samples in the atlas

Seven thousand six hundred eighty-two RNA-seq libraries, each with a unique SRA sample accession from 363 BioProjects, were obtained by the pipeline as described in the 'Materials and Methods' section and used to create a global atlas of gene expression. Metadata for the individual BioProjects are summarized in Supplementary Table S1. For comparative tissue analysis and the core atlas, expression across libraries was averaged by tissue, age and BioProject. This reduced the dataset to 585 different averaged samples of rat tissues and cells summarized in Supplementary Table S2A. For a separate analysis of liver, kidney, musculoskeletal, cardiovascular and central nervous systems to extract tissue-specific co-expression signatures, individual RNA-seq datasets from within each BioProject were used.

Network analysis of the rat transcriptome

Initially, we performed a sample-to-sample correlation to assess whether there were likely to be batch effects resulting in outlier samples that were unrelated to tissue type. To include all 585 samples, it was necessary to use a sample-to-sample $r \geq 0.21$. An image of the resulting network graph is shown in Figure 1. This visualization is analogous to a principal component analysis. Since BioProjects tended to focus on one strain, age, sex and tissue/treatment, some BioProject-specific clustering was expected. However, illustrating the robustness of the sampling and downsizing approach, the same or related tissues analysed in different BioProjects generally clustered together (compare Figure 1A where nodes are coloured by organ system and Figure 1B where they are coloured by BioProject). Note, for example, the tight clustering of liver sample (olive) generated by multiple independent laboratories. At a more stringent correlation coefficient threshold of 0.7, only 15 samples of relatively low connectivity were removed, but the association of nodes by organ system rather than BioProject becomes even more clear-cut (Figure 1C and D). No clear outliers or BioProject-specific clusters (batch effects) were identified, so all averaged samples were included in the subsequent gene-centred network analysis.

The threshold correlation coefficient for the gene-to-gene network was chosen empirically to maximize the number of nodes (genes included) while minimizing the number of edges (correlations between them) (Supplementary Figure S1). At the chosen correlation coefficient of $r \geq 0.75$, the graph contained 14 848 nodes (genes) connected by 1 152 325 edges. The full set of averaged expression profiles is provided as a web resource at http://biogps.org/ratatlas. On this site, a gene name query opens a display of the expression profiles across all 585 samples, links to rat genomic resources and a gene wiki connected to data related to the orthologous human gene. A click on the 'Correlation' button enables the user to define a correlation threshold and to generate a ranked list of correlated transcripts. This function can enable confirmation of relationships inferred from the clustering described below. It can also identify potential co-regulated transcripts for genes that were excluded from the network at the *r* value used for clustering.

Supplementary Table S2A shows all of the clusters detected for transcripts with a minimum expression of 10 TPM in at least one sample. In comparison to previous network analysis of mouse, human, pig, chicken, sheep and water buffalo transcriptomes (7–11) at this relatively stringent correlation coefficient, the much larger and more diverse rat transcriptomic dataset has a more fine-grained distribution with >1300 clusters having two nodes or more. In the published RNA-seq transcriptional atlas of 11 rat organs (32) that is included in the current data, around 40% of transcripts were expressed in all organs, in both sexes and at all development stages. In this larger set of averaged data, reflecting the much greater diversity of tissues and isolated cells sampled here, only 95 genes (0.38%) were detected above the 10 TPM minimal threshold in all 585 samples. These are shown in Supplementary Table S2B, with calculated maximum, minimum and variance. There is an obvious enrichment for mitochondrial and ribosomal subunit genes. There is still considerable variation among tissues and samples, but these genes have potential as controls for qRT-PCR. Commonly used controls such as *Actb*, *Gapdh*, *Tbp* and *Hprt* are also widely expressed but very low or absent in selected tissues that can be identified in the BioGPS site.

Significantly enriched GO terms (with associated corrected *P*-values) for clusters discussed later are included in Supplementary Table S2C. Consistent with previous analysis, there are clusters that show no evidence of tissue specificity but are clearly enriched for genes involved in defined biological functions. For example, clusters 11, 54 and 69 are associated with the cell cycle, DNA synthesis and repair.

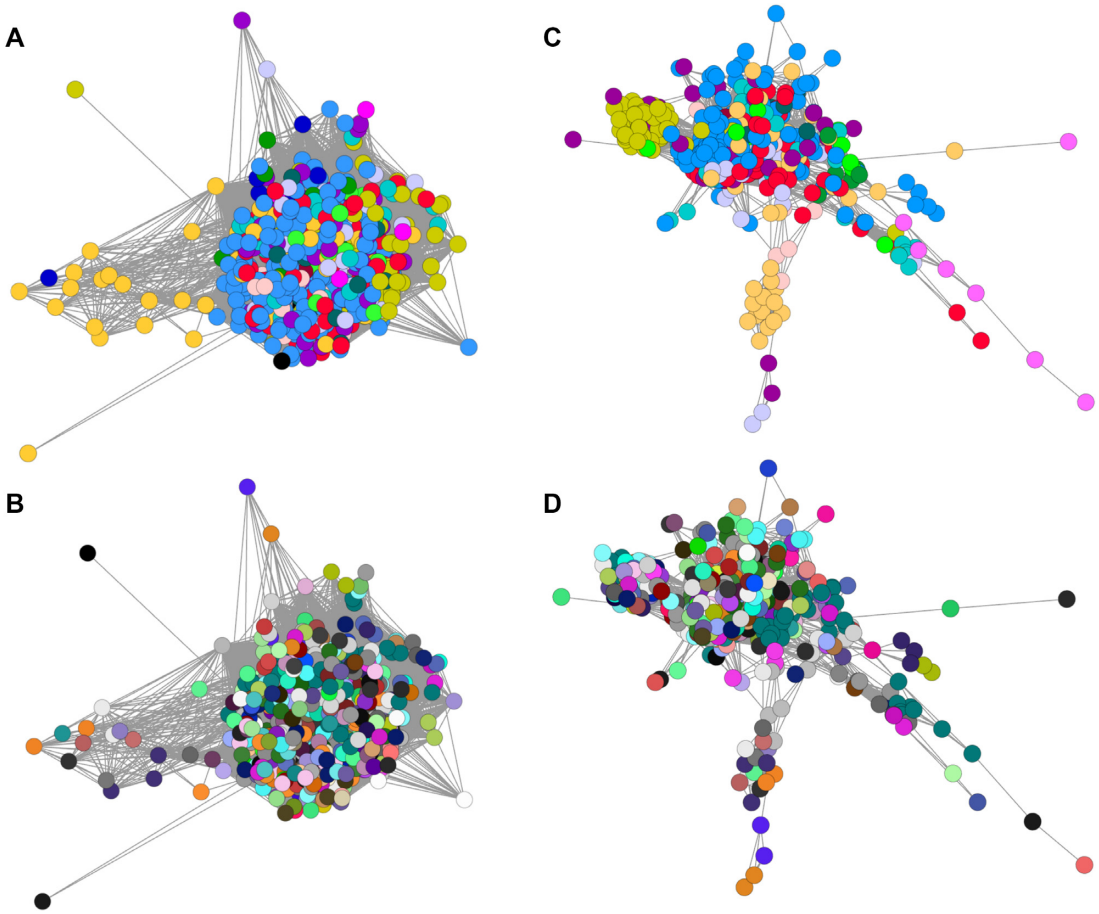

Figure 1. Sample-to-sample network graph for samples averaged by BioProject, age and tissue type. The 585 averaged RNA-seq profiles generated as described in the 'Materials and Methods' section. A pairwise sample-to-sample Pearson correlation coefficient (r) was calculated. The resulting matrix is displayed as a network graph using BioLayout. The individual samples (nodes, coloured balls) are connected by edges (lines) that reflect the chosen r value threshold. For panels (**A**) and (**B**), a correlation coefficient threshold of $r > 0.21$ was used to include all samples. For panels (**C**) and (**D**), the threshold was increased to a more stringent $r > 0.7$, which removed 15 nodes that make no connection at this r threshold and increased the separation of the remaining nodes. In panels (A) and (C), the nodes are coloured by organ system: dark red, auditory system; light red, cardiovascular system; salmon, digestive system; orange, endocrine system; olive, liver; bright green, female reproductive system; teal, immune system; dark teal, integumentary system; dark green, male reproductive system; black, mixed tissues; light blue, nervous system; dark blue, primordia/early development; purple, renal system; pink, respiratory system; mauve, skeletomuscular system; and grey, whole body (embryo). In panels (B) and (D), nodes are coloured by BioProject, data being generated by different laboratories. Note that in panels (A) and (C), where nodes are coloured by organ system, nodes of the same colour cluster together, whereas there is no pattern of association when the same nodes are coloured based on BioProject in panels (B) and (D).

Cluster 41 (see also Supplementary Table S2D) is made up almost entirely of histone-encoding transcripts, likely due to incomplete removal of non-polyadenylated transcripts in some of the RNA-seq libraries. This cluster is not specific to any BioProject. The 18 transcripts within this cluster identified by LOCID also have provisional annotation as histones. Although this cluster is the product of a technical error, it also highlights the power of the clustering approach to extract signatures of co-expression.

Table 1 summarizes the expression patterns and biological processes associated with clusters of transcripts showing evidence of tissue or cell type enrichment. The largest cluster of transcripts (cluster 1), >1500 in total, is expressed almost exclusively in the testis. A smaller cluster 70 is also testis specific. More than 500 of the transcripts in clusters 1 and 70 are identified only by a LOCID, RGD or other uninformative annotation and many more are identified only by structural motif (e.g. 50 members of the Ccdc family, 35 undefined Fams, 20 testis-expressed (Tex) and 15 Tmem protein genes). The complexity of the testis transcriptome in all mammalian species has been widely recognized [reviewed in (33)]. The set of testis-enriched transcripts with functional annotations encodes proteins associated with meiosis, sperm differentiation, structure and motility, and acrosomes. Unannotated genes are likely to be involved in male fertility. For example, mutation of *Dlec1*, a putative tumour suppressor gene, was recently shown to cause male infertility in mice (34). LOC498675 is a predicted 1:1 ortholog of mouse testis-specific gene 1700102P08Rik, which is expressed in spermatocytes and is essential for male fertility (35,36). Other smaller testis-enriched clusters include cluster 29, which contains Sertoli cell markers such as *Aard* and *Tsx* (37,38), cluster 72, which contains *Fshr* and the essential testis-specific transcription factor *Taf7l* (39,40), and cluster 88, which includes the male-determining transcription factor *Sry*.

