Combination of novel and public RNA-seq datasets to generate an mRNA expression atlas for the domestic chicken

Stephen J. Bush[1], Lucy Freem[1], Amanda J. MacCallum[1], Jenny O'Dell[1, 2], Chunlei Wu[3], Cyrus Afrasiabi[3], Androniki Psifidi[1, 4], Mark P. Stevens[1], Jacqueline Smith[1], Kim M. Summers[1,5], David A. Hume[1, 5 *]

[1] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Edinburgh, Midlothian, EH25 9RG, UK

[2] Current address: Rivers and Lochs Institute, Inverness College, University of the Highlands and Islands, Research Hub, 1 Inverness Campus, Inverness, IV2 5NA, UK

[3] Department of Integrative and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

[4] Current address: The Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire, AL9 7TA, UK

[5] Current address: Mater Research-University of Queensland, Translational Research Institute, 37 Kent Street, Woolloongabba, Queensland 4102, Australia

[*]Corresponding author: david.hume@uq.edu.au, david.hume@roslin.ed.ac.uk

Email addresses:

Stephen J. Bush	stephen.bush@roslin.ed.ac.uk
Lucy Freem	lucy.freem@roslin.ed.ac.uk

Amanda J. MacCallum	amanda.maccallum@roslin.ed.ac.uk
Jenny O'Dell	jenny.odell.ic@uhi.ac.uk
Chunlei Wu	cwu@scripps.edu
Cyrus Afrasiabi	cyrus@scripps.edu
Androniki Psifidi	apsifidi@rvc.ac.uk
Mark P. Stevens	mark.stevens@roslin.ed.ac.uk
Jacqueline Smith	jacqueline.smith@roslin.ed.ac.uk
Kim M. Summers	kim.summers@roslin.ed.ac.uk
David A. Hume	david.hume@uq.edu.au, david.hume@roslin.ed.ac.uk

ABSTRACT

Background

The domestic chicken (*Gallus gallus*) is widely used as a model in developmental biology and is also an important livestock species. We describe a novel approach to data integration to generate an mRNA expression atlas for the chicken spanning major tissue types and developmental stages, using a diverse range of publicly-archived RNA-seq datasets and new data derived from immune cells and tissues.

Results

Randomly down-sampling RNA-seq datasets to a common depth and quantifying expression against a reference transcriptome using the mRNA quantitation tool Kallisto ensured that disparate datasets explored comparable transcriptomic space. The network analysis tool Miru was used to extract clusters of co-expressed genes from the resulting expression atlas, many of which were tissue or cell-type restricted, contained transcription factors that have previously been implicated in their regulation, or were otherwise associated with biological processes, such as the cell cycle. The atlas provides a resource for the functional annotation of genes that currently have only a locus ID. We cross-referenced the RNA-seq atlas to a publicly available embryonic Cap Analysis of Gene Expression (CAGE) dataset to infer the developmental time course of organ systems, and to identify a signature of the expansion of tissue macrophage populations during development.

Conclusion

Expression profiles obtained from public RNA-seq datasets – despite being generated by different laboratories using different methodologies – can be made comparable to each other. This meta-analytic approach to RNA-seq can be extended with new datasets from novel tissues, and is applicable to any species.

INTRODUCTION

Aggregation and meta-analysis of multiple large gene expression datasets based upon common microarray platforms is relatively commonplace in many species (e.g. [1-3]). Although RNA-seq is rapidly supplanting microarrays for gene expression profiling, it is not yet clear whether data from multiple different labs can be analysed together in an informative manner. Confounding variables reflect the many technical – and bias-prone – aspects of library preparation and sequencing (see reviews [4, 5]), with RNA-seq datasets often differing in read length [6], depth of coverage [7], strand specificity [8], RNA extraction and library selection methods [9], sequencing platform [10, 11] and the choice to sequence single- or paired-end reads [12]. For a given dataset, these variables can together affect both the number and type of genes detectable and the accuracy of their expression level estimates. Expression quantification is also affected by sample quality [13] and storage method [14], irrespective of sequencing technique: RNA degrades with lengthier post-mortem intervals [15] (the extent of which is tissue-dependent [16]) with degradation resulting in inaccurate quantification, particularly for shorter transcripts [17]. Sequencing composite biological structures (those with internal structures that have distinct functions), whether intentionally or inadvertently, can mask the signal of structure-specific differential expression [18]. Despite these variables, meta-analysis combining mammalian gene expression datasets [19-21] suggests that RNA-seq datasets are generally robust to inter-study variation, with the expression profiles of homologous tissues clustering more closely with each other than with different samples from the same study or species [22].

Expression atlases are valuable resources for functional genomics. Groups of transcripts – members of which will have similar expression profiles – can be associated with a shared function, such as a particular pathway or biological process. This principle is known as ‘guilt by association’ [23] and has previously been used to annotate genes of unknown function in human [2, 24, 25], pig [26], sheep [27] and mouse [28, 29] datasets. Co-expression information is also informative in genome-wide association studies (GWAS) of complex traits and disease susceptibility. The simple principle, that genes involved in the same trait or phenotype tend to be expressed in the same cell type or tissue, or otherwise participate in the same pathway, has been confirmed in multiple datasets [28, 30].

Because of the ease of access *in ovo*, the chicken (*Gallus gallus*) embryo has been widely used as a model system in cell and developmental biology, constrained only by methods for genomic manipulation *in situ*, or in the germ line. These constraints were largely overcome through the sequencing of the genome, and technological developments such as *in vivo* electroporation, more than 15 years ago [31, 32]. More recent innovations including the generation of reporter transgenes [33] and genome editing via primordial germ cells [34-36] have transformed the utility of the chicken as a model organism. However, the current genome build still has many unannotated or minimally annotated genes about which very little is known [28]. Of the 18,347 protein-coding genes in version GalGal5 of the chicken genome in Ensembl89, 7275 (40%) have only been assigned an Ensembl placeholder ID.

The domestic chicken is also a major source of animal protein worldwide, with different lines heavily selected for optimal production traits such as increased egg production or rapid weight gain. The molecular basis for these traits is increasingly being associated with genomic loci through genome-wide association studies based upon high density SNP platforms [37]. Both the application of the chick as a model organism, and for candidate gene analysis in genomic intervals associated with trait variation, would be expedited by improvements in functional genome annotation. In particular, it would be useful to identify the sets of protein-coding genes that share transcriptional regulation between the chick and the mouse, the most widely-studied mammalian model organism. For this purpose, we aimed to generate a comprehensive atlas of mRNA expression for the chicken.

With the removal of antibiotics from the food chain and threats from emerging diseases, there is also interest in the selection of birds with increased disease resistance and/or resilience [38]. To support this activity, we were particularly interested in identifying and annotating genes expressed specifically at high levels in cells of the innate immune system. Such gene sets have been identified in previous studies of human [2, 24, 25], pig [26], sheep [27] and mouse [28].

The current version of the chicken assembly was largely derived from high-throughput (i.e. comparatively cheap but imprecise) short read sequencing and primarily contains protein-coding gene models. The recent use of long-read – PacBio SMRT Iso-Seq – data has demonstrated that the transcriptomic complexity of chickens is comparable to humans, with many additional lncRNA models (among others) scheduled for inclusion in future Ensembl annotations [39].

To identify the set of genes expressed in innate immune cells in both unchallenged and activated conditions, we generated pure cultures of bone marrow-derived macrophages (BMDMs) grown in the presence of recombinant chicken macrophage colony-stimulating factor (CSF1), and stimulated them with the archetypal microbial agonist, lipopolysaccharide (LPS) [40]. To complement the data generated from macrophages *in vitro*, we also obtained RNA-seq libraries from the caecal tonsils of birds infected with *Campylobacter*, as well as from previous studies of macrophage, dendritic cell and heterophil populations. A global expression atlas for the chicken transcriptome was created by combining our immune-related data with 20 publicly archived RNA-seq datasets. Some were collated by the Avian RNA-seq Consortium [41], while others are drawn from a diverse range of existing publications, including studies that characterised the genetic basis of retinogenesis [42], the genetic determinants of meat tenderness [43], the morphological diversity of skin appendages [44], visceral fat metabolism [45], the transition between laying and brooding phases [46], the effect of heat stress upon pituitary development [47] and spleen function [48], the pathways involved in avian influenza resistance [49], the role of lncRNAs in the development of muscle [50], liver and adipose [51], and the transcriptional landscape of mRNA editing [52]. In total, 279 RNA-seq libraries were obtained, representing 48 distinct tissue and cell types at developmental stages spanning early embryonic (5 days) to mature adult (70 weeks post-hatching). In addition, we accessed a recently published transcriptional analysis of chick development generated by Cap Analysis of Gene Expression (CAGE)

[53], a technique which can be used to quantify gene expression based on the transcript start site [54]. We show that the 'guilt by association' approach to functional annotation is viable even when combining disparate RNA-seq datasets, and utilise the meta-dataset to identify macrophage-specific and other informative co-expression clusters, providing a resource for genetic and genomic study of avian trait variation.

RESULTS

Selecting samples for inclusion in an RNA-seq meta-dataset

Many chicken RNA-seq datasets are available in public repositories, as detailed in [41]. Robust co-expression clustering of any two genes depends upon sampling tissues and cells in which both vary across the widest possible range. To maximise the co-expression signal, we chose datasets to represent the greatest possible diversity of tissues and organ systems. Not all studies contain links to a publicly archived dataset, such as a study of induced ochratoxicosis in the kidney cortex [55] and two studies of the bursa of Fabricius [56, 57]. Samples containing less than 10 million reads were not used, such as those from a study of the follicular transcriptome throughout the ovulation cycle [58].

Datasets used are detailed in Table S1, and have few commonalities: they were sequenced using a variety of Illumina instruments (HiSeq 2000/2500/3000/4000, Genome Analyzer II/IIx, NextSeq 500 and HiScanSQ), and include single- and paired-end, strand-specific and non-specific, polyA-selected (mRNA-seq) and rRNA-depleted (total RNA-seq) libraries at different read lengths and depths. For 12 tissues, independently sequenced RNA-seq datasets for the same tissue (Table S2) also allow for internal tests of the validity of aggregating the data. Throughout this text studies are referred to by their NCBI BioProject ID.

Quantifying expression by iteratively revising a reference transcriptome

Expression was quantified – as transcripts per million (TPM) – using an RNA-seq processing pipeline [59] which iteratively runs the quantification tool Kallisto [60] with each iteration using an incrementally revised transcriptome. Kallisto requires that the user provide a set of transcripts, which are decomposed into k-mers. The expression of each transcript is quantified by matching this set of k-mers to the k-mers of the reads. For the first iteration of Kallisto, a non-redundant transcriptome (57,234 transcripts, representing 17,680 Ensembl protein-coding genes) was obtained by combining Ensembl transcript models with NCBI mRNA RefSeqs (see Materials and Methods).

The output was first parsed for library quality. The reverse cumulative distribution of TPM per gene was plotted on a log-log scale (Figure 1). The distributions generally approximate a power-law with an exponent of approximately -1 (Table S3), consistent with Zipf's law (that the probability of an observation is inversely proportional to its rank) [61, 62]. Four samples with exponents < -0.8 or > -1.2, i.e. deviating > 20% from the optimal value of -1 – were excluded from further analysis (i.e. the next iteration of Kallisto) (Table S3). Using only data from the useable samples, we created a revised reference transcriptome. During the first iteration of Kallisto, 55,027 of 57,234 transcripts (96%) were detectably expressed (average TPM > 1 in at least one tissue, where the average is the median TPM across all replicates, per BioProject, of that tissue), representing 17,313 Ensembl protein-coding genes (Table S4). After excluding 2207 transcripts with TPM < 1 in all tissues (Table S5) and those detectable only in the 4 excluded samples (n = 57), a revised transcriptome was generated containing 54,970 transcripts. For the second iteration of Kallisto, expression was re-quantified using this revised transcriptome, creating a final set of gene-level TPM estimates. The overall meta-dataset provides gene-level expression for 23,864 gene models (both Ensembl and NCBI) as median TPM across all replicates, per BioProject, per tissue (Table S6). Of these gene models, 43% (10,090) were unannotated, having only either an Ensembl placeholder ID or an NCBI locus ID.

Randomly down-sampling RNA-seq datasets does not quantitatively alter their expression profiles

Higher resolution expression profiles are dependent upon higher sequencing depths [63] with diminishing returns – after approximately 10 million reads – on the power to detect genes differentially expressed between conditions [64]. For the purpose of functional annotation, it is more important to minimise variation between samples than to comprehensively capture transcripts. Accordingly, all datasets were randomly down-sampled to exactly 10 million reads before quantification.