Table 1. Gene expression clusters from rat tissues and cells

Cluster number	Number of transcripts	Specificity	Index genes and TFs	Functional annotation
1 and 70	1514 and 27	Testis	Acr, Amhr2, Ccna1, Fshr, Meioc, Spata16, Tnp1/2, Rec8, Stag3, **Nr6a1**, **Pbx4**, **Rfx2/8**, **Sox5**, **Sox30**, **Tcfl5**, **Taf7l**	Spermatogenesis, motility, meiosis
2	1303	CNS neurons	Amigo1, Camk2a, Cx3cl1, Gabbr1/2, Grik1–5, Nfasc, Snca, **Atf2**, **Bcl7a**, **Cbx6**, **Hdac11**, **Hivep2**, **Lmo3**, **Pou6f1**, **Rfx3**, **Tcf25**	Neurotransmission, neural development
3	583	Non-specific variable	Atm, Birc6, Ccnt1/2, Cdk12/13, Ddx5/6, Fancb, Herc1/2, Hipk1, **Arid2**, **Creb1**, **Kdm5a**, **Nf1**, **Nfe2l3**, **Nr2c2**, **Smad4/5**	Misfolded protein/stress response, tumour suppressors
5	342	Liver	Afm, Alb, Apoc1–4, C3, Cfb, Cth, Cyp2a1, F2, Fetub, Gcgr, Ghr, Hpx, Igf1, Plg, Serpina1, **Creb3l3**, **Foxa3**, **Meox2**, **Nr0b2**, **Nr1h3/i2/i3**, **Rxra**	Hepatocyte secretory products, xenobiotic metabolism
6	310	Oocyte	Axin2, Bmp15, Bub1b, Ccnb3, Dlgap5, Esrp1, Eya1/3, Gdf9, Gpr1, Zp1–4, **Cbx2**, **Dux4**, **Foxn4**, **Foxr1**, **Gata3**, **Lhx8**, **Nobox**, **Sall3**, **Taf4b**, **Taf5**, **Tead4**	Oocyte-specific transcription, zona pellucida structure, meiosis
7	213	Skeletal muscle	Acta1, Casq1, Ckm, Des, Mb, Myh2, Myl1, Pfkm, Ryr1, **Lbx1**, **Myf6**, **Pou6f2**, **Six1**, **Snai3**, **Zfp106**	Muscle contraction, calcium signalling
8	211	Kidney	Aco1, Adm2, Cyp4a2/a8, Klk1, Nox4, Pth1r, Slc5a2	Tubule function, resorption, metabolism
9	194	Oocyte	Aurkc, Ccnb1, Magoh, Mnd1, Mos, Nanos2, Ooep, **Brdt**, **Dazl**, **Gsc**, **Nr5a2**, **Pcgf1/6**, **Sall4**, **Sox15**, **Tcf15**, **Tcl1a**, **Zfp57**	Stem cell renewal, meiosis
11	188	Variable, not tissue specific	Bub1, Ccna2, Cdk1/2, Cenpk, Lig1, Mki67, Orc1, Pcna, Pola1, **E2f8**, **Foxm1**	Cell division cycle, DNA synthesis/repair, mitosis
12	165	ES cells	Dppa3/a4, Dusp10, Fgf17, Fzd6, Slc2a3, **Deaf1**, **Ferd3l**, **H2az1**, **Lefty1**, **Lmo2**, **Mybl2**, **Nanog**, **Nkx2–8**, **Tbx3**	Stem cell maintenance
14	124	Intestine	Ace2, Cdh17, Cldn7, Defa family, Dgat1, Heph, Il20ra, Krt20, Lgals4, Muc13, Vil1, **Hnf4g**	Intestinal barrier function
15	111	Stimulated T cells	Cd2, Cd3e, Cd69, Dock2, Il2rg, Ltb, Ptprc, Sla, Was, **E2f2**, **Ets1**, **Gfi1**, **Ikzf1/3**, **Limd2**	T-cell function
17	96	Pineal gland/retina	Aanat, Arr3, Asmt, Gch1, Opn1sw, **Bsx**, **Crx**, **Isl2**, **Lhx4**, **Mitf**, **Neurod4**, **Tafa3**	Pineal function, melatonin synthesis
18	95	Retina/pineal gland	Cnga1, Gabbr1/2, Opn1mw, Pde6a/b/g/h, Rd3, Rdh8, Rp1, Rtbdn, **Bhlhe23**, **Pax4**, **Prdm13**	Retinal function
19	94	Thymus	Ccl25, Cd3d, Cd8a/b, Fas, Rag1, Tap2, Tbata, **Foxn1**, **Ikzf2**, **Myb**, **Pax1**, **Rorc**, **Tcf7**, **Themis**	Thymic differentiation, selection
20	94	Liver, kidney	Cyp2c23, Dcxr, Fbp1, G6pc, Gk, H6pd, Pck1, Slc22a1, Slc37a4, **Hnf1a/4a**, **Nr1h4**	Gluconeogenesis
21	94	Macrophage, microglia	C1qa/b/c, Csf1r, Ctss, Gpr84, Hexb, Mpeg1, P2ry12/13, Siglec5, Tgfbr1, Trem2, Tyrobp, **Bhlhe41**, **Irf5**	Innate immune function, microglial differentiation
22	90	Skin	Cdsn, Csta, Klk9/10/12, Krt4/13/23, Lce3d/e, Lipk, Ppl, Trex2, Vsig8, **Barx2**	Skin barrier function
23	87	T cells, NK cells	Ccl1, Ccr4/5/8, Cd40lg, Gpr183, Ifng, Il17a, Il2, Il2ra/b, Lta, Zap70, **Batf**, **Icos**, **Runx3**, **Stat4**	Activation, cytokine secretion
24	85	Dorsal root ganglia	Acp3, Calca/b, Grik1, Htr1d, Nfeh/l/m, Nmb, Piezo2, Prokr1, Ret, **Drgx**, **Hoxd1**, **Pou4f1/f2**, **Smad9**, **Tlx3**	Ganglion cell differentiation
27, 28 and 33	75, 74 and 65	Skin	Adgrf4, Ces4a, Col17a1, keratins, Krtaps, Lce family, Lgals7, Lipm, Perp, **Tp63**, **Tprg1**	Skin barrier function
29	69	Testis	Aard, Clec12b, Gk5, Hormad1, Inca1, Shbg, Sycp1/2, **Msh4**, **Nkx3-1**, **Rhox8**, **Tbx22**, **Tsx**	Sertoli cell differentiation, synaptonemal complex
30	68	B cell	Btla, Cd19, Cd79a/b, Cxcr5, Fcna, Gpr174, Ighm, Jchain, **Ciita**, **Pax5**, **Pou2af1**, **Spib**, **Tlx1**	B-cell differentiation, immunoglobulin production
34	65	Prostate	Andpro, Cyss, Dach2, Eaf2, Fut4, Lao1, Lyc2, Mc5r, Pbsn, Sbp, Semg1, **Bhlha15**, **Creb3l4**, **Esr2**	Prostate differentiation, secretion
35	64	Adrenal	Cbr1, Cyp11a1/b2/b3, Cyp1b1, Fdx1, Kcnk3/9, Mc2r, Pcsk5, Pnmt, Soat1, Star, **Ar**, **Nr5a1**	Steroid hormone production, adrenalin
36 and 40	64 and 59	Placenta	Ceacam3/9/11/12, Cts7/8, Faslg, Fcrla/b, Ifnk, Il17f, Il23a, Lcn9, Mmp1, Peg10, Prl family, Wnt8a, **Elf5**, **Hand1**, **Rhox9**	Trophoblast differentiation, secretion
38	60	Brain	Crmp1, Ephb2, Gpc2, Gpr85, Marcksl1, Mdga1, Mex3b, **Dcx**, **Hmgb3**, **Lhx6**, **Mycl**, **Neurog2**, **Runx1t1**, **Sox11**	Neurogenic progenitor cell differentiation
42	56	Variable	Bub3, Ddx39a, Dkc1, Srsf2/3, Trip13, **Mycn**	Genotoxic damage response, tumour suppressors
43	52	Cochlea, middle ear	Cd164l2, Chrna9/10, Cldn9, Fbxo2, Grxcr1/2, Kncn, Loxhd1, Otoa/r/s	Hearing, cochlear function

Table 1. Continued

Cluster number	Number of transcripts	Specificity	Index genes and TFs	Functional annotation
44	51	Blood	Cxcr2, Gp9, Gypa, Kel, Pf4, S100a9, Tpt1, Tspo2	Platelets, granulocytes
46	49	Lung	Ager, Aqp5, Clec14a, Cyp2a3, Dram1, Fmo2, Lamp3, Lyz2, Scgb1a1/3a1/3a2, Sftpa1/b/c/d, Wnt3a, **Hopx**, **Nkx2-1**, **Smad6**, **Tbx4**	Alveolar type 1 and type II cell function and secretion
47 and 83	48 and 24	Heart	Actc1, Cav3, Fgf16, Myh7, Myl2, Palld, Ryr2, Tnnc1, **Ehd4**, **Irx4**, **Nkx2–5**, **Pdlim5**, **Tbx20**	Cardiac-specific muscle contraction.
48	48	Monocyte, macrophage	C5ar1, Ccr1, Cd14, Csf2ra, Cyba, Fcgr1a, Itgam, Msr1, Ncf1/2/4, Nlrp3, Slc11a1	Innate immune function, free radical production
49	46	Kidney	Acre2, Aqp2/3, Cldn8, Insrr, Kcne1, Oxgr1, **Foxi1**, **Hmx2**, **Hoxd3**	Distal tubule, collecting duct, water resorption
51	45	ES cells	Fgf4, Fgf19, Gdf3, **Nodal**, **Pou5f1**, **Prdm14**	Regulation of pluripotency
55	38	Granulocytes	Camp, Ctsg, Elane, Fncb, Mpo, Prg2/3, S100a8	Neutrophil granule proteins
63	33	Brain	Aqp4, Edil3, Gpr37/62 Mag, Mbp, Mobp, Opalin, Plp1, Sema4d, **Nkx6-2**	Myelination, oligodendrocytes
64	33	Pancreas	Amy2a3, Cel, Cela1/2a/3b, Cpa1/2, Ctrc/l Pnlip, Pnliprp1/2	Pancreatic enzymes, secretion
66	29	Stomach	Atp4a/b, Chia, Ctse, Cym, Ghrl, Gkn1/2, Pgc	Acidification, digestive enzymes
68	27	Brain, PC12 cells	P2rx2, Prph, Th, Vgf, **Gata2**, **Hand2**, **Phox2a**	Sympathetic neurons?
77	26	Mast cell?, lymphatic	Adgrg5, Cma1, Cpa3, Lilrb3a, Lyve1, Selp, Sirpd, Slpi, Timd4, **Cebpe**	
82	24	Adipose	Adipoq, Fabp4, Lep, Lipe, Lpl, Oxtr, Plin1, Pnpla2, Retn, Sucnr1, Tshr, **Pparg**	Fat storage, lipolysis, adipokines
87	21	Lens	Cryb family, Cryg family, Lim2, Opn4	Lens structural proteins
88	20	Macrophage	Adam8, Cd68, Ctsb, Ctsd, Gpnmb, P2rx4	Endosome/lysosome
90	20	Colon	Krt19, Lypd8, Phgr1, Pla2g10, Tspan1, **Cdx2**	Colon epithelium differentiation, secretion
92	19	Cerebellum	Ca8, Cbln1/3, Chn2, Fat2, Gabra6, Grm4, **En2**, **Hes3**	Purkinje cell differentiation, granule proteins
95	19	Variable in many tissues	Adgrl4, Cd93, Cdh5, Dll4, Egfl7, Kdr, Pcdh12, Pecam1, Tie1, **Erg**, **Myct1**	Endothelial cell differentiation
97	19	Cartilage growth plate	Acan, Clec11a, Col9a1/2/3, Loxl3, Rflna, **Alx1**, **Nkx3-2**	Cartilage structural proteins
98	18	Activated T cells, thymus	Ccr7, Cd7, Cd96, Heca, **Foxp3**	Immune cell activation
101	18	Macrophage	Acod1, Cxcl10, Il1a/b, Nos2	Response to LPS
106	16	Cartilage, tendon	Col2a1, Col10a1, Col11a1/2, Myh3, Ptx4, **Zfp648**, **Zim1**	Cartilage structural proteins

Clusters were generated at $r \geq 0.75$ and MCL inflation value 2.2. Selected transcripts encoding transcription factors are highlighted in bold. The full lists of transcripts in these clusters and the average expression profiles are provided in Supplementary Table S2. Index genes were chosen for illustrative purpose based upon known function in the indicated tissue confirmed by a PubMed search on gene name AND tissue. Where two cluster numbers are shown, the two clusters are in the same region of the network graph and show closely related expression profiles.