To ensure the resulting co-expression signals are reproducible, it is necessary to establish that there are no significant differences in expression profiles introduced by sampling. For instance, the LPS-stimulated BMDM datasets were sequenced at depths of 37.5 to 52.6 million reads, such that when down-sampling, the BMDM expression profile as quantified for the meta-dataset was obtained using approximately one fifth to one quarter of the original reads (Table S7). To validate the approach, we randomly down-sampled each BMDM dataset to 10 million reads 100 times, using seqtk (https://github.com/lh3/seqtk, downloaded 29th November 2016) seeded with a random integer between 0 and 10,000 (Dataset S1). After performing an all-against-all correlation of the 100 sets of data, the average Spearman's *rho* was > 0.96 (Table S8), with the absolute difference, per gene, between maximum and minimum expression level averaging approximately 8 TPM (Figure 2 and Table S9). 70-75% of the genes detectably expressed (TPM > 1) in at least one of the 100 random samples were detected in all 100 samples (Table S8). Conversely, <5% of the genes were detectable in <5% of the samples (Table S8). The detection of these genes was stochastic, as they were expressed at very low levels – on average, 1.3 TPM (Table S8).

Biologically meaningful expression profiles are identified even after combining disparate RNA-seq datasets

If a meta-analytic approach to RNA-seq is valid, subsets of transcripts enriched in a given tissue should have annotations functionally appropriate to that tissue. To test this, we calculated a preferential expression measure (PEM) for each gene [65], essentially the median expression divided by the mean. We then obtained the set of Gene Ontology (GO) terms enriched in each subset of genes with the highest PEM associated with a particular tissue (Table S10) (see Materials and Methods). Consistent with the function of each tissue, the bursa of Fabricius (the site of B cell synthesis [66]) showed tissue-specificity for the expression of genes enriched for 'defence response to bacterium' ($p = 8.3x10^{-5}$), breast muscle for 'striated muscle contraction' ($p = 1.9x10^{-6}$), cerebrum for 'synaptic transmission' ($p = 1.5x10^{-4}$), claw epithelium for 'bone mineralisation' ($p = 6.4x10^{-4}$), heart for both 'muscle contraction' ($p = 8.8x10^{-6}$) and 'cellular respiration' ($p = 4.6x10^{-15}$), kidney for 'oxidation-reduction process' ($p = 5.3x10^{-5}$), pancreas for 'proteolysis' ($p = 0.001$), pituitary gland for 'endocrine system development' ($p = 2x10^{-4}$), retina for 'visual perception' ($p = 7.2x10^{-17}$), spleen for 'immune response' ($p = 2.2x10^{-6}$), and trachea for 'cilium morphogenesis' ($p < 1x10^{-30}$) (Table S10). In an all-against-all correlation matrix (Pearson's *r*) (Table S11), the expression profiles of like tissues were correlated regardless of their BioProject of origin (Table S12). A sample-to-sample network graph also demonstrates that samples of the same or related tissues cluster together (Figure 3). Taken together, these results validate the aggregation of data from multiple sources to create an informative expression atlas.

Signals of co-expression allow for informative functional annotation

Network analysis of the meta-dataset was performed using Miru, a commercial version of BioLayout *Express*^{3D} [67, 68], previously applied to pig [26], sheep [27] and mouse [28] microarray datasets and CAGE data from the FANTOM5 consortium [24, 25]. A Pearson's correlation matrix for each gene-to-gene comparison was visualised as a network graph of 18,127 nodes (genes) linked by 632,038 edges (correlations above a certain threshold; in this case, $r = 0.8$). Clusters of interconnected nodes represent sets of genes that share a signal of co-expression. These clusters were identified by applying the Markov clustering (MCL) algorithm [69] to the network graph, at an inflation value (which determines cluster granularity) of 2.2. The contents of each cluster are given in Table S13.

Many of the co-expression clusters comprised genes with a tissue- or process-specific expression profile. Table S14 summarises the highest PEM value for a tissue in each of the clusters with >25 members. Cluster 2 was largely brain-specific: of the 655 genes in this cluster, 281 (43%) had their highest PEM in the hypothalamus, 155 (24%) had their highest PEM in the cerebrum and 115 (18%) had their highest PEM in the cerebellum. Other clusters contained genes with expression enriched in liver (cluster 6), ovary (cluster 7), trachea (cluster 8), testis (cluster 10), retina (clusters 13 and 24), feather epithelium (cluster 14), breast muscle (cluster 16), kidney (cluster 17), pituitary gland (clusters 19 and 25), *Campylobacter*-infected caecal tonsils (cluster 20), spleen (clusters 21 and 22) and adipose (cluster 23).

The tissues in some of these clusters were represented by multiple independent projects combined in this meta-atlas. For instance, cluster 6 comprises genes that were enriched in the liver, with data from three separate BioProjects. Some variation in expression estimates between these independent liver samples did not affect their inclusion in the same co-expression cluster. Furthermore, the GO terms enriched in each cluster are functionally consistent with its observed tissue-specificity (Table S15).

Some clusters were associated with processes shared by multiple tissues. The largest cluster, cluster 1, was enriched in embryo-derived samples, and the GO terms are associated strongly with the cell division cycle and DNA repair (Table S15). The genes within this list include the key transcriptional regulator, *FOXM1*, and multiple cyclins (*CCNA2/B2B3/C/E1/F* and *J*), and overlap substantially with cell cycle-associated lists derived from previous cluster analysis [2, 70].

We used the ‘guilt by association’ principle to contextualise individual gene annotations – obtained by protein-level alignment and of varying quality (see Materials and Methods) – as there is an *a priori* expectation that by virtue of being co-expressed, the genes within a given cluster have related (that is, tissue- or process-specific) functions. In this respect, we can increase confidence in otherwise lower-quality alignments. Some examples and proposed annotations are summarised in Table S13. The co-expression profile is especially informative for clusters with few known genes. For instance, cluster 14 contains 210 genes expressed largely in the feather epithelium (Table S13). 93% of the genes within this cluster are unannotated, with only 14 genes having a known function (Table 1). Collectively, the functions of these genes are biologically consistent with an epithelium-enriched expression profile. Of the 196 unannotated genes, 86% can be aligned to feather keratins (representing 86 of the 96 genes with only an Ensembl ID and 83 of the 100 genes with only an NCBI RefSeq ID) (Table S13). Other unannotated genes include paralogues of existing genes in the cluster (ENSGALG00000004358 shares homology with *AMZ1*, ENSGALG00000029002 with *XG* and LOC428538 with *SDR16C5*), probable members of the keratin-associated protein family, which have essential roles in hair shaft formation [71] (ENSGALG00000018878, ENSGALG00000044257, LOC101751162, LOC101751279, LOC107055127, LOC107055128 and LOC107055130), a gene with homology to the tight junction protein claudin 4 (ENSGALG00000035131) [72], and several transcripts with homology to uricases (LOC101747367, LOC107056676 and LOC107056678), enzymes which degrade uric acid (the end point of purine metabolism) [73], notable because purines act as pigments in avian feathers [74].

Annotation of co-expression clusters associated with innate and acquired immunity and macrophage biology

The most prominent set of genes co-expressed in macrophages was cluster 4 (n = 458 genes; 129 [28%] are unannotated), in which > 60% of the genes have their highest PEM for BMDMs 24 hours post-LPS stimulation (Figure 4 and Table S14). This cluster is internally validated by the presence of transcripts encoding numerous known myeloid effectors/receptors (e.g. *C3AR1*, *CCR2*, *CD40*, *CYBB*, *CLEC5A*, *DCSTAMP*, *NLRC5*, *METRK*, *MYD88*, *TLR4*), lysosomal components (e.g. *CTSB*, *LAMP1*, *M6PR*) and multiple transcription factors (*BATF3*, *CEBPB*, *IRF1*, *NFE2L2*, *NRR1H3* [also known as *LXRA*], *SPI1* [also known as *PU.1*], *STAT1*, *TFEC*) that are also macrophage-enriched in mouse and human [75]. Their co-expression strongly indicates that basic macrophage transcriptional regulation is conserved between birds and mammals. Accordingly, the provisional annotations of genes that lack an informative name in this cluster, shown in Table S13, are given extra weight by their association. Other macrophage clusters include cluster 34 (n = 93 genes; 72 [77%] are unannotated) and cluster 37 (n = 79 genes; 16 [20%] are unannotated), in both of which the majority of genes had their highest PEM for the HD11 immortalised macrophage cell line (from BioProject PRJEB1406): 98% and 90%, respectively (Table S14). The smallest macrophage-specific cluster was cluster 84 (n = 26 genes; 19 [73%] are unannotated), in which every gene had its highest PEM for BMDMs treated with CSF1 (from BioProject PRJEB7662) (Table S14).

The *CSF1R* gene was contained within cluster 27 (n = 129 genes, of which 32 [25%] are unannotated), which had an expression profile shared by both dendritic cells and macrophages. 36% of the genes in cluster 27 had their highest PEM for dendritic cells and 26% for untreated BMDMs (both samples from BioProject PRJEB7475), with the remaining 26% for BMDMs treated with CSF1 (from BioProject PRJEB7662) (Table S14). This cluster also contained the lipopolysaccharide receptor and commonly used monocyte marker, *CD14*, several genes (*C1QA/B/C*, *MARCO*, *P2RY12/13*, and *STAB1*) that are associated with tissue-specific macrophage populations in mice [76], and a single myeloid-associated transcription factor, *MAFB*, which is required for tissue macrophage development in mice [77]. The cells referred to as dendritic cells are bone marrow cells grown in *GM-CSF* (*CSF2*), rather than *CSF1*. As noted in previous analyses of mouse [78] and human [79] transcriptomes, cells differentiated in *GM-CSF* have much more in common with macrophages than with classical dendritic cells dependent upon *FLT3*-ligand.

The clusters associated with the acquired immune response, predominantly B and T cells, are somewhat smaller and poorly-annotated (clusters 20, 21, 22, 29 and 78). Cluster 21, expressed most highly in spleen, contains *TIMD4* (ENSGALG00000003876), which promotes T-cell expansion and survival [80], and is enriched with B cell-associated genes, including the B cell transcription factors *BATF*, *IRF4*, *PAX5*, *RUNX3*, and *SPIC*, as well as the class II trans-activator *CIITA*, class II subunit *CD74* and the class II MHC gene *BLB2*. The thymus-enriched cluster 29 contains *CD4*, the recombination activating genes *RAG1* and *RAG2*, and the T cell transcription factors *LEF1*, *RORC* and *TCF7*.

Integrating gene expression and protein-protein interaction networks

Biological systems can be functionally organised into many different (and intersecting) networks based on the nature of their interaction, including – aside from gene co-expression networks – metabolic/biochemical networks, signal transduction networks, regulatory networks, and protein-protein interaction (PPI) networks [81]. Data from different networks can be integrated: for instance, subunits of the same protein complex are known to be co-expressed [82], with those genes present in both a co-expression and PPI network having a high probability of performing similar functions [83]. We therefore determined the set of genes present in both the same co-expression cluster and a PPI network (Table S16), obtaining chicken PPI data by mapping human PPIs to orthologous chicken genes (see Materials and Methods). The PPI and co-expression data are mutually supportive. For example, there were 32 PPIs among the genes in the macrophage-specific cluster 4. These include *STAT1* (signal transducer and activator of transcription 1-alpha/beta) – a critical mediator of the pro-inflammatory response of macrophages to LPS [84] – and the transcription factors *ATF3*, a known inducer of *STAT1* [85], and *SPI1/PU.1*, which is essential for macrophage differentiation [86]. Also in the network are the tyrosine kinase *LYN*, which is activated alongside *STAT1* in response to *IL5* (a key mediator of eosinophil activation [87]), and the adaptor protein *GRB2*, which facilitates the activation of *ERK* by tyrosine kinases [88] (*ERK* signalling is essential to macrophage development [89]). In addition, the network contained *SOCS3*, a negative regulator of cytokine signalling that inhibits the nuclear translocation of *STAT1* in response to *IFN* stimulation [90], with this stimulation a key constituent of classical macrophage activation [91].

Integrating gene expression and promoter expression networks

Relatively few RNA-seq datasets were available for chicken embryonic development. Lizio *et al.* [53] have recently analysed the time course of chicken development using Cap Analysis of Gene Expression (CAGE). Their dataset complements a CAGE-based analysis of gene expression in multiple tissues of the mouse during embryonic development [92]. Network analysis of the mouse dataset revealed a signature of the expansion of the tissue macrophage populations during embryonic development, and the inverse relationship between cell proliferation and tissue-specific differentiation in each organ [93]. Analysis of a macrophage-specific transgene in birds revealed that, as in mammals, macrophages are first produced by the yolk sac, progressively infiltrate the embryo and expand in number to become a major cell population in every organ [33, 94]. The expression atlas we have developed provides a complementary resource for adult tissues and includes a time course of embryonic development. By combining the atlas with the CAGE data, it would be possible to infer the developmental time course of organ systems in the chicken. We obtained the chicken CAGE data of Lizio *et al.* [53] and clustered the promoter-based expression levels in the same manner as for the RNA-seq atlas. Figure 5 shows the resulting network graph, and the average expression profiles of a subset of clusters. Table S17 provides a full list of promoters in each of the co-expression clusters and their average expression profiles. As discussed by Lizio, *et al.*, the embryonic CAGE data identify transcription start sites for many tissue-specific and regulated genes, including developmental regulators such as *brachyury*. The intersection of the CAGE and RNA-seq clusters is presented in Table S18. Not surprisingly, the largest promoter cluster overlapped substantially with cluster 1 in the RNA-seq atlas which was embryo-enriched in expression. It contained numerous developmental regulators, anabolic/cell cycle, and mitochondria-associated genes with an average profile of down-regulation during development (Figure 5). Aside from the whole embryo profiles, the CAGE data contains several additional samples, including bone marrow-derived mesenchymal stem cells (MSC), aortic smooth muscle cells (ASMC), hepatocytes, extra-embryonic tissues and both leg and wing buds. Each of the samples was enriched for specific promoters that also varied during development and accordingly defined clusters. Clusters 2 and 10 of the CAGE data were enriched in MSC and ASMC, and contained many mesenchyme-associated genes including multiple collagens and other connective tissue-associated transcripts. CAGE clusters 4 and 9 were hepatocyte-enriched and most likely track the development of the liver during development. Cluster 4, shown in Figure 5, contains the transcription factor *HNF1A,* and many of the transcripts within it encode secreted proteins such as complement components and clotting factors. CAGE cluster 5 (Figure 5) contains the muscle-specific transcription factors *MYOD1*, *MYOG* and *SOX2*, and numerous skeletal muscle-associated genes in common with cluster 16 from the RNA-seq atlas, and increases in expression throughout development. The transcripts within cluster 5 are not expressed in the aortic smooth muscle cells. CAGE clusters 7, 16, 18 and 19 contained transcripts that were expressed transiently at different stages of embryonic development, including multiple members of the *HOX* and *CDX* families. CAGE clusters 8 and 25 both contained promoters of multiple genes that are expressed specifically in macrophages in the RNA-seq atlas (clusters 4 and 27). The average expression profiles are shown in Figure 5, with representative genes indicated. The macrophage-specific transcription factor *SPI1*, and most other macrophage-enriched genes within CAGE clusters 8 and 25, fall within the larger macrophage-associated clusters (4, 27 and 31) within the RNA-seq atlas. Interestingly, CAGE cluster 25 appears to be enriched for genes expressed specifically in brain macrophages (microglia), including *CSF1R*, *C1QA*, *C1QB*, *C1QC*, *CTSS*, *DOCK2*, *HAVCR1 LAPTM5*, *LY86*, *MPEG1*, and *P2RY13* [95], which in mice appear to develop from yolk sac progenitors rather than definitive haematopoiesis [96]. Several other microglia/macrophage-associated transcripts, notably *CX3CR1*, *P2RY12*, *TIMD4*, and *TREM2*, are detectable in the CAGE data at the same embryonic stage, but did not cluster because their expression differs in the cell populations. In each of the macrophage-associated clusters, there were numerous promoters currently with uninformative annotation, which by inference are likely to be macrophage-related. Consistent with the location of *CSF1R* mRNA and the *CSF1R*-reporter gene in the chicken [33], *CSF1R* and *SPI1* were both first detectable in the embryo at between HH12 and HH14 (day 2), and both increased in parallel during embryonic development. Figure 6 shows the ZENBU (http://fantom.gsc.riken.jp/zenbu/) view of the chicken *CSF1R* locus, identifying the transcription start site downstream of the *PDGFRB* locus, and the time course of appearance of *CSF1R* transcripts in the embryo and their expression in isolated cells. The reason that CAGE clusters 8 and 25 genes separate in the dataset is that they were also detected at high levels in “mesenchymal stem cells” and to a varying extent in “hepatocytes” (Figure 5). In mice, macrophages were shown to be a major contaminant of bone marrow-derived osteoblast cell cultures [97]. Based upon this cluster analysis in the embryo (which reveals separate mesenchyme and hepatocyte-specific clusters), and the atlas data, where these genes were clearly macrophage-enriched, the expression of macrophage-associated genes is almost certainly a reflection of the presence of large numbers of macrophages in these cell populations. Indeed, the set of promoters active in “mesenchymal stem cells” was found to be enriched for binding sites for *SPI1* and *CEBPA*, transcription factors that can induce the transdifferentiation of lymphoid precursors into macrophages [98].

DISCUSSION

RNA-seq is a multi-step process of reverse transcription, amplification, fragmentation, purification, adaptor ligation and sequencing, with each step subject to error [99]. Such laboratory-specific variation is also independent of intrinsic sequencing biases, which can influence the nucleotide composition of the reads [100] (leading to mismatches between the sequenced read and the original RNA fragment [101]), the GC content of the reads [102], and the sequencing error rate [11]. Despite all of these constraints, Figure 3 shows that in a sample-to-sample network graph of many independently sequenced tissues, the signal of co-expression clearly outweighs the noise. The critical step in reducing the noise, and making the datasets comparable, was to down-size the RNA-seq libraries so that the depth of coverage of the transcriptome was the same in each case. This has the effect of removing a great deal of the stochastic detection of more lowly-expressed transcripts. Figure 2 and Table S9 show that the random sampling used to down-size does not substantially alter the relative expression estimates of any two genes within any given sample, with equivalent expression profiles reconstructed for each of 100 random samples. Combined with the use of Kallisto to quantify expression, which maps a common depth of k-mers to a standardised reference transcriptome, the method we have developed effectively ensured that each RNA-seq library was exploring an equivalent transcriptomic space.

The success of the aggregation of public domain data in terms of genome annotation is evident from the analysis of the membership of co-expression clusters in Table S13. Each cluster clearly contains genes of known function, shows evidence of very strong GO enrichment, and as noted in similar array-based studies [2, 26] commonly contains the transcription factors that regulate the other members of the cluster. On that basis, it would be reasonable to provisionally assign the same GO terms to genes of unknown function, at least within the larger clusters. For example, the genes within cluster 1 that are not currently functionally annotated or assigned a clear orthologue are likely to be involved in some way in the cell cycle. Indeed, the provisional annotations of many of them shown in Table S13 indicate this is very likely to be the case. Similarly, the genes we have identified that were enriched in innate and acquired immune cells are likely to be associated with heritable variation in disease resistance/susceptibility.

Detailed examination of individual clusters can provide significant biological insights. Cluster 8, enriched in trachea, and with the second highest expression in lung, was strongly enriched with GO terms associated with cilium, microtubule binding, motor activity and the actin cytoskeleton (Table S15), and includes, for example, multiple members of the cilia and flagella-associated protein (*CFAP*), dynein regulatory complex (*DRC*) and other dynein-related gene families. Mutations in many of these genes have been associated with human ciliopathies [103]. This cluster also contained the transcription factor *FOXJ1*, which is essential for the formation of motile cilia in mice [104]. Provisional annotations of genes of unknown function in this cluster are consistent with the overall enrichment for genes associated with motility. The presence of the epithelial transcription factors *ELF5* and *PAX9* in this cluster suggests both could have a role in regulation of this key gene set, providing a possible reason for the embryonic lethality of the knockouts of each gene [105, 106]. Interestingly, *KIAA0586*, which is also known as *TALPID3*, is in a separate smaller cluster – number 139 – that is more widely expressed. The *TALPID3* protein encodes a centromeric component, and mutation affects the formation of primary, non-motile cilia and signaling by the morphogen sonic hedgehog [107, 108]. Many of the genes that are apparently co-regulated with *TALPID3* have been associated in some way with regulatory functions of primary cilia, including *CEP120* which, like *KIAA0596*, is mutated in human Joubert syndrome [109]. Other members of the cluster may be candidate interactors with *TALPID3*.

The validity of the approach, and of the clusters generated, was established by comparing tissue- and function-specific clusters obtained by an alternate method of quantifying RNA expression levels, CAGE, using a public dataset of chicken embryo development. This showed that tissue-specific developmental gene expression can be detected using whole embryos (as we have previously shown for mouse [93]), and that the genes in the developmental stage clusters matched those found in the adult tissue atlas.

The clustering we have presented is based upon an arbitrary correlation threshold. For every gene of interest, it can be informative to identify its transcriptional companions. To this end, as we have done previously for human [2], pig [26], sheep [27] and mouse [28], we have made the current version of this atlas available as a searchable database using the gene annotation portal BioGPS [110] (http://www.biogps.org/chickenatlas), where one can utilise a simple "find correlated" function to identify genes with similar expression profiles. In turn, this resource allows a rapid comparative assessment of the expression of a gene of interest in mammals and birds and the extent to which functional information is likely to be transferable across species.

The advantage of the aggregation method we have applied is that it is can be extended with new data from tissues and cell types we have not currently included. The larger the dataset, and the greater the transcriptional space sampled, the more stringent the correlations that will be generated and the more likely they are to produce new biological insights.

MATERIALS AND METHODS

Animals

To obtain bone marrow-derived macrophages, nine chickens of approximately 8 weeks of age (3 female and 3 male Ross 308 broilers, and 3 female CSF1R-MacApple transgenic NOVOgen Brown layers) were euthanized by cervical dislocation and confirmed dead by decapitation. Likewise were euthanized 23 broiler chickens, each 5 weeks of age, to obtain the caecal tonsils. All animal work was conducted in accordance with guidelines of the Roslin Institute and the University of Edinburgh and carried out under the regulations of the Animals (Scientific Procedures) Act 1986. Approval was obtained from the Roslin Institute's and the University of Edinburgh's Protocols and Ethics Committees.

Macrophage cell culture and RNA isolation

Bone marrow-derived macrophage (BMDM) culture and challenge *in vivo* were performed as previously described [111]. Chicken bone marrow was cultured for 7 days with 350 ng/µl chicken CSF1 on Sterilin plastic to differentiate BMDMs. Adherent cells were then transferred to tissue culture plastic and cells plated at 80% confluence. BMDMs were challenged with the addition of LPS at 100 ng/ml to culture medium and then harvested after 0 (null condition), and 24 hours. Cells were harvested in TRIzol® (15596018; Thermo Fisher Scientific) and extraction performed with the RNeasy Mini Kit (74106; Qiagen Hilden, Germany) according to manufacturer's instructions.

Collection of Campylobacter-infected caecal tonsils

Birds were naturally exposed to *Campylobacter* spp. under commercial farm conditions. Caeca and caecal tonsil samples were collected in RNAlater (AM7021; Thermo Fisher Scientific, Waltham, USA). *Campylobacter* load in caeca was determined by selective culture as previously described [112]. Seven serial ten-fold dilutions of caecal content were prepared in phosphate-buffered saline and 100 µl plated to mCCDA (modified cefoperazone-deoxycholate agar) supplemented with cefoperazone (32 mg/L) and amphotericin B (10 mg/L; Oxoid), followed by incubation for 48 hours under microaerophilic conditions (5% O2, 5% CO2, and 90% N2) at 41C. Dilutions were plated in duplicate and colonies with morphology typical of *Campylobacter* detected in all samples. RNA was extracted from the caecal tonsils using the RNeasy Mini Kit (74106; Qiagen Hilden, Germany) according to manufacturer's instructions. As chickens were exposed naturally rather than being explicitly challenged with *Campylobacter*, bacterial load varied considerably between individuals. Accordingly, tonsil samples were partitioned into two broad subsets: those from chickens whose caecum has high *Campylobacter* load (>= 10,000 CFU/g), and those with low *Campylobacter* load (< 10,000 CFU/g).

RNA-sequencing

For both BMDM and caecal tonsil samples, library preparation was performed by Edinburgh Genomics. Total RNA (for BMDMs) and mRNA (for caecal tonsils) was, in both cases, sequenced by Edinburgh Genomics at a depth of >40 million strand-specific 75bp paired-end reads per sample, using an Illumina HiSeq 4000. The raw data is deposited in the European Nucleotide Archive under accessions PRJEB22373 (BMDMs) and PRJEB22580 (caecal tonsils).