Clusters 17 and 18 contain transcripts expressed in both the retina and the pineal gland, both intimately involved in chronobiology and light sensing. Chang *et al.* (41) recently produced an aggregated resource describing the shared and divergent transcriptomes of these structures. Cluster 17 contains *Opn1sw*, the pineal-enriched transcription factor *Crx* and its target *Aanat* encoding the rate-limiting enzyme in melatonin synthesis (42). One unexpected inclusion in cluster 17, enriched in pineal, is the transcript encoding the transcription factor MITF. *MITF* in humans may be driven by as many as seven distinct promoters, including one used specifically by melanocytes. A unique transcription start site is shared by retinal pigment epithelial cells and pineal gland. *Mitf* overexpression in mouse pineal gland relative to other tissues has been noted previously (42,43) and in humans also *MITF* is most highly expressed in pineal gland (http://biogps.org). However, whereas targets of MITF have been identified in melanocytes and many other cell types (44) and mutations impact many complex phenotypes in mice and humans, there appears to be no literature on its role in the pineal gland. To illustrate the utility of the data, in Supplementary Table S2D we have reviewed the annotation of transcripts in clusters 17 and 18. Several novel transcripts of unknown function [e.g. *Katnip*, also annotated as *LOC361646*; *KIAA0586* (*Talpid3*), encoding a highly conserved ciliary protein associated with the human genetic disease Joubert syndrome (45); and *Lrtm1* (*LOC102547963*), a novel membrane protein] are also almost uniquely expressed in the human pineal gland (http://biogps.org).

Many smaller clusters detailed in Supplementary Table S2A are enriched in tissues, cell types or activation states that were not analysed in the existing rat atlases or indeed in any previous atlas project in other species. They can be annotated based upon known markers. For example, cluster 145 with 12 nodes contains transcripts encoding major secreted products of the pituitary (*Cga*, *Gh1*, *Fshb*, *Lhb*, *Tshb*) and the transcription factors that regulate their expression (*Pitx1*, *Six6*, *Tbx19*). Cluster 180 contains a subset of known immediate early genes (*Egr1*, *Fos*, *Jun*) mostly associated with isolated primary cells, and likely reflects cell

activation during isolation or tissue processing (20). Other known genes in the immediate early class cluster separately, or not at all, because they are constitutively expressed by specific cell types. Similarly, groups of inducible genes in innate immune cells are all expressed by LPS-stimulated macrophages but divide into at least three clusters (cluster 101, including *Il1a*; cluster 112, including *Ifit2* and other interferon targets; and cluster 126, including *Tnf*) because of expression by non-immune cells.

Other smaller clusters in Supplementary Table S2A group genes that share functions. The large protocadherin family of cell adhesion molecules is broadly divided into the clustered (α, β, γ) and non-clustered (δ) subgroups (46). The δ protocadherins are predominantly expressed in the nervous system and indeed *Pcdh1*, *Pcdh8*, *Pcdh9* and *Pcdh20* are brain restricted and part of the second largest cluster (cluster 2). However, cluster 81 includes *Pcdhb22* and 16 members of the *Pcdhg* (A and B) families, which are collectively enriched in the CNS but also widely expressed in other tissues. In addition, LOC108353166 within this cluster is annotated as protocadherin gamma-B2-like. Further members are more brain restricted and grouped together in cluster 250.

Nine of the 13 known mitochondrially encoded peptides group together in cluster 212, whereas clusters 61 and 76 group nuclear-encoded mitochondrial genes involved in the TCA cycle and oxidative phosphorylation (as expected, most highly expressed in heart and kidney). Cluster 102 groups 18 transcripts encoding proteins involved in mitochondrial β-oxidation of fatty acids. Several of the genes in this cluster are mutated in multiple acyl-CoA dehydrogenase deficiency (also known as glutaric aciduria type II) and related metabolic disorders (47). One additional gene involved in this pathway, *Etfb*, does not form part of a cluster. The web server (http://biogps.org/ratatlas) shows that *Etfb* is significantly correlated with many other genes associated with mitochondrial β-oxidation (e.g. with *Etfa* at $r = 0.59$ and with *Etfdh* at $r = 0.54$) but is expressed at lower levels in certain tissues, including the pineal gland.

Cluster 127, with 14 nodes, contains two markers of neurogenic cells [*Sstr2*, *Mpped1* (48,49)] and a candidate regulator, *Tiam2* (50), and is otherwise made up of 11 brain-specific transcriptional regulators, each of which has been shown to be essential for neurogenesis and likely interacts with the others. Clusters 125 and 332 contain 20 genes encoding proteins that have all been implicated as molecular chaperones, including multiple components of the TRIC chaperone complex (*Tcp1*, *Cct2*, *Cct3*, *Cct4*, *Cct5*). Cluster 557 with only four nodes contains the oligodendrocyte transcription factors, *Olig1* and *Olig2*, as well as *Sox 8*, which has non-redundant function in oligodendrocyte differentiation (51). The fourth node in this cluster, LOC103692025, is predicted by the RGD to be an ortholog of *Lhfpl3*, which in mouse is a marker of oligodendrocyte lineage commitment (52). The two calmodulin-encoding genes (*Calm1* and *Calm2*) are co-expressed (cluster 673) as are three genes involved in cholesterol synthesis (*Fdft1*, *Hmgcr*, *Hmgcs1*) (cluster 742). *Ins1* and *Ins2*, encoding insulin, are co-expressed with pancreatic polypeptide (*Ppy*) (cluster 751) but not with glucagon (*Gcg*). Although *Ppy* is normally expressed by rare gamma cells in pancreatic islets, a recent study indicated that gamma cells can produce insulin following beta-cell injury (53).

Each of the clusters contains genes that are identified only as LOCID or other numerical designation. These are obviously the subject of ongoing curation and in some cases LOCID transcripts duplicate named transcripts in the same cluster. In Supplementary Table S2, we have included an update on candidate annotations from the RGD and the http://biogps.org/ratatlas server provides a link to RGD besides the expression profile. Clearly, the co-expression information can provide additional assurance that putative orthology relationships with known mouse or human genes are likely to be correct.

Transcripts that do not form clusters

The first step in network analysis is the generation of a pairwise correlation matrix, and for any gene of interest one can immediately identify others with the most similar expression patterns. By lowering the inclusion threshold (*r* value), it is possible to include a larger proportion of transcripts, but the associations may become less informative biologically. For each gene of interest, the correlation function of the BioGPS site enables extraction of transcripts that are correlated at lower *r* values, which may provide some insight into function. Genes with unique expression profiles across the samples will not correlate with any other and therefore will not be included in the network graph. In many cases, the unique expression profile of a gene of interest arises because the gene product is 'multi-tasking' in different locations. Figure 2 shows the individual profiles of selected examples discussed later.

Mutations in *FBN1*, encoding the extracellular matrix protein fibrillin-1, are associated with Marfan syndrome that has complex impacts on musculoskeletal development, adiposity, vascular function and the eye. Distinct 3′ truncation mutations are associated with a neonatal progeroid lipodystrophy syndrome (54). Consistent with these phenotypes, *Fbn1* mRNA is highly expressed uniquely in the rat eye, aorta and cardiovascular tissues and cartilage/tendons and to a lesser extent in fibroblasts and adipose. There is also moderate expression in spinal cord and dorsal root ganglia, lung and testis. Dural ectasia, enlargement of the neural canal, is a common feature of Marfan syndrome (55). Expression in the lung may underlie the pulmonary emphysema observed in mouse models of fibrillinopathy (56); patients with Marfan syndrome frequently show apical blebs in the lung and are prone to pneumothorax (collapsed lung). Although *Fbn1* does not form part of a cluster at $r > 0.75$, the BioGPS correlation function reveals 52 genes correlated at $r > 0.6$, mostly associated with mesenchyme and extracellular matrix biology (e.g. *Adamts2*, *Bgn*, *Col5a1*, *Loxl1*, *Pdgfrb*, *Tgfb3*) (57).

The gene encoding dystrophin (*DMD*) associated in humans with mutations causing Duchenne muscular dystrophy is also not clustered. As expected, it is expressed in rat cardiac, skeletal and uterine muscle, but is also expressed in multiple brain regions at similar levels. This expression may be related to the neuropsychiatric impacts of the disease in both affected individuals and mouse models (58). In this case, FANTOM5 data indicate that *DMD* has at least

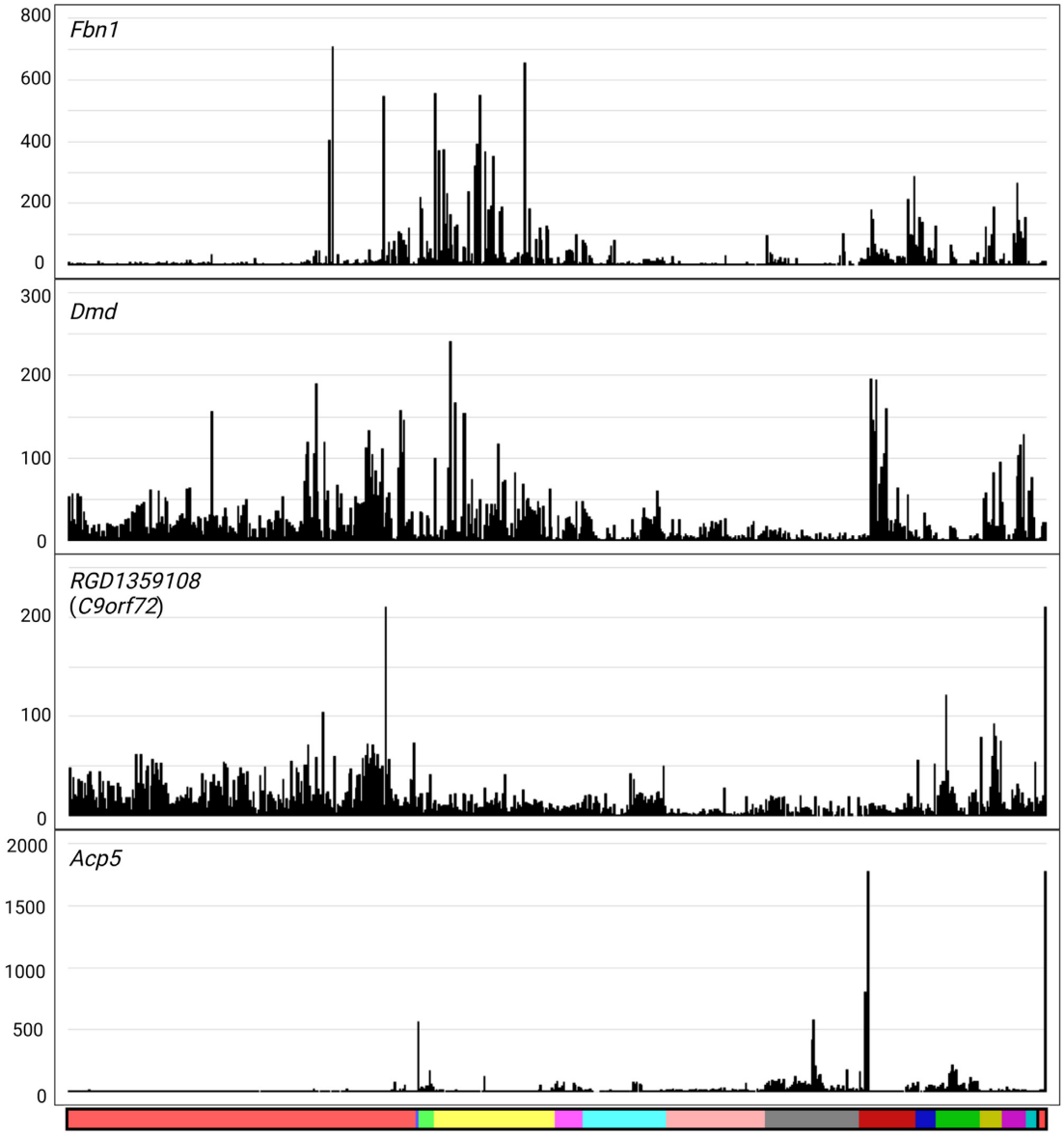

Figure 2. Gene expression profiles for genes that did not fall within a cluster. *Y*-axis shows the expression level in TPM. *X*-axis shows the organ system, coloured as in Supplementary Table S2. Reading from left to right: light red, nervous system; blue, auditory system; light green, respiratory system; yellow, cardiovascular system; pink, digestive system; turquoise, endocrine system; salmon, liver; grey, renal system; dark red, skeletomuscular system; dark blue, integumentary system; dark green, immune system; olive, male reproductive system; dark pink, female reproductive system; dark turquoise, primordia/early development; black, whole body (embryo); and red, mixed tissues.

two independent promoters (6). Nevertheless, several other genes associated with muscular dystrophy (*Sgcd*, *Lama2*, *Dst*) are correlated with *Dmd* at $r > 0.6$ (see BioGPS site).