Public RNA-seq datasets

Publicly accessible datasets used in this study are described in Table S1. The meta-atlas aggregating these data details, per tissue, the associated NCBI BioProject and Sequence Read Archive (SRA) sample IDs (Table S6). All public datasets for this study are available via the SRA, a public repository for sequence data maintained by the International Nucleotide Sequence Database Collaboration (INSDC) and accessible from the websites of its constituent members: known as the SRA if via the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/sra), the DRA (DDBJ Read Archive) if via the DNA Data Bank of Japan (DDBJ) (http://trace.ddbj.nig.ac.jp/dra/), and the European Nucleotide Archive (ENA) if via the European Bioinformatics Institute (EBI) (www.ebi.ac.uk/ena) [113]. For retrieving the raw files used in this study or for expanding this work with new datasets from novel tissues, note that data are directly accessible in fastq format from the ENA and DDBJ but only in a binary .sra format from the NCBI. Decompiling the latter into fastq files – using the fastq-dump tool within the SRA Toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software) –

is far slower than analysing fastq files with Kallisto, and so forms a bottleneck in the expression atlas creation pipeline. For this reason, obtaining fastq files in bulk from NCBI is not recommended unless necessary.

Defining a reference transcriptome and quantifying expression

Prior to expression level quantification, all RNA-seq datasets were randomly down-sampled to 10 million reads using seqtk (https://github.com/lh3/seqtk, downloaded 29th November 2016) with parameter -s 100 (to seed the random number generator). Expression level was then estimated, as transcripts per million (TPM), using the high-speed quantification tool Kallisto v0.43.1 [60] and default parameters. For datasets comprising single-end reads, we used parameters -l 100 -s 10; estimates of the average fragment length and standard deviation of the fragment length, respectively. Kallisto quantifies expression at the transcript level by building an index of k-mers from a set of reference transcripts and then mapping the RNA-seq reads to it, matching k-mers generated from the reads with the k-mers present in the index. Transcript-level TPM estimates are then summarised to the gene level. A critical aspect of this method is in selecting an appropriate set of reference transcripts for which expression is quantified. An appropriate value of *k* for the index is also required because if *k* is too large relative to read length, there is a higher chance the k-mers of the reads will contain errors (as read quality decreases towards the 3' end of reads [4]). If the reads generate erroneous k-mers, they will not match the k-mers of the index. We used a value of *k* = 21, which lies – approximately – between half the length of the shortest read and a third the length of the longest read.

As a reference transcriptome, we obtained from Ensembl v89 the set of GalGal5 protein-coding transcripts, parsing the batch release (ftp://ftp.ensembl.org/pub/release-89/fasta/gallus_gallus/cds/Gallus_gallus.Gallus_gallus-5.0.cds.all.fa.gz, accessed 21st June 2017) to retain only those transcripts with the 'protein-coding' biotype (n=28,768 transcripts, representing 10,846 genes). To this was added the CDS of 28,466 NCBI mRNA RefSeqs that had neither been assigned Ensembl transcript IDs, nor whose sequence was already present in the Ensembl release (under any other identifier). To reduce the likelihood of spurious read mapping, CDS < 300 bp were excluded from analysis. Erroneous expression level estimates are more likely when fewer possible reads can be derived from a gene, i.e. if the CDS is short [59]. While this approach arguably improves accuracy, it unavoidably excludes certain families, for instance the gallinacins [114], antimicrobial peptides known for their short chain lengths [115].

Although the Ensembl and NCBI sets of transcripts overlap, there are many unique entries in each. For example, RefSeqs XM_015294055 and XM_015294059 are both predicted transcripts of the macrophage-marker gene *CD163* [116], although Ensembl refers to this gene only by the numerical ID '418303'. RefSeq records beginning 'XM' are produced by the NCBI genome annotation pipeline and can lack transcript or protein homology support; by contrast, 'NM' records are validated [117]. Consequently, neither of the *CD163* RefSeqs are assigned Ensembl transcript IDs, and so they are excluded from the Ensembl batch release.

The RefSeq mRNA set also includes predictions of novel transcript sequences for existing Ensembl genes. For instance, the chicken *BF1* gene (classical MHC class 1; Ensembl gene ID ENSGALG00000033932) has 7 transcripts (Ensembl v89), encoding proteins of length 228, 323, 345, 346, 350, 354 and 360 amino acids (aa). However, *BF1* has only 3 associated mRNA RefSeqs, 1 validated and 2 predicted: NM_001044683, XM_015294995, and XM_015294996. These RefSeqs do not necessarily encode different proteins to those present in Ensembl – rather, the RefSeq mRNAs incorporate untranslated regions (UTRs) and so can encapsulate Ensembl CDS. For instance, the validated RefSeq mRNA NM_001044683 encodes the same 360aa protein as Ensembl CDS ENSGALT00000066783 (i.e. the same transcript model is independently available from both resources), but the RefSeq nucleotide sequence extends 17 bases upstream (the 5' UTR) and 146 bases downstream (the 3' UTR) of the coding ORF. By contrast, XM_015294995 encodes a putative 356aa peptide (XP_015150481) and XM_015294996 a 349aa peptide (XP_015150482), neither of which are available from Ensembl. As the XM_015294996 mRNA – an automated prediction – fully incorporates ENSGALT00000086848 (the CDS encoding the 228aa *BF1* protein), we considered the sequence better supported by the Ensembl model, as Ensembl takes a conservative approach to annotation [118], and the predicted peptide spurious. By contrast, the XM_015294995 mRNA does not contain any existing Ensembl CDS and so encodes a protein absent from Ensembl.

Overall, we retained RefSeq 'XM' mRNAs only if they can be assigned to a gene not yet present in the Ensembl annotation, or, if that gene is present, they do not incorporate a CDS from any of that gene's Ensembl transcript models. UTRs were trimmed from each RefSeq mRNA by excluding all sequence outside the longest ORF. This combined set of Ensembl and RefSeq transcripts constitutes a standardised RNA space against which expression can be quantified, as in [59].

After quantifying expression with this initial transcriptome, a revised transcriptome was created, excluding those transcripts whose average TPM was < 1 in all tissues (Table S5), or which were only detectable in one tissue (as these may be artefacts of differential sequencing depth). Tissues whose distribution of TPM estimates does not comply with Zipf's law (see below) were not counted. The revised transcriptome contains 28,276 Ensembl transcripts (representing 10,826 Ensembl genes) and 26,694 NCBI transcripts (which account for only 4665 existing Ensembl genes).

Compliance of RNA-seq datasets with Zipf's law

In a correctly prepared RNA-seq dataset, a minority of reads will produce the majority of reads and so its distribution of gene-level TPM estimates should comply, to a reasonable approximation, with Zipf's law (which states that the probability of an observation is inversely proportional to its rank). A custom Perl script was used to identify, per sample, the number of unique TPM values and the number of genes with a TPM at or exceeding this level. After excluding, for robustness, data from the first and last order of magnitude (as in [119]) and all values of TPM < 5 (which have a higher likelihood of transcriptional noise), the data was log-transformed and a linear regression model fitted using R v3.2.0

[120]. Samples whose exponents deviated too greatly from -1 (by ± 20%, i.e. if the exponent is < -0.8 or > -1.2) were considered erroneous.

Tissue specificity

For each gene, we calculated a preferential expression measure (PEM) in a manner similar to [65]. PEM relates the average expression of that gene in a given tissue to the average expression of that gene in all tissues. For each gene *i*, then for tissue t_i, PEM(t_i) = S-A, where S = expression of gene *i* in tissue t_i, and A = arithmetic mean expression of gene *i* across the set of all tissues. Prior to calculation, all TPM values < 1 were considered to be 1, and a $\log_2$-transformation applied. This is to ensure that genes with expression indistinguishable from noise (TPM < 1) will have a PEM of 0. Each gene will have a distribution of PEM values, one for each tissue in the meta-datasets. Genes with higher PEM values for a given tissue are more tissue-specific in their expression profile.

Gene Ontology (GO) term enrichment

GO term enrichment was assessed using the R package topGO [121], which utilises the 'weight' algorithm to account for the nested structure of the GO tree [122]. topGO requires a reference set of GO terms, which was built manually from the GalGal5 set (obtained from Ensembl BioMart v89 [123]) and filtered to remove those terms with evidence codes NAS (non-traceable author statement) or ND (no biological data available), and those assigned to fewer than 10 genes in total. Significantly enriched GO terms ($p < 0.05$) are reported only if the observed number per tissue exceeds the expected by 2-fold or greater.

Gene annotation

Unannotated genes in GalGal5 – those with only an Ensembl placeholder ID, rather than an HGNC name – are annotated by reference to the NCBI non-redundant (nr) peptide database v77 [124], with each annotation assigned a quality category of 1 to 8 (highest to lowest quality, respectively), as previously described [27]. For each unannotated gene, we took the longest encoded peptide and obtained the set of blastp alignments [125] against NCBI nr, at a scoring threshold of $p <= 1e^{-25}$. These alignments are a set of possible gene descriptions, of which only one can be selected as the annotation of that gene. The lowest quality category, 8, is the blastp hit with the lowest E-value. All subsequent quality categories require higher-quality hits, which: (a) have a % identity within the aligned region of >= 90%, (b) have an alignment length >= 90% of the length of the query protein, (c) have an alignment length >= 50 amino acids, (d) have no gaps, and (e) are not to a protein labelled either ‘low quality’, ‘hypothetical’, ‘unnamed’, ‘uncharacterized’ or ‘putative’, or otherwise having a third-party annotation (as these can be by inference and not experiment). Quality category 7 is the best-scoring (i.e. lowest E-value) of these higher quality hits. Category 6 is as above, but with at least one identifiable hit to the human proteome. Category 5 requires that the set of alignments span at least 4 different genera (excluding *Gallus*). At this point, if >= 75% of the alignments have the same description, the gene is named for the associated HGNC name (according to ftp://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/locus_types/gene_with_protein_product.txt, downloaded 24th August 2016). However, as NCBI nr aggregates multiple sources of data, gene descriptions have numerous synonyms and so it is not always possible to automatically assign an HGNC symbol. The highest quality categories, 1 to 4, not only meet the above criteria but have degrees of reciprocal % identity to the human proteome. The highest quality category, 1, is if there is also a near-perfect match to an existing, related, peptide (alignment length >= 90% of the length of a human protein). Other quality categories, in descending order, are: 2 (alignment length >= 75% of the length of a human protein), 3 (>= 50%), and 4 (< 50%). Human protein sequences were obtained from genebuild GRCh38.p8 (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000001405.34_GRCh38.p8/GCF_000001405.34_GRCh38.p8_protein.faa.gz, downloaded 30th August 2016).

Network analysis

Network analysis was performed using Miru (Kajeka Ltd, Edinburgh, UK), a commercial version of BioLayout *Express*^{3D} [67, 68]. Miru determines the similarities between individual expression profiles by building a correlation matrix for both gene-to-gene and sample-to-sample comparisons. This matrix is then filtered to remove all correlations below a certain threshold (for the gene-to-gene comparison in the RNA-seq atlas, Pearson's $r < 0.8$). A network graph is constructed by connecting nodes (genes) with edges (correlations above the threshold), and its local structure interpreted by applying the Markov clustering (MCL) algorithm [69] at an inflation value (which determines cluster granularity) of 2.2.

Protein-protein interactions

Protein-protein interaction data was obtained from the IID (Integrated Interactions Database) version 2017-04 (http://iid.ophid.utoronto.ca/iid, accessed 25th July 2017) [126], a resource which combines computationally predicted PPIs with experimentally determined PPIs drawn from multiple databases. These include BIND (Biomolecular Interaction Network Database) [127], BioGRID (Biological General Repository for Interaction Datasets) [128], DIP (Database of Interacting Proteins) [129], HPRD (Human Protein Reference Database) [130], IntAct [131], I2D (Interologous Interaction Database) [132], InnateDB [133] and MINT (Molecular Interaction Database) [134]. The format of the PPI data is as a list of UniProt IDs, with one of three evidence types for the interaction: 'exp' (experimentally determined in this species), 'pred' (an *in silico* prediction from one of four previous studies [135-138]) and 'ortho' (predicted by mapping experimentally determined PPIs from another species to orthologous protein pairs in this species). As chicken PPI data is unavailable, we obtained human PPIs from the IID, and considered only those PPIs that (a) involve genes that each have a one-to-one orthologue to the chicken with an orthology confidence score of 1 (using data from Ensembl Compara [139], a score of 1 indicates compliance with the gene tree), a reciprocal % gene identity of >= 75%, a whole genome alignment score of >= 75%, and a gene order conservation score of >= 75% (indicating a high degree of contiguity around the gene of interest), (b) have UniProt IDs that are unambiguously assigned to only one human gene ID (and thereby only one orthologous chicken gene ID), and (c) have PPI evidence type 'exp' or 'pred'.

Availability of datasets

To test whether down-sampling quantitatively alters the expression profile of an RNA-seq dataset, we randomly down-sampled each of the 18 BMDM datasets (+/- LPS) to 10 million reads 100 times, using seqtk seeded with a random integer between 0 and 10,000. These sets of expression estimates are available as Dataset S1, hosted on the University of Edinburgh DataShare portal (http://dx.doi.org/10.7488/ds/2137). The meta-atlas of chicken gene expression is available in full as Table S6 and via the cross-species annotation portal BioGPS (http://biogps.org/dataset/BDS_00031/chicken-atlas/). To compare genes between species and to visualise expression profiles, BioGPS requires that each gene have an Entrez ID, although this is not the case for all genes in GalGal5. The expression profiles of those genes without Entrez IDs can be found in Table S6.

Analysis of chicken developmental samples

The expression data derived from CAGE [53] were obtained from http://fantom.gsc.riken.jp/5/suppl/Lizio_et_al_2017/data; the expression file is named galGal5.cage_peak_tpm.osc.txt.gz and the annotation file galGal5.cage_peak_ann.txt. The annotation and expression files were emerged based on chromosomal location of the promoter. All promoters where no sample exceeded 10 tags per million (tagsPM) were excluded from the analysis. The expression data were then entered into Miru (as described above), using a correlation coefficient threshold of 0.75. 22,839 nodes joined by 5,035,102 edges were entered into the analysis and clustered with an MCL inflation value of 2.2, resulting in 132 clusters of at least 10 nodes.

DECLARATIONS

Funding

This work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) project grant (BB/M011925/1) to DAH and institute strategic program grants ‘Farm Animal Genomics’ (BBS/E/D/20211550) and ‘Transcriptomes, Networks and Systems’ (BBS/E/D/20211552). RNA-seq of the *Campylobacter*-infected chickens was supported by funding from the Rural and Environmental Science and Analytical Sciences Division of the Scottish Government, and from a BBSRC LINK project with Aviagen Ltd. (BB/J006815/1). Edinburgh Genomics is partly supported through core grants from the BBSRC (BB/J004243/1), Natural Environmental Research Council (R8/H10/56), and Medical Research Council (MR/K001744/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The datasets generated during this study are available in the European Nucleotide Archive under accessions PRJEB22373 and PRJEB22580. All data analysed during this study are included in this published article (and its supplementary information files). The atlas of chicken gene expression is also available via the cross-species annotation portal BioGPS (http://biogps.org/dataset/BDS_00031/chicken-atlas/).

Authors’ contributions

DAH coordinated the study. LF, AJM, and JOD performed macrophage cell culture and RNA extraction. AP, JS and MS, funded, generated and provided RNA-seq data from the caecal tonsils of *Campylobacter*-infected birds. CW and CA prepared data for visualisation with BioGPS. SJB performed all bioinformatic analyses with the exception of the CAGE analysis. KMS performed the CAGE analysis. SJB and DAH wrote the manuscript. All authors read, contributed to, and approved the final manuscript.

Competing interests

The authors declare they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Approval was obtained from The Roslin Institute's and the University of Edinburgh's Protocols and Ethics Committees. All animal work was carried out under the regulations of the Animals (Scientific Procedures) Act 1986.

REFERENCES

1. He F, Yoo S, Wang D, Kumari S, Gerstein M, Ware D, Maslov S: **Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis**. *Plant J* 2016, **86**(6):472-480.
2. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA: **An expression atlas of human primary cells: inference of gene function from coexpression networks**. *BMC Genomics* 2013, **14**(1):1-13.
3. Doig TN, Hume DA, Theocharidis T, Goodlad JR, Gregory CD, Freeman TC: **Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment**. *BMC Genomics* 2013, **14**:469.

4. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X *et al*: **A survey of best practices for RNA-seq data analysis**. *Genome Biology* 2016, **17**:13.

5. van Dijk EL, Jaszczyszyn Y, Thermes C: **Library preparation methods for next-generation sequencing: Tone down the bias**. *Experimental Cell Research* 2014, **322**(1):12-20.

6. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA: **The impact of read length on quantification of differentially expressed genes and splice junction detection**. *Genome Biology* 2015, **16**(1):131.

7. Sinha R, Lenser T, Jahn N, Gausmann U, Friedel S, Szafranski K, Huse K, Rosenstiel P, Hampe J, Schuster S *et al*: **TassDB2 - A comprehensive database of subtle alternative splicing events**. *BMC Bioinformatics* 2010, **11**(1):1-7.

8. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, von Schack D, Zhang B: **Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap**. *BMC Genomics* 2015, **16**(1):675.

9. Sultan M, Amstislavskiy V, Risch T, Schuette M, Dökel S, Ralser M, Balzereit D, Lehrach H, Yaspo M-L: **Influence of RNA extraction methods and library selection schemes on RNA-seq data**. *BMC Genomics* 2014, **15**(1):1-13.

10. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers**. *BMC Genomics* 2012, **13**(1):341.

11. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y *et al*: **Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study**. *Nat Biotech* 2014, **32**(9):915-925.

12. González E, Joly S: **Impact of RNA-seq attributes on false positive rates in differential expression analysis of de novo assembled transcriptomes**. *BMC Research Notes* 2013, **6**(1):503.

13. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T *et al*: **Comparative analysis of RNA sequencing methods for degraded or low-input samples**. *Nat Methods* 2013, **10**(7):623-629.

14. Esteve-Codina A, Arpi O, Martinez-García M, Pineda E, Mallo M, Gut M, Carrato C, Rovira A, Lopez R, Tortosa A *et al*: **A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples**. *PloS one* 2017, **12**(1):e0170632.

15. Gallego Romero I, Pai AA, Tung J, Gilad Y: **RNA-seq: impact of RNA degradation on transcript quantification**. *BMC Biology* 2014, **12**(1):42.

16. Seear PJ, Sweeney GE: **Stability of RNA isolated from post-mortem tissues of Atlantic salmon (Salmo salar L.)**. *Fish Physiol Biochem* 2008, **34**(1):19-24.

17. Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beißbarth T, Gaedcke J: **Impact of RNA degradation on gene expression profiling**. *BMC Medical Genomics* 2010, **3**(1):36.

18. Johnson BR, Atallah J, Plachetzki DC: **The importance of tissue specificity for RNA-seq: highlighting the errors of composite structure extractions**. *BMC Genomics* 2013, **14**(1):586.

19. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M *et al*: **The evolution of gene expression levels in mammalian organs**. *Nature* 2011, **478**(7369):343-348.

20. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC *et al*: **Comparison of the transcriptional landscapes between human and mouse tissues**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(48):17224-17229.

21. Merkin J, Russell C, Chen P, Burge CB: **Evolutionary dynamics of gene and isoform regulation in Mammalian tissues**. *Science (New York, NY)* 2012, **338**(6114):1593-1599.

22. Sudmant PH, Alexis MS, Burge CB: **Meta-analysis of RNA-seq expression data across species, tissues and studies**. *Genome Biology* 2015, **16**(1):287.

23. Oliver S: **Guilt-by-association goes global**. *Nature* 2000, **403**(6770):601-603.

24. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T *et al*: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**(7493):455-461.

25. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M *et al*: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**(7493):462-470.

26. Freeman TC, Ivens A, Baillie JK, Beraldi D, Barnett MW, Dorward D, Downing A, Fairbairn L, Kapetanovic R, Raza S *et al*: **A gene expression atlas of the domestic pig**. *BMC Biology* 2012, **10**(1):1-22.

27. Clark EL, Bush SJ, McCulloch MEB, Farquhar IL, Young R, Lefevre L, Pridans C, Tsang HG, Wu C, Afrasiabi C *et al*: **A high resolution atlas of gene expression in the domestic sheep (Ovis aries)**. *PLOS Genetics* 2017, **13**(9):e1006997.

28. Hume DA, Summers KM, Raza S, Baillie JK, Freeman TC: **Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations**. *Genomics* 2010, **95**(6):328-338.

29. Carpanini SM, Wishart TM, Gillingwater TH, Manson JC, Summers KM: **Analysis of gene expression in the nervous system identifies key genes and novel candidates for health and disease**. *Neurogenetics* 2017, **18**(2):81-95.

30. Eising E, Huisman SM, Mahfouz A, Vijfhuizen LS, Anttila V, Winsvold BS, Kurth T, Ikram MA, Freilinger T, Kaprio J *et al*: **Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas**. *Human genetics* 2016, **135**(4):425-439.

31. Stern CD: **The chick; a great model system becomes even greater**. *Dev Cell* 2005, **8**(1):9-17.

32. Intarapat S, Stern CD: **Chick stem cells: current progress and future prospects**. *Stem Cell Res* 2013, **11**(3):1378-1392.

33. Balic A, Garcia-Morales C, Vervelde L, Gilhooley H, Sherman A, Garceau V, Gutowska MW, Burt DW, Kaiser P, Hume DA *et al*: **Visualisation of chicken macrophages using transgenic reporter genes: insights into the development of the avian macrophage lineage**. *Development* 2014, **141**(16):3255-3265.

34. Han JY, Lee HJ: **Genome Editing Mediated by Primordial Germ Cell in Chicken**. *Methods in molecular biology (Clifton, NJ)* 2017, **1630**:153-163.

35. Woodcock ME, Idoko-Akoh A, McGrew MJ: **Gene editing in birds takes flight**. *Mamm Genome* 2017.

36. Taylor L, Carlson DF, Nandi S, Sherman A, Fahrenkrug SC, McGrew MJ: **Efficient TALEN-mediated gene targeting of chicken primordial germ cells**. *Development* 2017, **144**(5):928-934.

37. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P *et al*: **Development of a high density 600K SNP genotyping array for chicken**. *BMC Genomics* 2013, **14**:59.

38. Cheng HH, Kaiser P, Lamont SJ: **Integrated genomic approaches to enhance genetic resistance in chickens**. *Annu Rev Anim Biosci* 2013, **1**:239-260.

39. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW: **Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human**. *BMC Genomics* 2017, **18**(1):323.

40. Kapetanovic R, Fairbairn L, Beraldi D, Sester DP, Archibald AL, Tuggle CK, Hume DA: **Pig Bone Marrow-Derived Macrophages Resemble Human Macrophages in Their Response to Bacterial Lipopolysaccharide**. *The Journal of Immunology* 2012, **188**(7):3382-3394.

41. Smith J, Burt DW, the Avian RC: **The Avian RNAseq Consortium: a community effort to annotate the chicken genome**. *Cytogenetic and genome research* 2015, **145**(2):78-179.

42. Langouet-Astrie CJ, Meinsen AL, Grunwald ER, Turner SD, Enke RA: **RNA sequencing analysis of the developing chicken retina**. *Scientific Data* 2016, **3**:160117.

43. Piórkowska K, Żukowski K, Nowak J, Połtowicz K, Ropka-Molik K, Gurgul A: **Genome-wide RNA-Seq analysis of breast muscles of two broiler chicken groups differing in shear force**. *Animal Genetics* 2016, **47**(1):68-80.

44. Wu P, Ng CS, Yan J, Lai Y-C, Chen C-K, Lai Y-T, Wu S-M, Chen J-J, Luo W, Widelitz RB *et al*: **Topographical mapping of α- and β-keratins on developing chicken skin integuments: Functional interaction and evolutionary perspectives**. *Proceedings of the National Academy of Sciences* 2015, **112**(49):E6770-E6779.

45. Resnyk CW, Chen C, Huang H, Wu CH, Simon J, Le Bihan-Duval E, Duclos MJ, Cogburn LA: **RNA-Seq Analysis of Abdominal Fat in Genetically Fat and Lean Chickens Highlights a Divergence in Expression of Genes Controlling Adiposity, Hemostasis, and Lipid Metabolism**. *PloS one* 2015, **10**(10):e0139549.

46. Shen X, Bai X, Xu J, Zhou M, Xu H, Nie Q, Lu X, Zhang X: **Transcriptome sequencing reveals genetic mechanisms underlying the transition between the laying and brooding phases and gene expression changes associated with divergent reproductive phenotypes in chickens**. *Molecular Biology Reports* 2016, **43**(9):977-989.

47. Pritchett EM, Lamont SJ, Schmidt CJ: **Transcriptomic changes throughout post-hatch development in Gallus gallus pituitary**. *Journal of Molecular Endocrinology* 2016, **58**(1):43-55.