RGD1359108 is a clear 1:1 ortholog of human *C9orf72*, associated with amyotrophic lateral sclerosis and frontal temporal dementia. O'Rourke et al. (59) reported that loss of function mutation in the orthologous gene in mice did not produce motor neuron dysfunction, but did lead to macrophage dysfunction, splenomegaly and lymphadenopathy. In rat, the ortholog of *C9orf72* is expressed widely in all CNS-associated tissues, most highly in spinal cord, but not enriched in any isolated CNS cell population. Outside the CNS, it is most highly expressed in stimulated macrophages and in testis.

A significant cohort of transcripts is excluded from co-expression clusters because they have alternative promoters, each with a distinct expression profile. One such gene is *Acp5*, encoding the widely used osteoclast (OCL) marker, tartrate-resistant acid phosphatase. *Acp5* forms part of a small cluster (cluster 179, 10 nodes) that is most highly expressed in the femoral diaphysis, and includes another OCL marker *Ctsk*, osteoblast-associated transcripts (*Bglap*, *Dmp1* and *Sp7*) and *Ifitm5*, mutated in a human bone-related genetic disease, osteogenesis imperfecta type V. It is surprising that so few transcripts are stringently associated with OCL; another small cluster (cluster 174, 11 nodes) that contains *Dcstamp*, *Ocstamp* (*Zfp334*) and *Mmp9* is enriched in the diaphysis sample but more widely expressed. Expres-

sion of *Acp5* in OCL in mice is initiated from an OCL-specific promoter (60). Aside from its function as a lysosomal enzyme in bone resorption, secreted ACP5 can function as a neutral ATPase and a growth factor for adipocytes (61,62). *Acp5* mRNA is expressed, albeit at lower levels than in bone, in rat adipose, lung (where it is expressed highly by alveolar macrophages), small and large intestine, kidney and spleen as well as isolated macrophages.

The transcriptome of the rat liver

The downloaded datasets included around 1900 individual RNA-seq libraries of liver, including whole liver from various ages, sexes, inbred and outbred rat strains, disease models, liver slice cultures and isolated cells. In principle, clustering of such diverse data could identify sets of co-expressed transcripts that are associated with cell types, locations or disease processes that are hidden in the averaged data of the complete sample set. To test that view, we clustered the entire liver-related dataset without averaging the replicates. As in the main atlas, the correlation threshold was chosen empirically at 0.75. The cluster list and the average profile of transcripts in each cluster are provided in Supplementary Table S3A and informative clusters are summarized in Table 2.

It is immediately evident that not all of the samples are pure liver. Liver cluster 31 contains a set of pancreas-specific genes, including *Cpa1* that overlaps with cluster 64 in the main atlas. This cluster arises because of random contamination with pancreatic tissue of liver samples in the large BodyMap project (32). Liver cluster 73 contains transcripts encoding all of the major secretory products of pancreatic islets (e.g. *Ins1* and *Gcg*). This cluster was detected only in liver from a study of enforced activity and sleep deprivation (63). It is not clear from the paper how these samples could have been selectively contaminated with islet mRNA unless they are mislabelled. Liver cluster 5 is detected in a rather random subset of samples from multiple BioProjects likely also indicating contamination. It includes the progenitor marker, *Lgr5*, but also various adhesion molecules (*Cldn10/18*) and neuroendocrine markers (*Chga/b*). There is little evidence of expression of these genes in normal liver in other species, and at least some of the genes (e.g. *Cckar* and *Cldn10/18*) are highly expressed in pancreas and/or stomach (e.g. see http://biogps.org). Liver cluster 21 is detected in a single sample, and contains smooth muscle-associated transcripts (*Actg2*, *Tpm2*).

The disadvantage of analysing a single tissue is that most transcripts do not vary greatly between datasets. In one sense, this provides a quality control for the efficacy of the random sampling approach we have used. In this dataset, the largest cluster by far (liver cluster 1) is relatively consistent with the exception of increased detection in all samples from a BioProject that profiled liver slices from a bile duct ligation model, cultured for 48 h *in vitro* and treated with various agents (64). It is not clear why this gene set would be expanded in that cellular system. Liver cluster 1 includes many transcripts expressed constitutively by hepatocytes. The most abundant hepatocyte-specific transcript encoding albumin (*Alb*) is not strictly correlated with any other transcript presumably reflecting its specific regulation (65).

Table 2. Gene expression clusters from rat liver

Liver cluster number	Number of nodes	Description and index genes
1	6292	Widely expressed, high in bile duct ligation model; growth, protein synthesis, inflammation, fibrosis, connective tissue
2	752	High in foetal liver; cell cycle, haematopoiesis, embryonic liver; *cyclins, Cdk1, Pcna, Igf2, Hbb, S100a8/9,* ***E2f2***, ***Klf1***, ***Myb***
3	414	General expression, metabolic regulation; *Bcl2l2, Cdk5, Cirbp,* ***Esrra***, ***Foxk1***, ***Hdac6***, ***Nfe2l1***, ***Nr1h2***, ***Nr2c1***, ***Pias3***, ***Rara***, ***Six5***, ***Tfe3***, ***Tfeb***
4	278	General expression, control of lipid metabolism; *Arid1a, Bcl9, Camta2, Crtc1/2, Fastk,* ***Foxj2***, ***Foxp4***, ***Hsf1***, ***Mef2d***, ***Rela***, ***Rfx1***, ***Rxrb***, ***Tp53***
5	206	Isolated samples, gall bladder, neuroendocrine; *Cckar, Chga/b, Cldn10/18, Inha, Krtap1–3, Lgr5, Scg3/5, Nmb, Nts*
6	166	E14 liver, fibrosis model; *Acta2, Cdh11, Epha4/7, Fbn2, Gpc2, Myh6/7, Sfrp1/2,* ***Alx***, ***Cited1***, ***Foxf1***, ***Gata5***, ***Shox2***, ***Tbx15/18***, ***Tgif2***, ***Twist1/2***, ***Wt1***
7	148	Foetal liver, fibrosis, Zucker rats: myeloid infiltration; *Axl, Cd4, Cd68, Clec4a1, Fcgr1a, Hk3, Lyz2, Ptprc,* ***Irf5***, ***Fli1***, ***Spi1***
10	98	Variable expression: proteasome complex, proteolysis; *Anxa7, Ctsd/l, Fbxo22, Prdx1/6, Psma, Psmb2, Psmc1, Psmd1, Tmx2, Usp5,* ***Creb3***
11	76	Variable, low in foetal liver, periportal hepatocytes, urea synthesis; *Agmat, Ass1, Ces1a, Cyp2e1, Gls2, Gcgr, Gpt, Hsd17b11, Pink1, Slc25a22;* ***Mlxipl***, ***Nr1i2***
13	67	Variable, low in foetal liver, fibrosis model, fatty acid β-oxidation; ***Acat1***, ***Acot1***, ***Crat***, ***Cyp4a1***, ***Etfdh***, ***Hadh***, ***Pank1***, ***Pdk4***, ***Slc22a5***, ***Vnn1***
16 and 70	105 and 10	Variable, cholesterol and fatty acid syntheses; *Aacs, Acaca, Acly, Dhcr7, Fads1/2, Fasn, Hmgcr, Hmgcs1, Lss, Mvd,* ***Nfe2***, ***Srebf1/2***
18	54	Fibrosis; *Angptl4, Col1a1/2, Col6a1/6, Gpc1, Lgals1, Loxl1, Lum, S100a4, Sfpr4,* ***Etv1***, ***Osr2***
24	41	Variable, mast cells; *Cpa3*, *Cpz*, *Mcpt2*, *Prss8*
25	41	Variable, interferon response; *Dhx58, Gbp1/4, Ifi44, Ifit1, Isg15, Mx1/2, Oas1/2,* ***Irf7***
26	41	Variable, mitochondrial; ***Atp5me/f/g***, ***Cox7ab***, ***Ndufa2/4/5/6***
31	34	One BioProject, pancreas contamination; ***Cela1***, ***Cpa1***, ***Klk1***, ***Pnlip***, ***Prrs1***
33	32	One BioProject, NK cells; ***Cd96***, ***Gzma***, ***Klra1***, ***Ly49***, ***Prf1***
34	31	Highly variable, hepatic stellate cell activation? *Acvr1c, Apob, Egfr, Fcgr2b, Klb, Mrc1, Stab2,* ***Klf12***, ***Nr3c2***
43	21	Variable, interferon response; ***Adar***, ***Ifih1***, ***Parp9/10/12/14***, ***Irf9***
56	12	Kupffer cell; ***Cd5l***, ***Csf1r***, ***Sdc3***, ***Siglec1***, ***Vsig4***
63	10	Endothelial cell; *Cd93, Cdh5, Flt1, Nrp1, Pecam1, Tgfbr3, Tie1,* ***Ets1***, ***Tbx20***
65	10	Class II MHC; *Aif1, Batf2, Cd74, Rt1-Ba/b, RT1-Da/b,* ***Irf8***, ***Ciita***
66	10	Male-specific; *Akr1c12, Cyp2a2, Hsd3b5, Sult1c3*
69	10	Xenobiotic-induced; *Ces2a, Gstm2, Ugt1a5*
84	9	Female-specific; *Akr1b7, Cyp2c12, Srd5a1, Sult2a1/6,* ***Cux2***, ***Trim24***

Clusters were generated at $r \geq 0.75$ and MCL inflation value 1.7. The full gene lists for each of the clusters are provided in Supplementary Table S3A. Transcription factors are highlighted in bold. Index genes were chosen for illustrative purpose based upon known function.

Liver cluster 1 also contains transcripts encoding markers of hepatic stellate cells (e.g. *Pdgfra/b*) and the corresponding growth factors (*Pdgfa/b/d*) as well as more general mesenchyme markers (e.g. *Vim*) and markers of cholangiocytes (e.g. *Krt7*) suggesting that their relative abundance is not highly variable among the samples.

The remaining liver clusters analyse differential development and activation states that distinguish the samples and BioProjects. These clusters are informative and consistent with prior knowledge. Liver cluster 2 is expressed specifically in embryonic liver and is a complex mix of transcripts reflecting both differentiation of hepatocytes and the function of the liver as a haematopoietic organ. Accordingly, it contains the cell cycle genes, the foetal growth factor *Igf2*, and markers of erythroid (e.g. *Hbb*) and myeloid (*S100a8/a9*) haematopoietic lineages. Liver clusters 3 and 4 are both expressed in almost all liver samples and the level of expression is not highly variable. Expression of each of the smaller clusters is much more variable between samples and BioProjects and known genes within those clusters indicate an association with specific cell types and processes as summarized in Table 2 and discussed later.