48. Van Goor A, Ashwell CM, Persia ME, Rothschild MF, Schmidt CJ, Lamont SJ: **Unique genetic responses revealed in RNA-seq of the spleen of chickens stimulated with lipopolysaccharide and short-term heat**. *PloS one* 2017, **12**(2):e0171414.

49. Wang Y, Lupiani B, Reddy SM, Lamont SJ, Zhou H: **RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens**. *Poultry Science* 2014, **93**(2):485-493.

50. Li Z, Ouyang H, Zheng M, Cai B, Han P, Abdalla BA, Nie Q, Zhang X: **Integrated Analysis of Long Non-coding RNAs (LncRNAs) and mRNA Expression Profiles Reveals the Potential Role of LncRNAs in Skeletal Muscle Development of the Chicken**. *Frontiers in Physiology* 2016, **7**:687.

51. Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, Désert C, Boutin M, Jehl F, Acloque H *et al*: **Long noncoding RNA repertoire in chicken liver and adipose tissue**. *Genetics, selection, evolution : GSE* 2017, **49**:6.

52. Roux P-F, Frésard L, Boutin M, Leroux S, Klopp C, Djari A, Esquerré D, Martin PGP, Zerjal T, Gourichon D *et al*: **The Extent of mRNA Editing Is Limited in Chicken Liver and Adipose, but Impacted by Tissular Context, Genotype, Age, and Feeding as Exemplified with a Conserved Edited Site in COG3**. *G3: Genes|Genomes|Genetics* 2016, **6**(2):321-335.

53. Lizio M, Deviatiiarov R, Nagai H, Galan L, Arner E, Itoh M, Lassmann T, Kasukawa T, Hasegawa A, Ros MA *et al*: **Systematic analysis of transcription start sites in avian development**. *PLoS biology* 2017, **15**(9):e2002887.

54. Deviatiiarov R, Lizio M, Gusev O: **Application of a CAGE Method to an Avian Development Study**. *Methods in molecular biology (Clifton, NJ)* 2017, **1650**:101-109.

55. Zeferino CP, Wells KD, Moura ASAMT, Rottinghaus GE, Ledoux DR: **Changes in renal gene expression associated with induced ochratoxicosis in chickens: activation and deactivation of transcripts after varying durations of exposure**. *Poultry Science* 2017, **96**(6):1855-1865.

56. Han D, Zhang Y, Chen J, Hua G, Li J, Deng X, Deng X: **Transcriptome analyses of differential gene expression in the bursa of Fabricius between Silky Fowl and White Leghorn**. *Scientific Reports* 2017, **7**:45959.

57. Liu X-d, Zhang F, Shan H, Wang S-b, Chen P-Y: **mRNA expression in different developmental stages of the chicken bursa of Fabricius**. *Poultry Science* 2016, **95**(8):1787-1794.

58. Zhu G, Mao Y, Zhou W, Jiang Y: **Dynamic Changes in the Follicular Transcriptome and Promoter DNA Methylation Pattern of Steroidogenic Genes in Chicken Follicles throughout the Ovulation Cycle**. *PloS one* 2016, **10**(12):e0146028.

59. Bush SJ, McCulloch MEB, Summers KM, Hume DA, Clark EL: **Integration of quantitated expression estimates from polyA-selected and rRNA-depleted RNA-seq libraries**. *BMC Bioinformatics* 2017, **18**(1):301.

60. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nat Biotech* 2016, **34**(5):525-527.

61. Lu T, Costello CM, Croucher PJ, Hasler R, Deuschl G, Schreiber S: **Can Zipf's law be adapted to normalize microarrays?** *BMC Bioinformatics* 2005, **6**:37.

62. Furusawa C, Kaneko K: **Zipf's law in gene expression**. *Phys Rev Lett* 2003, **90**(8):088102.

63. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A: **Differential expression in RNA-seq: A matter of depth**. *Genome Research* 2011, **21**(12):2213-2223.

64. Liu Y, Zhou J, White KP: **RNA-seq differential expression studies: more sequence or more replication?** *Bioinformatics* 2014, **30**(3):301-304.

65. Huminiecki L, Lloyd A, Wolfe K: **Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases**. *BMC Genomics* 2003, **4**(1):31.

66. Glick B: **Historical perspective: the bursa of Fabricius and its influence on B-cell development, past and present**. *Vet Immunol Immunopathol* 1991, **30**(1):3-12.

67. Freeman TC, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock RJ, Freilich S, Thornton J, Enright AJ: **Construction, visualisation, and clustering of**

transcription networks from microarray expression data. *PLoS computational biology* 2007, **3**(10):2032-2042.

68. Theocharidis A, van Dongen S, Enright AJ, Freeman TC: **Network visualization and analysis of gene expression data using BioLayout Express(3D)**. *Nature protocols* 2009, **4**(10):1535-1550.
69. van Dongen S, Abreu-Goodger C: **Using MCL to extract clusters from networks**. *Methods in molecular biology (Clifton, NJ)* 2012, **804**:281-295.
70. Bar-Joseph Z, Siegfried Z, Brandeis M, Brors B, Lu Y, Eils R, Dynlacht BD, Simon I: **Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(3):955-960.
71. Wu D-D, Irwin DM, Zhang Y-P: **Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair**. *BMC evolutionary biology* 2008, **8**(1):241.
72. Eckelhoefer HA, Rajapaksa TE, Wang J, Hamer M, Appleby NC, Ling J, Lo DD: **Claudin-4: Functional Studies Beyond the Tight Junction**. In: *Claudins: Methods and Protocols.* Edited by Turksen K. Totowa, NJ: Humana Press; 2011: 115-128.
73. So A, Thorens B: **Uric acid transport and disease**. *The Journal of Clinical Investigation* 2010, **120**(6):1791-1799.
74. Galvan I, Solano F: **Bird Integumentary Melanins: Biosynthesis, Forms, Function and Evolution**. *Int J Mol Sci* 2016, **17**(4):520.
75. Hume DA, Summers KM, Rehli M: **Transcriptional Regulation and Macrophage Differentiation**. *Microbiol Spectr* 2016, **4**(3).
76. Mass E, Ballesteros I, Farlik M, Halbritter F, Gunther P, Crozet L, Jacome-Galarza CE, Handler K, Klughammer J, Kobayashi Y *et al*: **Specification of tissue-resident macrophages during organogenesis**. *Science (New York, NY)* 2016, **353**(6304).
77. Aziz A, Soucie E, Sarrazin S, Sieweke MH: **MafB/c-Maf deficiency enables self-renewal of differentiated functional macrophages**. *Science (New York, NY)* 2009, **326**(5954):867-871.
78. Hume DA, Mabbott N, Raza S, Freeman TC: **Can DCs be distinguished from macrophages by molecular signatures?** *Nature immunology* 2013, **14**(3):187-189.
79. Joshi A, Pooley C, Freeman TC, Lennartsson A, Babina M, Schmidl C, Geijtenbeek T, Michoel T, Severin J, Itoh M *et al*: **Technical Advance: Transcription factor, promoter, and enhancer utilization in human myeloid cells**. *Journal of leukocyte biology* 2015, **97**(5):985-995.
80. Rodriguez-Manzanet R, Meyers JH, Balasubramanian S, Slavik J, Kassam N, Dardalhon V, Greenfield EA, Anderson AC, Sobel RA, Hafler DA *et al*: **TIM-4 Expressed on APCs Induces T Cell Expansion and Survival**. *The Journal of Immunology* 2008, **180**(7):4706.
81. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG: **Using graph theory to analyze biological networks**. *BioData Mining* 2011, **4**:10-10.
82. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **12**(1):37-46.
83. Tornow S, Mewes HW: **Functional modules by relating protein interaction networks and gene expression**. *Nucleic acids research* 2003, **31**(21):6283-6289.
84. Kovarik P, Stoiber D, Novy M, Decker T: **Stat1 combines signals derived from IFN-gamma and LPS receptors during macrophage activation**. *The EMBO Journal* 1998, **17**(13):3660-3668.

85. Kim JY, Song EH, Lee S, Lim JH, Choi JS, Koh IU, Song J, Kim WH: **The induction of STAT1 gene by activating transcription factor 3 contributes to pancreatic beta-cell apoptosis and its dysfunction in streptozotocin-treated mice**. *Cell Signal* 2010, **22**(11):1669-1680.

86. Celada A, Borras FE, Soler C, Lloberas J, Klemsz M, van Beveren C, McKercher S, Maki RA: **The transcription factor PU.1 is involved in macrophage proliferation**. *J Exp Med* 1996, **184**(1):61-69.

87. Pazdrak K, Justement L, Alam R: **Mechanism of inhibition of eosinophil activation by transforming growth factor-beta. Inhibition of Lyn, MAP, Jak2 kinases and STAT1 nuclear factor**. *J Immunol* 1995, **155**(9):4454-4458.

88. Frühbeck G: **Intracellular signalling pathways activated by leptin**. *Biochemical Journal* 2006, **393**(Pt 1):7-20.

89. Richardson ET, Shukla S, Nagy N, Boom WH, Beck RC, Zhou L, Landreth GE, Harding CV: **ERK Signaling Is Essential for Macrophage Development**. *PloS one* 2015, **10**(10):e0140064.

90. Song MM, Shuai K: **The suppressor of cytokine signaling (SOCS) 1 and SOCS3 but not SOCS2 proteins inhibit interferon-mediated antiviral and antiproliferative activities**. *J Biol Chem* 1998, **273**(52):35056-35062.

91. Su X, Yu Y, Zhong Y, Giannopoulou EG, Hu X, Liu H, Cross JR, Ratsch G, Rice CM, Ivashkiv LB: **Interferon-gamma regulates cellular metabolism and mRNA translation to potentiate macrophage activation**. *Nature immunology* 2015, **16**(8):838-849.

92. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drablos F, Lennartsson A, Ronnerblad M, Hrydziuszko O, Vitezic M *et al*: **Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells**. *Science (New York, NY)* 2015, **347**(6225):1010-1014.

93. Summers KM, Hume DA: **Identification of the macrophage-specific promoter signature in FANTOM5 mouse embryo developmental time course data**. *Journal of leukocyte biology* 2017.

94. Garceau V, Balic A, Garcia-Morales C, Sauter KA, McGrew MJ, Smith J, Vervelde L, Sherman A, Fuller TE, Oliphant T *et al*: **The development and maintenance of the mononuclear phagocyte system of the chick is controlled by signals from the macrophage colony-stimulating factor receptor**. *BMC Biol* 2015, **13**:12.

95. Gautier EL, Shay T, Miller J, Greter M, Jakubzick C, Ivanov S, Helft J, Chow A, Elpek KG, Gordonov S *et al*: **Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages**. *Nature immunology* 2012, **13**(11):1118-1128.

96. Epelman S, Lavine KJ, Randolph GJ: **Origin and functions of tissue macrophages**. *Immunity* 2014, **41**(1):21-35.

97. Chang MK, Raggatt LJ, Alexander KA, Kuliwaba JS, Fazzalari NL, Schroder K, Maylin ER, Ripoll VM, Hume DA, Pettit AR: **Osteal tissue macrophages are intercalated throughout human and mouse bone lining tissues and regulate osteoblast function in vitro and in vivo**. *J Immunol* 2008, **181**(2):1232-1244.

98. Feng R, Desbordes SC, Xie H, Tillo ES, Pixley F, Stanley ER, Graf T: **PU.1 and C/EBPα/β convert fibroblasts into macrophage-like cells**. *Proceedings of the National Academy of Sciences* 2008, **105**(16):6057-6062.

99. Li X, Nair A, Wang S, Wang L: **Quality control of RNA-seq experiments**. *Methods in molecular biology (Clifton, NJ)* 2015, **1269**:137-146.

100. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming**. *Nucleic acids research* 2010, **38**(12):e131-e131.

101. van Gurp TP, McIntyre LM, Verhoeven KJF: **Consistent Errors in First Strand cDNA Due to Random Hexamer Mispriming**. *PloS one* 2013, **8**(12):e85583.

102. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-Content Normalization for RNA-Seq Data**. *BMC Bioinformatics* 2011, **12**(1):480.

103. Reiter JF, Leroux MR: **Genes and molecular pathways underpinning ciliopathies**. *Nat Rev Mol Cell Biol* 2017.

104. Stauber M, Boldt K, Wrede C, Weidemann M, Kellner M, Schuster-Gossler K, Kuhnel MP, Hegermann J, Ueffing M, Gossler A: **1700012B09Rik, a FOXJ1 effector gene active in ciliated tissues of the mouse but not essential for motile ciliogenesis**. *Dev Biol* 2017.