One signature that was not detected is that of the specialized centrilobular population that is adapted to clear ammonia generated by the urea cycle. In mice, the rate-limiting enzyme, glutamate ammonia lyase (also known as glutamine synthetase, *Glul* gene), is expressed exclusively in a band of cells surrounding the central vein. Liver-specific deletion of *Glul* leads to pathological hyperammonaemia (66). In mice, this population of cells co-expressed *Rhgb* (encoding an ammonia transporter) and ornithine aminotransferase (*Oat*) and was enriched for a number of *Cyp* genes (e.g. *Cyp2e1* and *Cyp1a2*). However, in the diverse rat liver dataset, there was only marginal correlation with other centrilobular-enriched transcripts.

The transcriptome of central nervous, renal, musculoskeletal and cardiovascular systems

Each of these systems also contributes hundreds of RNA-seq datasets including isolated cells and specific regions or structures. To further examine the utility of these large datasets for the analysis of cell type- and process-specific signatures, the data from each of these biological systems were clustered separately in Supplementary Table S4 (nervous), Supplementary Table S5 (renal), Supplementary Table S6 (cardiovascular) and Supplementary Table S7 (musculoskeletal). The clusters are annotated in the tables and to avoid confusion with multiple cluster numbers, each system is discussed separately in Supplementary Text. Broadly speaking, as in the liver, network analysis of individual organ systems enables a more fine-grained extraction of cell type-, region- and process-specific expression signatures.

The transcriptome of rat macrophages

The transcriptome of rat macrophages has been analysed previously based upon microarrays (67) and the RNA-seq data included here (68). Macrophages adapt to perform specific functions in specific tissues (20). Cluster 21 (Table 1 and Supplementary Table S2), which includes *Csf1r*, is most highly expressed in brain and brain-derived cells and includes transcripts that are enriched in microglia compared to macrophages from other tissues (e.g. *P2ry12*). Around two-thirds of these transcripts are contained within a set of 119 transcripts depleted in all brain regions of *Csf1r*-knockout rats (69). Cluster 47 (Supplementary Table S2) contains transcripts that may be shared with microglia (e.g. *Itgam*, encoding CD11b) but are common to monocytes and many tissue macrophage populations. Cell surface markers of other macrophage populations cluster idiosyncratically as shown in Supplementary Table S2, indirectly supporting tissue macrophage heterogeneity; *Clec4f*, the Kupffer cell marker, is within the liver cluster, *Vsig4* and *Marco* (cluster 1239), *Clec10a* , *Mrc1* (CD206) and *Stab1* (cluster 168), *Lyve1* and *Timd4* (cluster 79), and *Adgre1* and *Clec4a1/3* (cluster 286) are correlated with each other, while others (e.g. *Cd163*, *Tnfrsf11a*, *Siglec1*) do not cluster at all at this threshold because each has a unique pattern of expression in tissue macrophages. Figure 3 shows the profiles of *Csf1r*, *Adgre1*, *Cd163*, *Vsig4* and *Mrc1* in the averaged data.

The network analysis of such a diverse set of cells and tissues also dissociates known macrophage transcriptional regulators (e.g. *Spi1*, *Spic*, *Nr1h3*, *Mafb*, *Irf8*, *Cebpa/b*, *Tfec*) (20) from macrophage expression clusters because none of these regulators is entirely macrophage restricted. For example, transcription factor SPIC in mice is required for splenic red pulp macrophage and splenic iron homeostasis (70). In the rat, *Spic* mRNA is most highly expressed in spleen as expected, but also detected in ES cells and germ cells (see profile on http://biogps.org/ratatlas). Macrophage differentiation and adaptation likely involve combinatorial interactions among multiple transcription factors as exemplified by the complex regulation of the transcription of the *Csf1r* gene (71).

Whereas macrophages express a diversity of endocytic receptors, there is not a corresponding large cluster of transcripts encoding endosome–lysosome components including the vacuolar ATPase (ATP6v) subunits and lysosomal hydrolases. Transcripts encoding endosome-associated CD68 and GPNMB proteins are co-expressed with *Ctsb* and *Ctsd*. Although CD68 is often used as a macrophage marker, it is clearly not macrophage restricted. Most transcripts encoding lysosomal acid hydrolases (e.g. *Acp1*, *Lipa*) are widely expressed and each varies independently.

Csf1r is strongly correlated with other macrophage-specific markers in cluster 21, consistent with strong evidence that expression is entirely restricted to the macrophage lineage in rats as it is in mice (72). It is also detected at relatively high levels in all tissues (around 5–10% of the level in isolated macrophages) consistent with the abundance of tissue macrophages detectable with a *Csf1r* reporter transgene (72) and with a study of tissue development in mice (73). However, expression was also detected in many isolated primary cell samples that are not meant to contain macrophages. For example, BioProjects PRJNA556360 and PRJNA552875 contain RNA-seq data derived from oligodendrocyte progenitors purified using the A2B5 marker, but this population has *Csf1r* expression at similar levels to purified macrophages. Another BioProject, PRJNA355082, describes expression profiling of isolated

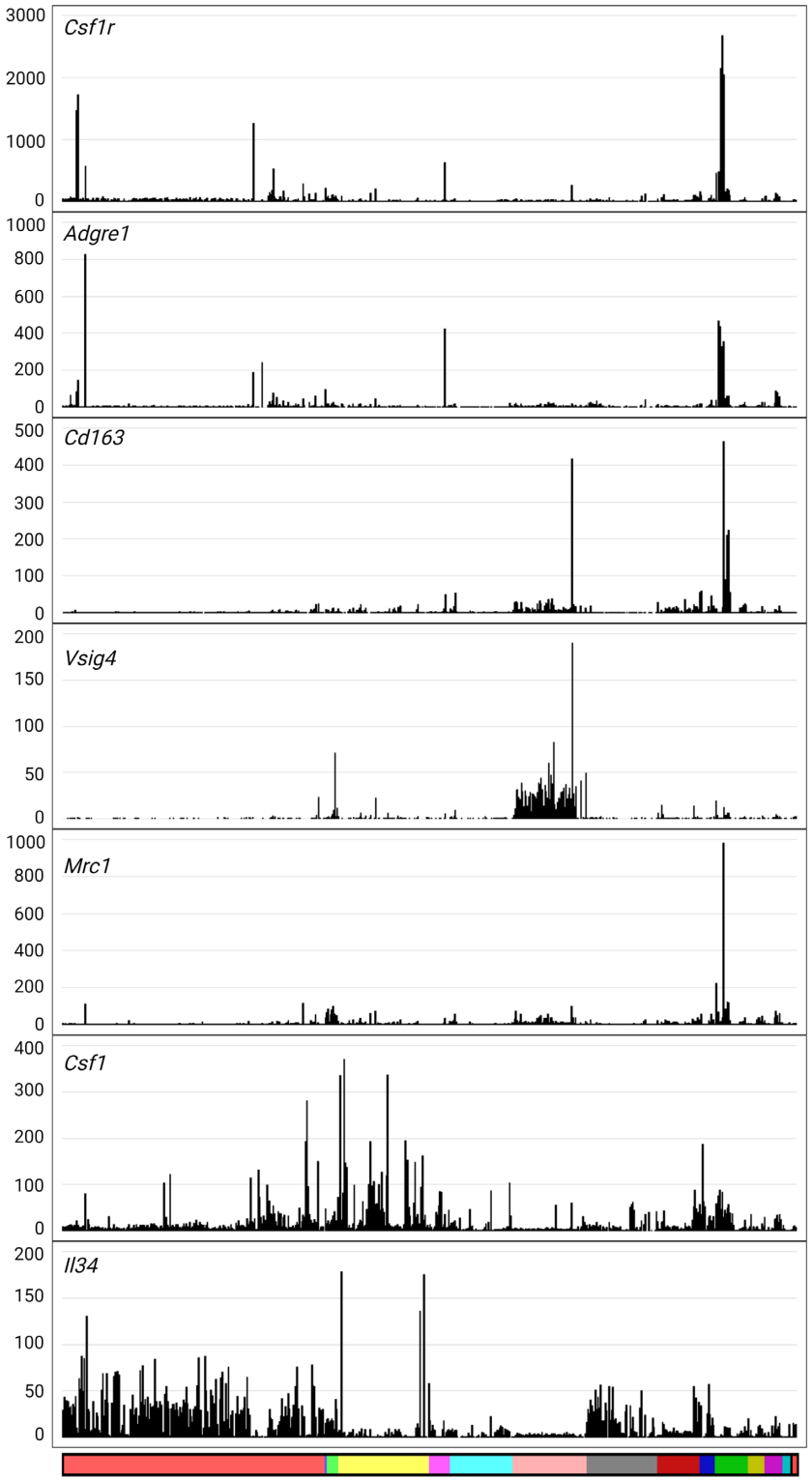

Figure 3. Gene expression profiles for macrophage-related genes. *Y*-axis shows the expression level in TPM. *X*-axis shows the organ system, coloured as in Supplementary Table S2. Reading from left to right: light red, nervous system; blue, auditory system; light green, respiratory system; yellow, cardiovascular system; pink, digestive system; turquoise, endocrine system; salmon, liver; grey, renal system; dark red, skeletomuscular system; dark blue, integumentary system; dark green, immune system; olive, male reproductive system; dark pink, female reproductive system; dark turquoise, primordia/early development; black, whole body (embryo); and red, mixed tissues.

astrocytes, but this dataset also has a similar level of *Csf1r* mRNA to pure macrophages. Other datasets from various ganglion cell populations, neuronal progenitor cells, cardiac fibroblasts and cardiomyocytes and hepatic stellate cells are clearly highly enriched in *Csf1r* and other macrophage-associated transcripts.

CSF1R has two ligands, CSF1 and IL34. In mice and rats, mutation of the *Csf1* gene leads to a global reduction in many tissue macrophage populations, whereas mutation of *Il34* in mice leads to selective reduction of microglia and Langerhans cells. Based upon the difference in phenotype between *Csf1* and *Csf1r* mutations in rats, we speculated that *Il34* could be more widely expressed and functional in rat macrophage homeostasis compared to mouse (68). Neither growth factor forms part of a cluster. Figure 3 also shows the profiles of *Csf1* and *Il34*. As expected, *Csf1* mRNA is widely expressed and enriched in isolated mesenchymal cells. *Il34* is expressed in all brain regions and isolated cells at similar levels and also in skin. However, in contrast to mouse, *Il34* is expressed at similar levels in many other tissues, notably aorta, adipose, kidney, lung and testis.

The tissue-specific analysis in Supplementary Tables S4–S7 enables the extraction of macrophage-specific signatures from resident populations that have not been isolated and characterized previously. For example, in the cardiovascular analysis, a cluster of 184 transcripts containing *Csf1r* as well as a smaller cluster containing *Adgre1* extracts a signature of cardiac resident macrophages distinct from blood leukocytes that form a separate cluster (see Supplementary Text).