105. Zhou J, Chehab R, Tkalcevic J, Naylor MJ, Harris J, Wilson TJ, Tsao S, Tellis I, Zavarsek S, Xu D *et al*: **Elf5 is essential for early embryogenesis and mammary gland development during pregnancy and lactation**. *EMBO J* 2005, **24**(3):635-644.

106. Kist R, Greally E, Peters H: **Derivation of a mouse model for conditional inactivation of Pax9**. *Genesis* 2007, **45**(7):460-464.

107. Bangs F, Antonio N, Thongnuek P, Welten M, Davey MG, Briscoe J, Tickle C: **Generation of mice with functional inactivation of talpid3, a gene first identified in chicken**. *Development* 2011, **138**(15):3261-3272.

108. Yin Y, Bangs F, Paton IR, Prescott A, James J, Davey MG, Whitley P, Genikhovich G, Technau U, Burt DW *et al*: **The Talpid3 gene (KIAA0586) encodes a centrosomal protein that is essential for primary cilia formation**. *Development* 2009, **136**(4):655-664.

109. Roosing S, Romani M, Isrie M, Rosti RO, Micalizzi A, Musaev D, Mazza T, Al-Gazali L, Altunoglu U, Boltshauser E *et al*: **Mutations in CEP120 cause Joubert syndrome as well as complex ciliopathy phenotypes**. *J Med Genet* 2016, **53**(9):608-615.

110. Wu C, Jin X, Tsueng G, Afrasiabi C, Su AI: **BioGPS: building your own mash-up of gene annotations and expression profiles**. *Nucleic acids research* 2016, **44**(D1):D313-D316.

111. Garcia-Morales C, Nandi S, Zhao D, Sauter KA, Vervelde L, McBride D, Sang HM, Clinton M, Hume DA: **Cell-autonomous sex differences in gene expression in chicken bone marrow-derived macrophages**. *J Immunol* 2015, **194**(5):2338-2344.

112. Psifidi A, Fife M, Howell J, Matika O, van Diemen PM, Kuo R, Smith J, Hocking PM, Salmon N, Jones MA *et al*: **The genomic architecture of resistance to Campylobacter jejuni intestinal colonisation in chickens**. *BMC Genomics* 2016, **17**:293.

113. Kodama Y, Shumway M, Leinonen R: **The Sequence Read Archive: explosive growth of sequencing data**. *Nucleic acids research* 2012, **40**(Database issue):D54-56.

114. Lynn DJ, Higgs R, Gaines S, Tierney J, James T, Lloyd AT, Fares MA, Mulcahy G, O'Farrelly C: **Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken**. *Immunogenetics* 2004, **56**(3):170-177.

115. Le C-F, Gudimella R, Razali R, Manikam R, Sekaran SD: **Transcriptome analysis of Streptococcus pneumoniae treated with the designed antimicrobial peptides, DM3**. *Scientific Reports* 2016, **6**:26828.

116. Fabriek BO, Dijkstra CD, van den Berg TK: **The macrophage scavenger receptor CD163**. *Immunobiology* 2005, **210**(2):153-160.

117. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic acids research* 2016, **44**(D1):D733-745.

118. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl Automatic Gene Annotation System**. *Genome Research* 2004, **14**(5):942-950.

119. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E: **Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data**. *Genome Biol* 2009, **10**(7):R79.

120. **R: A Language and Environment for Statistical Computing** [http://www.R-project.org]

121. **topGO: Enrichment analysis for Gene Ontology** [http://www.bioconductor.org/packages/release/bioc/html/topGO.html]

122. Alexa A, Rahnenführer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure**. *Bioinformatics* 2006, **22**(13):1600-1607.

123. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A *et al*: **Ensembl BioMarts: a hub for data retrieval across taxonomic space**. *Database : the journal of biological databases and curation* 2011, **2011**:bar030.

124. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic acids research* 2005, **33**(Database Issue):D501-D504.

125. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic acids research* 1997, **25**(17):3389-3402.

126. Kotlyar M, Pastrello C, Sheahan N, Jurisica I: **Integrated interactions database: tissue-specific view of the human and model organism interactomes**. *Nucleic acids research* 2016, **44**(D1):D536-541.

127. Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database**. *Nucleic acids research* 2003, **31**(1):248-250.

128. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L *et al*: **The BioGRID interaction database: 2015 update**. *Nucleic acids research* 2015, **43**(Database issue):D470-478.

129. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic acids research* 2004, **32**(Database issue):D449-451.

130. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic acids research* 2009, **37**(Database issue):D767-772.

131. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A *et al*: **IntAct: an open source molecular interaction database**. *Nucleic acids research* 2004, **32**(Database issue):D452-D455.

132. Brown KR, Jurisica I: **Unequal evolutionary conservation of human protein interactions in interologous networks**. *Genome Biol* 2007, **8**(5):R95.

133. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan TH, Shah N *et al*: **InnateDB: facilitating systems-level analyses of the mammalian innate immune response**. *Mol Syst Biol* 2008, **4**:218.

134. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database**. *Nucleic acids research* 2007, **35**(Database issue):D572-D574.

135. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network**. *Nature biotechnology* 2005, **23**(8):951-959.

136. Elefsinioti A, Sarac OS, Hegele A, Plake C, Hubner NC, Poser I, Sarov M, Hyman A, Mann M, Schroeder M *et al*: **Large-scale de novo prediction of physical protein-protein association**. *Molecular & cellular proteomics : MCP* 2011, **10**(11):M111 010629.

137. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T *et al*: **Structure-based prediction of protein-protein interactions on a genome-wide scale**. *Nature* 2012, **490**(7421):556-560.

138. Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, Naranian T, Niu Y, Ding Z, Vafaee F *et al*: **In silico prediction of physical protein interactions and characterization of interactome orphans**. *Nat Methods* 2015, **12**(1):79-84.

139. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates**. *Genome Res* 2009, **19**(2):327-335.

140. Diaz-Perales A, Quesada V, Peinado JR, Ugalde AP, Alvarez J, Suarez MF, Gomis-Ruth FX, Lopez-Otin C: **Identification and characterization of human archaemetzincin-1 and -2, two novel members of a family of metalloproteases widely distributed in Archaea**. *J Biol Chem* 2005, **280**(34):30367-30375.

141. Jiang TX, Tuan TL, Wu P, Widelitz RB, Chuong CM: **From buds to follicles: matrix metalloproteinases in developmental tissue remodeling during feather morphogenesis**. *Differentiation* 2011, **81**(5):307-314.

142. Takeda M, Obara N, Suzuki Y: **Keratin filaments of epithelial and taste-bud cells in the circumvallate papillae of adult and developing mice**. *Cell and Tissue Research* 1990, **260**(1):41-48.

143. Plowman GD, Green JM, McDonald VL, Neubauer MG, Disteche CM, Todaro GJ, Shoyab M: **The amphiregulin gene encodes a novel epidermal growth factor-related protein with tumor-inhibitory activity**. *Mol Cell Biol* 1990, **10**(5):1969-1981.

144. Günzel D, Yu ASL: **Claudins and the Modulation of Tight Junction Permeability**. *Physiological Reviews* 2013, **93**(2):525-569.

145. Quinn LM, Kilpatrick LM, Latham SE, Kalionis B: **Homeobox genes DLX4 and HB24 are expressed in regions of epithelial-mesenchymal cell interaction in the adult human endometrium**. *Mol Hum Reprod* 1998, **4**(5):497-501.

146. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschachler E, Eckhart L: **Immunolocalization of a Histidine-Rich Epidermal Differentiation Protein in the Chicken Supports the Hypothesis of an Evolutionary Developmental Link between the Embryonic Subperiderm and Feather Barbs and Barbules**. *PloS one* 2016, **11**(12):e0167789.

147. Lopes Ricardo J, Johnson James D, Toomey Matthew B, Ferreira Mafalda S, Araujo Pedro M, Melo-Ferreira J, Andersson L, Hill Geoffrey E, Corbo Joseph C, Carneiro

M: **Genetic Basis for Red Coloration in Birds**. *Current Biology* 2016, **26**(11):1427-1434.

148. Strasser B, Mlitz V, Hermann M, Rice RH, Eigenheer RA, Alibardi L, Tschachler E, Eckhart L: **Evolutionary Origin and Diversification of Epidermal Barrier Proteins in Amniotes**. *Molecular Biology and Evolution* 2014, **31**(12):3194-3205.

149. Holmes RS: **Vertebrate patatin-like phospholipase domain-containing protein 4 (PNPLA4) genes and proteins: a gene with a role in retinol metabolism**. *3 Biotech* 2012, **2**(4):277-286.

150. Long AC, Bomser JA, Grzybowski DM, Chandler HL: **All-Trans Retinoic Acid Regulates Cx43 Expression, Gap Junction Communication and Differentiation in Primary Lens Epithelial Cells**. *Current Eye Research* 2010, **35**(8):670-679.

151. Wasmeier C, Romao M, Plowright L, Bennett DC, Raposo G, Seabra MC: **Rab38 and Rab32 control post-Golgi trafficking of melanogenic enzymes**. *J Cell Biol* 2006, **175**(2):271-281.

152. Coppola U, Annona G, D’Aniello S, Ristoratore F: **Rab32 and Rab38 genes in chordate pigmentation: an evolutionary perspective**. *BMC evolutionary biology* 2016, **16**(1):26.

153. Wang F, Feng Y, Li P, Wang K, Feng L, Liu Y-F, Huang H, Guo Y-B, Mao Q-S, Xue W-J: **RASSF10 is an epigenetically inactivated tumor suppressor and independent prognostic factor in hepatocellular carcinoma**. *Oncotarget* 2016, **7**(4):4279-4297.

154. Lee S-A, Belyaeva OV, Kedishvili NY: **Biochemical characterization of human epidermal retinol dehydrogenase 2**. *Chemico-Biological Interactions* 2009, **178**(1):182-187.

155. Johnson NC: **XG: the forgotten blood group system**. *Immunohematology* 2011, **27**(2):68-71.

FIGURES

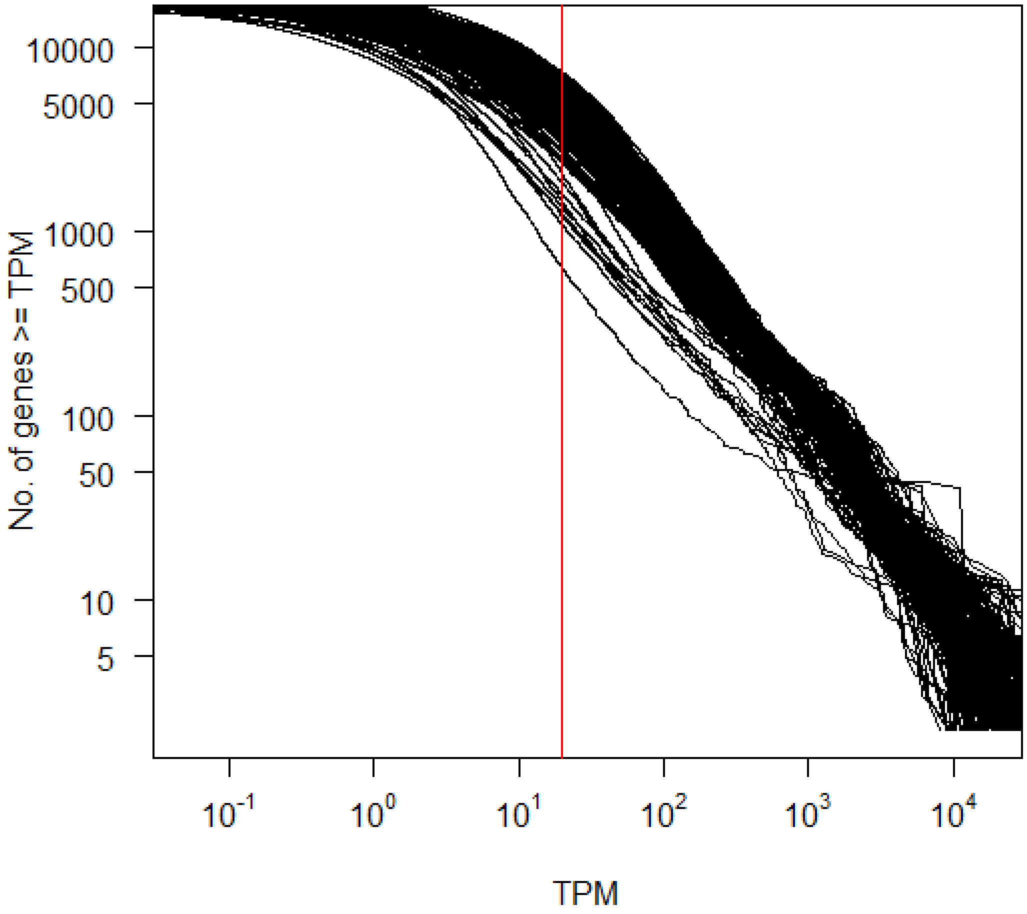

Figure 1. Reverse cumulative distribution of the number of genes that have at least a given TPM. Both axes are logarithmic. Each line represents data from an individual SRA sample ID, quantified using the first iteration Kallisto transcriptome (i.e. a non-redundant set of Ensembl protein-coding CDS plus trimmed RefSeq mRNAs). Samples are not otherwise distinguished as in general, most relationships approximate the same power-law: a minority of genes account for the majority of reads. These relationships are piecewise linear because the capture of lowly expressed genes is noisy, an artefact of random transcriptome sampling. The vertical red line denotes TPM = 5. At higher values of TPM, the

majority of samples have a log-linear relationship. Those that do not are erroneous, and are excluded from subsequent analysis. Exponents of each sample's log-log plot are given in Table S3.