DISCUSSION

Overview

The extraction and normalization of published RNA-seq data have enabled the generation of a comprehensive rat expression atlas that samples transcriptional diversity on a comparable scale to the FANTOM5 data for human and mouse (6) and massively extends the BodyMap generated from 11 rat tissues (32). The user-friendly display at http://biogps.org/ratatlas enables a gene-specific query to visualize the expression of any gene of interest across the full dataset and use of the correlation function allows the identification of transcripts with similar expression profiles. BioGPS also hosts large expression datasets for mouse, human, sheep and pig for comparative analysis. The validity of the downsampling normalization and the utility and information content of the atlas have been exemplified by gene-centred network analysis of the averaged core dataset. The primary data are available for download by users in a form that enables local regeneration of the networks and addition of user-generated datasets. In comparison to rat, there are orders of magnitude more total RNA-seq datasets from mouse and human cells and tissues in public repositories. We previously identified and analysed 470 RNA-seq datasets from mouse resident tissue macrophages alone, excluding data from cells stimulated *in vitro* or in disease models (20). The approach we have used is extensible to even larger datasets in mouse and human.

Analysis of liver-specific transcriptional network

The assembled dataset includes multiple BioProjects and thousands of RNA-seq datasets related to the liver, central nervous system, heart and cardiovascular system, and kidney. Each has been analysed independently to identify signatures of individual cell types and processes (Supplementary Tables S3–S7). To illustrate the ability of network analysis to extract biologically informative expression signatures, we analysed the liver data in greater detail and considered other tissue-specific analysis in Supplementary Text.

Liver gene expression is regulated in response to numerous physiological stimuli and chronic disease processes, including fatty liver disease. Aside from hepatic parenchymal cells, the liver contains several non-parenchymal populations. To identify co-regulated clusters within the liver transcriptome, we analysed the liver samples separately using the same GCN approach used for the overall atlas. The liver is the major source of plasma protein and performs many functions in energy homeostasis, lipid and protein synthesis, and biotransformation of xenobiotics and endogenous by-products. The function of the liver depends on its structure, which comprises small units called lobules, each composed of concentric layers of hepatocytes expanding from the central vein towards the periportal vein. The metabolic function of hepatocytes varies along the periportal–central axis, a phenomenon referred to as metabolic zonation (74). In principle, if there was significant heterogeneity in metabolic state or development among the liver samples, a gene-to-gene clustering might reveal sets of genes associated with portal versus centrilobular regions of liver lobules. Halpern *et al.* (75) performed single-cell RNA-seq (scRNA-seq) analysis of mouse hepatocyte diversity and concluded that zonation impacts as many as 50% of transcripts. However, this analysis was limited to 8-week-old fasted male C57BL/6 mice and does not necessarily capture coordinated regulation of the metabolic domains, including diurnal oscillations and response to feeding (76). Broadly speaking, the single-cell analysis indicated a periportal bias for major secretory products of hepatocytes and a pericentral concentration of expression of genes involved in xenobiotic metabolism.

Network analysis shown in Supplementary Table S3A and summarized in Table 2 revealed a large co-regulated cluster (liver cluster 11) that includes *Gls2*, an archetypal periportal marker in mice, other enzymes and transporters associated with the urea cycle (*Ass1*, *Acy3*, *Agmat*, *Cbs*, *Gpt*, *Slc25a22*, *Nags*) and the glucagon receptor, *Gcgr*. Cheng *et al.* showed that glucagon is a regulator of zonation in mouse liver, in that glucagon deficiency led to reduced expression of periportal-enriched transcripts (77). There are candidate transcriptional regulators within this cluster with known functions in hepatic transcriptional regulation: the xenobiotic sensor *Nr1i2* and the glucose-sensing transcription factor *Mlxipl* (78,79). A smaller liver cluster 88 contains additional key enzymes of urea synthesis, *Arg1*, *Cps1* and *Gpt2*, as well as the amino acid transporter, *Slc38a4*.

The analysis of the liver samples does not reveal a corresponding pericentral expression cluster. *Glul*, which appears strictly restricted to a single layer of cells surrounding

the central vein in mice, rats and humans (74), showed limited heterogeneity among the liver datasets and did not form part of this cluster. This suggests that *Glul* is not highly regulated, whereas other centrilobular-enriched transcripts alter their expression in response to external stimulus. Another putative landmark pericentral gene, *Cyp2e1*, is actually part of liver cluster 11, redistributed in at least some of the experimental models sampled herein, as observed in a model of paracetamol exposure that forms part of this dataset. Other transcripts that are biased to centrilobular also form separate clusters because of their independent regulation in response to stimulation. For example, *Cyp1a2* was identified as a pericentral marker (74). Liver cluster 54 (Supplementary Table S3A) is elevated in a dataset from a BioProject studying the effects of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, a potent aryl hydrocarbon receptor (AhR). It includes the detoxifying enzymes *Cyp1a1*, *Cyp1a2* and *Cyp1b1*, the AHR repressor gene (*Ahrr*) and transcription factor *Cdx2*, a known AHR target gene (80). A distinct set of xenobiotic metabolizing genes, *Ces2a*, *Gstm2* and *Ugt1a5*, is coregulated in liver cluster 69, and *Ephx1*, *Gsta2*, *Gsta4*, *Gsta5* and *Gstm1* are co-regulated in liver cluster 146. The proteasome subunit *Psmd4* was also pericentral in mice (75), but it is found in liver cluster 10 stringently co-regulated as one might expect with numerous other components of the proteasome complex. Liver cluster 10 contains the transcription factor *Creb3* and likely reflects the activation of the Golgi stress response in a subset of samples or BioProjects (81).

The regulation of lipid metabolism is of particular interest given the current epidemic of non-alcoholic fatty liver disease. There is some evidence of zonation of fatty acid metabolism in the liver, fatty acid β-oxidation being enriched in periportal hepatocytes and lipogenesis in pericentral hepatocytes (75), but these pathways are independently regulated in this dataset. Liver cluster 13 is highly enriched for genes involved in lipolysis and fatty acid β-oxidation. It overlaps the smaller cluster in the full atlas (cluster 101) but includes many additional genes that have tissue-specific enrichment (e.g. *Acot7* in CNS). Conversely, liver clusters 16 and 70 comprise enzymes of cholesterol and fatty acid syntheses and the known transcriptional regulators, *Nfe2* and *Srebf1/2*. Liver cluster 26 contains multiple genes involved more generally in mitochondrial oxidative phosphorylation, including multiple genes encoding NADH-ubiquinone oxidoreductase subunits. We are not aware of any heterogeneity in mitochondrial distribution in the liver.

The various metabolic and inflammatory disease models, with distinct effects on non-parenchymal cells, enable deconvolution of signatures of specific cell types and disease processes. Liver cluster 6, which includes the classical fibrosis marker, *Acta2* (smooth muscle alpha actin), is elevated in fibrosis models, but highest in E14 liver, which may indicate that myofibroblast activation in fibrosis recapitulates the phenotype of embryonic mesenchyme. Liver cluster 18 captures transcripts associated with more advanced fibrotic disease and includes multiple collagen genes and two candidate transcriptional regulators, *Etv1* and *Osr2*. This cluster also contains the mesenchymal gene *Olfml3*, which is also expressed in microglia in the mouse (see http://biogps.org) and human (82) but is not associated with microglia in the rat (69). This highlights the problems with assuming that genes have similar expression patterns and functions across species.

The fibrosis-associated clusters are clearly separated from liver cluster 7 that captures the phenotype of infiltrating $CD45^+$ (*Ptprc*) myeloid cells in various models. Supplementary Table S3B summarizes distinct GO term enrichment for liver clusters 7 and 18. Two sets of interferon-responsive transcripts including key regulators *Irf7* and *Irf9* cluster separately (liver clusters 25 and 43) as do transcripts associated with expression of class II MHC (liver cluster 65). These clusters are separated also from the signatures of endothelial cells (liver cluster 63) and of Kupffer cells, the resident macrophages (liver cluster 56) (Supplementary Table S3A). The latter cluster includes the transcript encoding the macrophage growth factor receptor, *Csf1r*, and many transcripts that were also downregulated in livers of *Csf1r*-knockout rats (83). *Clec4f*, which is expressed exclusively by Kupffer cells in mice, and is in the liver-specific cluster in the extended atlas, is in a separate cluster (liver cluster 95) with the three C1q subunits (*C1qa/b/c*), *Cfp*, *Ctss*, *Pld4* and *Tifab*. There is emerging interest in the last gene, a forkhead-associated domain protein, in immune cell function and inflammation (84).

Finally, in rodents, there is a set of transcripts that is expressed in the liver in a sex-specific manner in part under the influence of growth hormone (85,86). The male- and female-specific liver transcriptomes are regulated by differential expression of specific transcription factors, CUX2 and ONECUT2 in females and BCL6 in males. The majority of samples are from males, but nevertheless liver cluster 66 is excluded from female livers, and liver cluster 84 contains *Cux2*, *Trim 24* and known female-specific transcripts.

The relationship between network analysis and scRNA-seq for the definition of cell types in tissues

As in the liver, the network analysis of other major organ systems enabled robust extraction of clusters of co-regulated transcripts often including the transcription factors that regulate them. In this case, the issue of tissue-specific promoters becomes less of an issue and genes that have multiple promoters (e.g. *Mitf*, *Acp5*) may form part of tissue-specific networks highlighting local functions. The deconvolution of large datasets by network analysis complements scRNA-seq analysis that has rapidly become a dominant approach to analysis of cellular heterogeneity. scRNA-seq is not quantitative. Typically, expression of <1000 genes is detected in each cell and even the most highly expressed transcripts are not detected in every cell (87). The output of scRNA-seq conflates two distinct types of zero values: those where a gene is expressed but not detected by the sequencing technology (stochastic sampling) and those that reflect genuine expression heterogeneity. Whereas we can readily separate entirely unrelated cells that share few markers in scRNA-seq, such as epithelia and haematopoietic cells, the identification of numerous subpopulations within individual lineages is tenuous at best (20). A second disadvantage of analysis of isolated cells by scRNA-seq or total RNA-seq is that cells are inevitably ac-

tivated during isolation and single cells can have attached remnants of other cells that contribute RNA (20,88).

Suo *et al.* (89) described computational analysis of mouse cell atlas to identify 202 regulons whose activities are highly variable across different cell types and predicted a small set of essential regulators for each major cell type in mouse. We have achieved the same outcome for the rat without the use of scRNA-seq. The advantage of network deconvolution as performed here is that one can explore a much wider diversity of states than can be contemplated with scRNA-seq and identify more robust co-regulatory modules. Any proposed pair of markers of a specific cell population defined by scRNA-seq should be strongly correlated with each other if both are detectable in whole tissue. The prediction was tested in a meta-analysis of mouse tissue macrophage populations that failed to support the existence of a specialized macrophage subset defined from scRNA-seq data by reciprocal expression of *Lyve1* and *Mrc1* (20). Herein, the detailed analysis of the liver data indicates that zonation of the liver is dynamic and individual pathways are regulated to a large extent independently of each other. So, the definition of subpopulations of hepatocytes is state dependent. The discussion of other systems in Supplementary Text casts doubt on the fine-grained definition of subsets of tissue-specific parenchymal/epithelial cells and more generic glial cells, fibroblasts, endothelial cells, parenchymal cells and macrophages in many published scRNA-seq analyses. Network analysis reveals regulons that may, or may not, be restricted to a defined cell population, but which are clearly linked to function. In that respect, one might reasonably question the value of defining cell types as an approach to understanding biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author contributions: S.J.B. developed the informatics pipeline and generated the primary expression data. K.M.S. performed the network and enrichment analyses and manual annotation of metadata. C.W. developed BioGPS and established the BioGPS viewer of the atlas. D.A.H. wrote the initial manuscript, reviewed all genes and clusters, and contributed to informatic analysis. S.J.B. and K.M.S. contributed to manuscript editing.

FUNDING

Mater Foundation, Brisbane [to D.A.H. and K.M.S.]; Translational Research Institute [to D.A.H. and K.M.S.].
Conflict of interest statement. None declared.

REFERENCES

1. Smith,J.R., Hayman,G.T., Wang,S.J., Laulederkind,S.J.F., Hoffman,M.J., Kaldunski,M.L., Tutaj,M., Thota,J., Nalabolu,H.S., Ellanki,S.L.R. *et al.* (2020) The year of the rat: the Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Res.*, **48**, D731–D742.
2. Gibbs,R.A., Weinstock,G.M., Metzker,M.L., Muzny,D.M., Sodergren,E.J., Scherer,S., Scott,G., Steffen,D., Worley,K.C., Burch,P.E. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
3. Atanur,S.S., Diaz,A.G., Maratou,K., Sarkis,A., Rotival,M., Game,L., Tschannen,M.R., Kaisaki,P.J., Otto,G.W., Ma,M.C. *et al.* (2013) Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell*, **154**, 691–703.
4. Szpirer,C. (2020) Rat models of human diseases and related phenotypes: a systematic inventory of the causative genes. *J. Biomed. Sci.*, **27**, 84.
5. The GTEx Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
6. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest,A.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J., Haberle,V., Lassmann,T. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
7. Bush,S.J., Freem,L., MacCallum,A.J., O'Dell,J., Wu,C., Afrasiabi,C., Psifidi,A., Stevens,M.P., Smith,J., Summers,K.M. *et al.* (2018) Combination of novel and public RNA-seq datasets to generate an mRNA expression atlas for the domestic chicken. *BMC Genomics*, **19**, 594.
8. Clark,E.L., Bush,S.J., McCulloch,M.E.B., Farquhar,I.L., Young,R., Lefevre,L., Pridans,C., Tsang,H.G., Wu,C., Afrasiabi,C. *et al.* (2017) A high resolution atlas of gene expression in the domestic sheep (*Ovis aries*). *PLoS Genet.*, **13**, e1006997.
9. Young,R., Lefevre,L., Bush,S.J., Joshi,A., Singh,S.H., Jadhav,S.K., Dhanikachalam,V., Lisowski,Z.M., Iamartino,D., Summers,K.M. *et al.* (2019) A gene expression atlas of the domestic water buffalo (*Bubalus bubalis*). *Front. Genet.*, **10**, 668.
10. Muriuki,C., Bush,S.J., Salavati,M., McCulloch,M.E.B., Lisowski,Z.M., Agaba,M., Djikeng,A., Hume,D.A. and Clark,E.L. (2019) A mini-atlas of gene expression for the domestic goat (*Capra hircus*). *Front. Genet.*, **10**, 1080.
11. Summers,K.M., Bush,S.J., Wu,C., Su,A.I., Muriuki,C., Clark,E.L., Finlayson,H.A., Eory,L., Waddell,L.A., Talbot,R. *et al.* (2019) Functional annotation of the transcriptome of the pig, *Sus scrofa*, based upon network analysis of an RNAseq transcriptional atlas. *Front. Genet.*, **10**, 1355.
12. Gillis,J. and Pavlidis,P. (2012) "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput. Biol.*, **8**, e1002444.
13. Ballouz,S., Weber,M., Pavlidis,P. and Gillis,J. (2017) EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*, **33**, 612–614.
14. Freeman,T.C., Ivens,A., Baillie,J.K., Beraldi,D., Barnett,M.W., Dorward,D., Downing,A., Fairbairn,L., Kapetanovic,R., Raza,S. *et al.* (2012) A gene expression atlas of the domestic pig. *BMC Biol.*, **10**, 90.
15. Giotti,B., Chen,S.H., Barnett,M.W., Regan,T., Ly,T., Wiemann,S., Hume,D.A. and Freeman,T.C. (2018) Assembly of a parts list of the human mitotic cell cycle machinery. *J. Mol. Cell Biol.*, **11**, 703–718.
16. Hume,D.A., Summers,K.M., Raza,S., Baillie,J.K. and Freeman,T.C. (2010) Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations. *Genomics*, **95**, 328–338.
17. Mabbott,N.A., Baillie,J.K., Brown,H., Freeman,T.C. and Hume,D.A. (2013) An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*, **14**, 632.
18. Singh,A.J., Ramsey,S.A., Filtz,T.M. and Kioussi,C. (2018) Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.*, **75**, 1013–1025.
19. Doig,T.N., Hume,D.A., Theocharidis,T., Goodlad,J.R., Gregory,C.D. and Freeman,T.C. (2013) Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment. *BMC Genomics*, **14**, 469.
20. Summers,K.M., Bush,S.J. and Hume,D.A. (2020) Network analysis of transcriptomic diversity amongst resident tissue macrophages and dendritic cells in the mouse mononuclear phagocyte system. *PLoS Biol.*, **18**, e3000859.
21. Jubb,A.W., Young,R.S., Hume,D.A. and Bickmore,W.A. (2016) Enhancer turnover is associated with a divergent transcriptional

response to glucocorticoid in mouse and human macrophages. *J. Immunol.*, **196**, 813–822.
22. Villar,D., Berthelot,C., Aldridge,S., Rayner,T.F., Lukk,M., Pignatelli,M., Park,T.J., Deaville,R., Erichsen,J.T., Jasinska,A.J. *et al.* (2015) Enhancer evolution across 20 mammalian species. *Cell*, **160**, 554–566.
23. Ji,X., Li,P., Fuscoe,J.C., Chen,G., Xiao,W., Shi,L., Ning,B., Liu,Z., Hong,H., Wu,J. *et al.* (2020) A comprehensive rat transcriptome built from large scale RNA-seq-based annotation. *Nucleic Acids Res.*, **48**, 8320–8331.
24. Sollner,J.F., Leparc,G., Hildebrandt,T., Klein,H., Thomas,L., Stupka,E. and Simon,E. (2017) An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Sci. Data*, **4**, 170185.
25. Wang,X., You,X., Langer,J.D., Hou,J., Rupprecht,F., Vlatkovic,I., Quedenau,C., Tushev,G., Epstein,I., Schaefke,B. *et al.* (2019) Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. *Nat. Commun.*, **10**, 5009.
26. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
27. Choudhary,S. (2019) pysradb: a Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. *F1000Research*, **8**, 532.
28. Wu,C., Jin,X., Tsueng,G., Afrasiabi,C. and Su,A.I. (2016) BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic Acids Res.*, **44**, D313–D316.
29. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
30. Theocharidis,A., van Dongen,S., Enright,A.J. and Freeman,T.C. (2009) Network visualization and analysis of gene expression data using BioLayout Express^{3D}. *Nat. Protoc.*, **4**, 1535–1550.
31. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res.***30**, 1575–1584
32. Yu,Y., Fuscoe,J.C., Zhao,C., Guo,C., Jia,M., Qing,T., Bannon,D.I., Lancashire,L., Bao,W., Du,T. *et al.* (2014) A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat. Commun.*, **5**, 3230.
33. Grimes,S.R. (2004) Testis-specific transcriptional control. *Gene*, **343**, 11–22.
34. Okitsu,Y., Nagano,M., Yamagata,T., Ito,C., Toshimori,K., Dohra,H., Fujii,W. and Yogo,K. (2020) Dlec1 is required for spermatogenesis and male fertility in mice. *Sci. Rep.*, **10**, 18883.
35. Li,M., Zheng,J., Li,G., Lin,Z., Li,D., Liu,D., Feng,H., Cao,D., Ng,E.H.Y., Li,R.H.W. *et al.* (2021) The male germline-specific protein MAPS is indispensable for pachynema progression and fertility. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2025421118.
36. Wu,X.L., Yun,D.M., Gao,S., Liang,A.J., Duan,Z.Z., Wang,H.S., Wang,G.S. and Sun,F. (2020) The testis-specific gene 1700102P08Rik is essential for male fertility. *Mol. Reprod. Dev.*, **87**, 231–240.
37. Cunningham,D.B., Segretain,D., Arnaud,D., Rogner,U.C. and Avner,P. (1998) The mouse Tsx gene is expressed in Sertoli cells of the adult testis and transiently in premeiotic germ cells during puberty. *Dev. Biol.*, **204**, 345–360.
38. Svingen,T., Beverdam,A., Verma,P., Wilhelm,D. and Koopman,P. (2007) Aard is specifically up-regulated in Sertoli cells during mouse testis differentiation. *Int. J. Dev. Biol.*, **51**, 255–258.
39. Akinloye,O., Gromoll,J., Callies,C., Nieschlag,E. and Simoni,M. (2007) Mutation analysis of the X-chromosome linked, testis-specific TAF7L gene in spermatogenic failure. *Andrologia*, **39**, 190–195.
40. Cheng,Y., Buffone,M.G., Kouadio,M., Goodheart,M., Page,D.C., Gerton,G.L., Davidson,I. and Wang,P.J. (2007) Abnormal sperm in mice lacking the Taf7l gene. *Mol. Cell. Biol.*, **27**, 2582–2589.
41. Chang,E., Fu,C., Coon,S.L., Alon,S., Bozinoski,M., Breymaier,M., Bustos,D.M., Clokie,S.J., Gothilf,Y., Esnault,C. *et al.* (2020) Resource: a multi-species multi-timepoint transcriptome database and webpage for the pineal gland and retina. *J. Pineal Res.*, **69**, e12673.
42. Rohde,K., Rovsing,L., Ho,A.K., Moller,M. and Rath,M.F. (2014) Circadian dynamics of the cone-rod homeobox (CRX) transcription factor in the rat pineal gland and its role in regulation of arylalkylamine *N*-acetyltransferase (AANAT). *Endocrinology*, **155**, 2966–2975.
43. Bailey,M.J., Coon,S.L., Carter,D.A., Humphries,A., Kim,J.S., Shi,Q., Gaildrat,P., Morin,F., Ganguly,S., Hogenesch,J.B. *et al.* (2009) Night/day changes in pineal expression of >600 genes: central role of adrenergic/cAMP signaling. *J. Biol. Chem.*, **284**, 7606–7622.
44. Goding,C.R. and Arnheiter,H. (2019) MITF—the first 25 years. *Genes Dev.*, **33**, 983–1007.
45. Fraser,A.M. and Davey,M.G. (2019) TALPID3 in Joubert syndrome and related ciliopathy disorders. *Curr. Opin. Genet. Dev.*, **56**, 41–48.
46. Morishita,H. and Yagi,T. (2007) Protocadherin family: diversity, structure, and function. *Curr. Opin. Cell Biol.*, **19**, 584–592.
47. Missaglia,S., Tavian,D. and Angelini,C. (2021) ETF dehydrogenase advances in molecular genetics and impact on treatment. *Crit. Rev. Biochem. Mol. Biol.*, **56**, 360–372.
48. Chen,C.M., Wang,H.Y., You,L.R., Shang,R.L. and Liu,F.C. (2010) Expression analysis of an evolutionarily conserved metallophosphodiesterase gene, Mpped1, in the normal and beta-catenin-deficient malformed dorsal telencephalon. *Dev. Dyn.*, **239**, 1797–1806.
49. Maubert,E., Slama,A., Ciofi,P., Viollet,C., Tramu,G., Dupouy,J.P. and Epelbaum,J. (1994) Developmental patterns of somatostatin-receptors and somatostatin-immunoreactivity during early neurogenesis in the rat. *Neuroscience*, **62**, 317–325.
50. Chu,C.H., Chen,J.S., Chuang,P.C., Su,C.H., Chan,Y.L., Yang,Y.J., Chiang,Y.T., Su,Y.Y., Gean,P.W. and Sun,H.S. (2020) TIAM2S as a novel regulator for serotonin level enhances brain plasticity and locomotion behavior. *FASEB J.*, **34**, 3267–3288.
51. Stolt,C.C., Lommes,P., Friedrich,R.P. and Wegner,M. (2004) Transcription factors Sox8 and Sox10 perform non-equivalent roles during oligodendrocyte development despite functional redundancy. *Development*, **131**, 2349–2358.
52. Artegiani,B., Lyubimova,A., Muraro,M., van Es,J.H., van Oudenaarden,A. and Clevers,H. (2017) A single-cell RNA sequencing study reveals cellular and molecular dynamics of the hippocampal neurogenic niche. *Cell Rep.*, **21**, 3271–3284.
53. Perez-Frances,M., van Gurp,L., Abate,M.V., Cigliola,V., Furuyama,K., Bru-Tari,E., Oropeza,D., Carreaux,T., Fujitani,Y., Thorel,F. *et al.* (2021) Pancreatic Ppy-expressing gamma-cells display mixed phenotypic traits and the adaptive plasticity to engage insulin production. *Nat. Commun.*, **12**, 4458.
54. Davis,M.R., Arner,E., Duffy,C.R., De Sousa,P.A., Dahlman,I., Arner,P. and Summers,K.M. (2016) Expression of FBN1 during adipogenesis: relevance to the lipodystrophy phenotype in Marfan syndrome and related conditions. *Mol. Genet. Metab.*, **119**, 174–185.
55. Attanasio,M., Pratelli,E., Porciani,M.C., Evangelisti,L., Torricelli,E., Pellicano,G., Abbate,R., Gensini,G.F. and Pepe,G. (2013) Dural ectasia and FBN1 mutation screening of 40 patients with Marfan syndrome and related disorders: role of dural ectasia for the diagnosis. *Eur. J. Med. Genet.*, **56**, 356–360.
56. Jespersen,K., Liu,Z., Li,C., Harding,P., Sestak,K., Batra,R., Stephenson,C.A., Foley,R.T., Greene,H., Meisinger,T. *et al.* (2020) Enhanced Notch3 signaling contributes to pulmonary emphysema in a murine model of Marfan syndrome. *Sci. Rep.*, **10**, 10949.
57. Summers,K.M., Raza,S., van Nimwegen,E., Freeman,T.C. and Hume,D.A. (2010) Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome. *Eur. J Hum. Genet.*, **18**, 1209–1215.
58. Anderson,J.L., Head,S.I., Rae,C. and Morley,J.W. (2002) Brain function in Duchenne muscular dystrophy. *Brain*, **125**, 4–13.
59. O'Rourke,J.G., Bogdanik,L., Yanez,A., Lall,D., Wolf,A.J., Muhammad,A.K., Ho,R., Carmona,S., Vit,J.P., Zarrow,J. *et al.* (2016) C9orf72 is required for proper macrophage and microglial function in mice. *Science*, **351**, 1324–1329.
60. Walsh,N.C., Cahill,M., Carninci,P., Kawai,J., Okazaki,Y., Hayashizaki,Y., Hume,D.A. and Cassady,A.I. (2003) Multiple tissue-specific promoters control expression of the murine tartrate-resistant acid phosphatase gene. *Gene*, **307**, 111–123.
61. Mitic,N., Valizadeh,M., Leung,E.W., de Jersey,J., Hamilton,S., Hume,D.A., Cassady,A.I. and Schenk,G. (2005) Human tartrate-resistant acid phosphatase becomes an effective ATPase upon proteolytic activation. *Arch. Biochem. Biophys.*, **439**, 154–164.
62. Lang,P., van Harmelen,V., Ryden,M., Kaaman,M., Parini,P., Carneheim,C., Cassady,A.I., Hume,D.A., Andersson,G. and Arner,P.

(2008) Monomeric tartrate resistant acid phosphatase induces insulin sensitive obesity. *PLoS One*, **3**, e1713.
63. Sengupta,A., Rhoades,S.D., Kim,E.J., Nayak,S., Grant,G.R., Meerlo,P. and Weljie,A.M. (2017) Sleep restriction induced energy, methylation and lipogenesis metabolic switches in rat liver. *Int. J. Biochem. Cell Biol.*, **93**, 129–135.
64. Huang,X., Cai,H., Ammar,R., Zhang,Y., Wang,Y., Ravi,K., Thompson,J. and Jarai,G. (2019) Molecular characterization of a precision-cut rat liver slice model for the evaluation of antifibrotic compounds. *Am. J. Physiol. Gastrointest. Liver Physiol.*, **316**, G15–G24.
65. Kimball,S.R., Horetsky,R.L. and Jefferson,L.S. (1995) Hormonal regulation of albumin gene expression in primary cultures of rat hepatocytes. *Am. J. Physiol.*, **268**, E6–E14.
66. Qvartskhava,N., Lang,P.A., Gorg,B., Pozdeev,V.I., Ortiz,M.P., Lang,K.S., Bidmon,H.J., Lang,E., Leibrock,C.B., Herebian,D. *et al.* (2015) Hyperammonemia in gene-targeted mice lacking functional hepatic glutamine synthetase. *Proc. Natl Acad. Sci. U.S.A.*, **112**, 5521–5526.
67. Pridans,C., Irvine,K.M., Davis,G.M., Lefevre,L., Bush,S.J. and Hume,D.A. (2020) Transcriptomic analysis of rat macrophages. *Front. Immunol.*, **11**, 594594.
68. Hume,D.A., Caruso,M., Keshvari,S., Patkar,O.L., Sehgal,A., Bush,S.J., Summers,K.M., Pridans,C. and Irvine,K.M. (2021) The mononuclear phagocyte system of the rat. *J. Immunol.*, **206**, 2251–2263.
69. Patkar,O.L., Caruso,M., Teakle,N., Keshvari,S., Bush,S.J., Pridans,C., Belmer,A., Summers,K.M., Irvine,K.M. and Hume,D.A. (2021) Analysis of homozygous and heterozygous Csf1r knockout in the rat as a model for understanding microglial function in brain development and the impacts of human CSF1R mutations. *Neurobiol. Dis.*, **151**, 105268.
70. Kohyama,M., Ise,W., Edelson,B.T., Wilker,P.R., Hildner,K., Mejia,C., Frazier,W.A., Murphy,T.L. and Murphy,K.M. (2009) Role for Spi-C in the development of red pulp macrophages and splenic iron homeostasis. *Nature*, **457**, 318–321.
71. Rojo,R., Pridans,C., Langlais,D. and Hume,D.A. (2017) Transcriptional mechanisms that control expression of the macrophage colony-stimulating factor receptor locus. *Clin. Sci. (Lond.)*, **131**, 2161–2182.
72. Irvine,K.M., Caruso,M., Cestari,M.F., Davis,G.M., Keshvari,S., Sehgal,A., Pridans,C. and Hume,D.A. (2020) Analysis of the impact of CSF-1 administration in adult rats using a novel Csf1r-mApple reporter gene. *J. Leukoc. Biol.*, **107**, 221–235.
73. Summers,K.M. and Hume,D.A. (2017) Identification of the macrophage-specific promoter signature in FANTOM5 mouse embryo developmental time course data. *J. Leukoc. Biol.*, **102**, 1081–1092.
74. Ben-Moshe,S. and Itzkovitz,S. (2019) Spatial heterogeneity in the mammalian liver. *Nat. Rev. Gastroenterol. Hepatol.*, **16**, 395–410.
75. Halpern,K.B., Shenhav,R., Matcovitch-Natan,O., Toth,B., Lemze,D., Golan,M., Massasa,E.E., Baydatch,S., Landen,S., Moor,A.E. *et al.* (2017) Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, **542**, 352–356.
76. Atger,F., Gobet,C., Marquis,J., Martin,E., Wang,J., Weger,B., Lefebvre,G., Descombes,P., Naef,F. and Gachon,F. (2015) Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver. *Proc. Natl Acad. Sci. U.S.A.*, **112**, E6579–E6588.
77. Cheng,X., Kim,S.Y., Okamoto,H., Xin,Y., Yancopoulos,G.D., Murphy,A.J. and Gromada,J. (2018) Glucagon contributes to liver zonation. *Proc. Natl Acad. Sci. U.S.A.*, **115**, E4111–E4119.
78. Heidenreich,S., Weber,P., Stephanowitz,H., Petricek,K.M., Schutte,T., Oster,M., Salo,A.M., Knauer,M., Goehring,I., Yang,N. *et al.* (2020) The glucose-sensing transcription factor ChREBP is targeted by proline hydroxylation. *J. Biol. Chem.*, **295**, 17158–17168.
79. Jiang,Y., Feng,D., Ma,X., Fan,S., Gao,Y., Fu,K., Wang,Y., Sun,J., Yao,X., Liu,C. *et al.* (2019) Pregnane X receptor regulates liver size and liver cell fate by yes-associated protein activation in mice. *Hepatology*, **69**, 343–358.
80. Gialitakis,M., Tolaini,M., Li,Y., Pardo,M., Yu,L., Toribio,A., Choudhary,J.S., Niakan,K., Papayannopoulos,V. and Stockinger,B. (2017) Activation of the aryl hydrocarbon receptor interferes with early embryonic development. *Stem Cell Rep.*, **9**, 1377–1386.
81. Taniguchi,M. and Yoshida,H. (2017) TFE3, HSP47, and CREB3 pathways of the mammalian Golgi stress response. *Cell Struct. Funct.*, **42**, 27–36.
82. Gosselin,D., Skola,D., Coufal,N.G., Holtman,I.R., Schlachetzki,J.C.M., Sajti,E., Jaeger,B.N., O'Connor,C., Fitzpatrick,C., Pasillas,M.P. *et al.* (2017) An environment-dependent transcriptional network specifies human microglia identity. *Science*, **356**, eaal3222.
83. Keshvari,S., Caruso,M., Teakle,N., Batoon,L., Sehgal,A., Patkar,O.L., Ferrari-Cestari,M., Snell,C.E., Chen,C., Stevenson,A. *et al.* (2021) CSF1R-dependent macrophages control postnatal somatic growth and organ maturation. *PLoS Genet.*, **17**, e1009605.
84. Niederkorn,M., Agarwal,P. and Starczynowski,D.T. (2020) TIFA and TIFAB: FHA-domain proteins involved in inflammation, hematopoiesis, and disease. *Exp. Hematol.*, **90**, 18–29.
85. Conforto,T.L., Zhang,Y., Sherman,J. and Waxman,D.J. (2012) Impact of CUX2 on the female mouse liver transcriptome: activation of female-biased genes and repression of male-biased genes. *Mol. Cell. Biol.*, **32**, 4611–4627.
86. Lau-Corona,D., Suvorov,A. and Waxman,D.J. (2017) Feminization of male mouse liver by persistent growth hormone stimulation: activation of sex-biased transcriptional networks and dynamic changes in chromatin states. *Mol. Cell. Biol.*, **37**, e00301-17.
87. Lahnemann,D., Koster,J., Szczurek,E., McCarthy,D.J., Hicks,S.C., Robinson,M.D., Vallejos,C.A., Campbell,K.R., Beerenwinkel,N., Mahfouz,A. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.
88. Millard,S.M., Heng,O., Opperman,K.S., Sehgal,A., Irvine,K.M., Kaur,S., Sandrock,C.J., Wu,A.C., Magor,G.W., Batoon,L. *et al.* (2021) Fragmentation of tissue-resident macrophages during isolation confounds analysis of single-cell preparations from mouse hematopoietic tissues. *Cell Rep.*, **37**, 110058.
89. Suo,S., Zhu,Q., Saadatpour,A., Fei,L., Guo,G. and Yuan,G.C. (2020) Revealing the critical regulators of cell identity in the mouse cell atlas. *Cell Rep.*, **25**, 1436–1445.