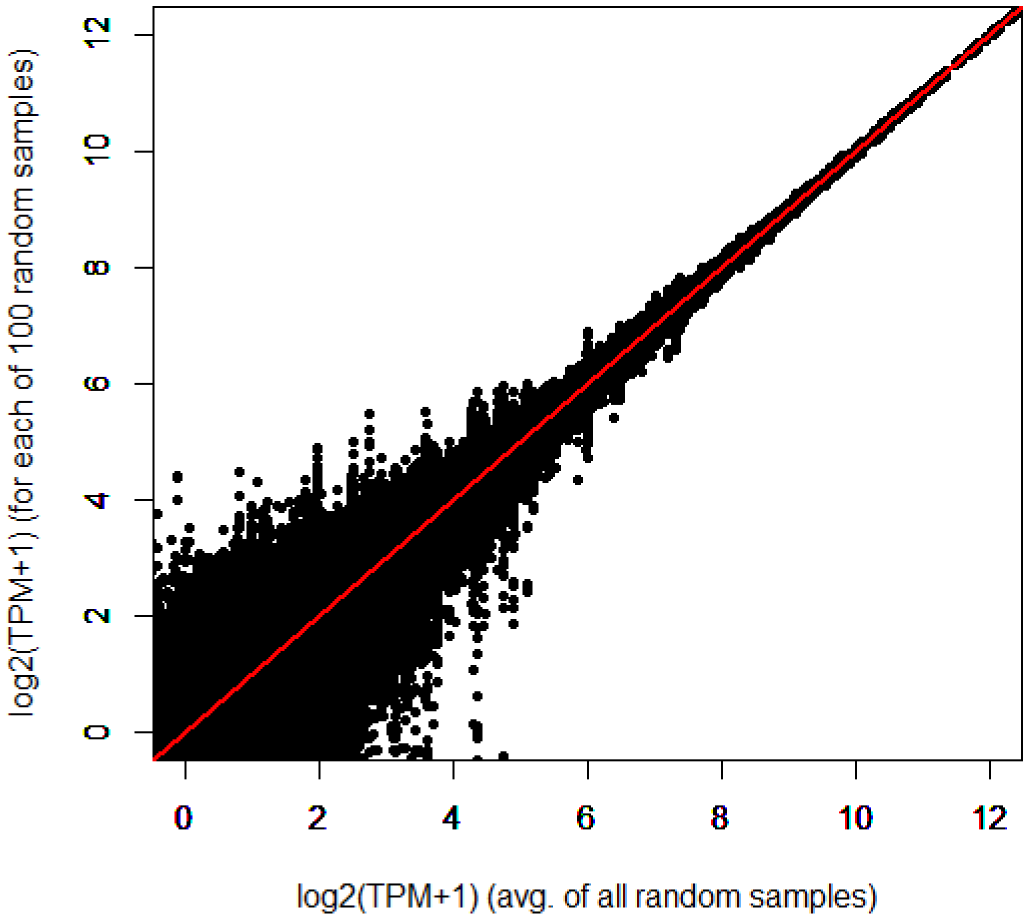

Figure 2. Randomly down-sampling RNA-seq reads has minimal impact on the overall expression profile, primarily affecting expression level estimates of lowly expressed genes. Data shown is from one dataset – unchallenged BMDMs from an adult female broiler (Ross 308) – although with quantitatively similar findings from other samples. The figure plots the average TPM per gene, taken after 100 random samples of 10 million reads, against the TPM obtained in each sample. The line *y* = *x* is shown in red.

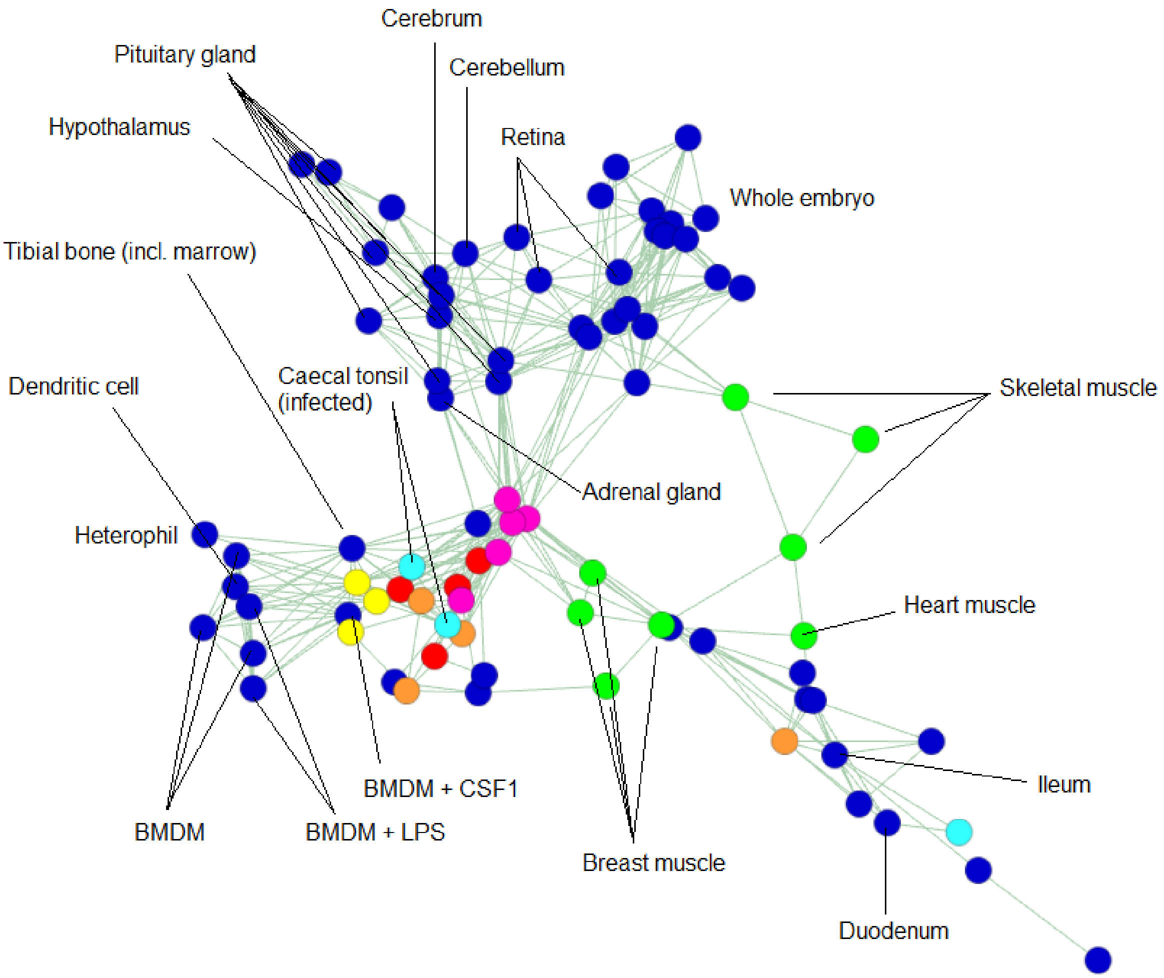

Figure 3. 2D representation of a sample-to-sample network graph, plotting Spearman's correlations between expression profiles. The graph was built using an RNA-seq meta-dataset with each sample distinct by tissue, developmental stage and BioProject of origin, and expression level per gene per sample averaged (where possible) across all replicates of that sample (dataset available as Table S6). Each node (circle) in the graph represents a sample, and each edge (line) a correlation exceeding a threshold (*rho* ≥ 0.82). The graph contains 82 nodes, connected by 243 edges. Selected nodes are labelled. Overall, like tissues tend to correlate more strongly with like, irrespective of BioProject of origin. Certain coloured nodes indicate tissues independently sequenced by multiple BioProjects (listed in Table S2), including liver (red), spleen (yellow), lung (orange), adipose (pink), caecal tonsil (light blue) and muscle (green). There are two notable idiosyncrasies: one of the four lung samples is

comparatively dissimilar to the others of its group, as is one of the three caecal tonsil samples. In the latter case, however, the two most closely correlated caecal tonsil samples are those infected with *Campylobacter*. Consistent with this, these samples cluster more closely with immune cells and tissues. The third caecal tonsil sample belongs to a healthy chicken.

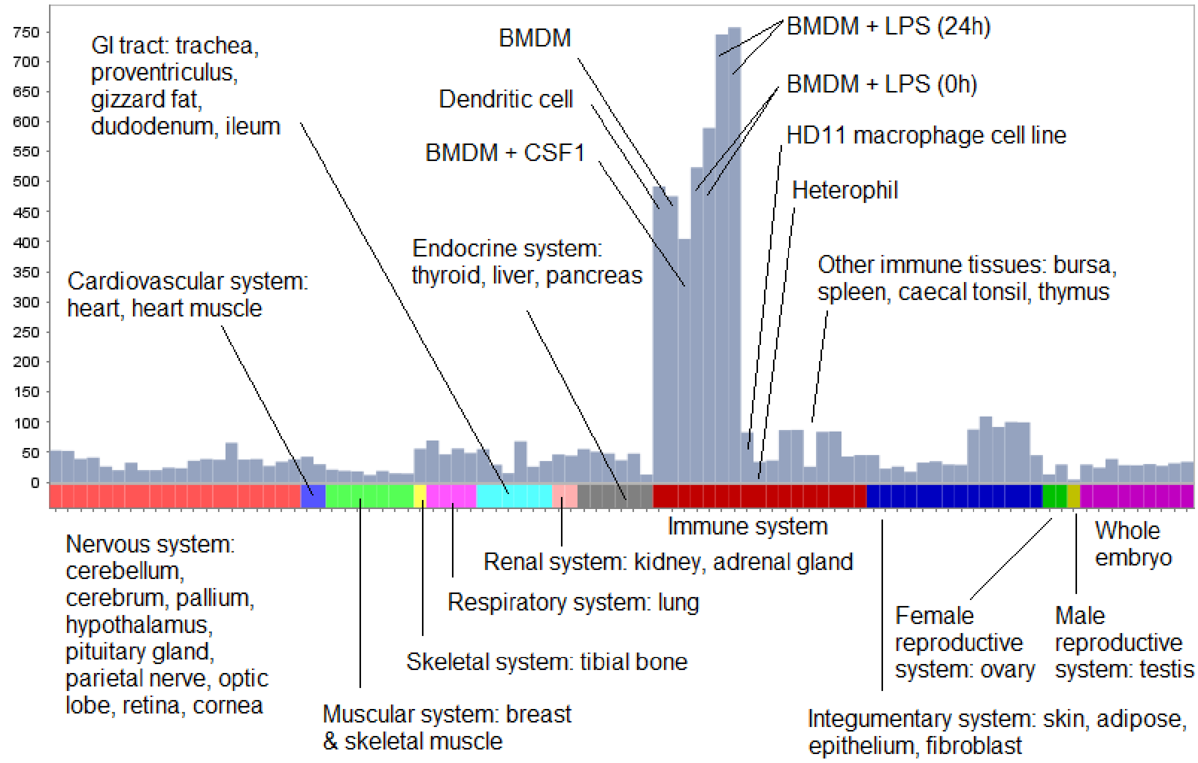

Figure 4. Expression profile of the macrophage-specific cluster 4. Histogram shows the average expression level of the 458 genes in the cluster, where expression level per gene is calculated as the median TPM across all replicates, per BioProject, per tissue. The expression level dataset is available as Table S6.

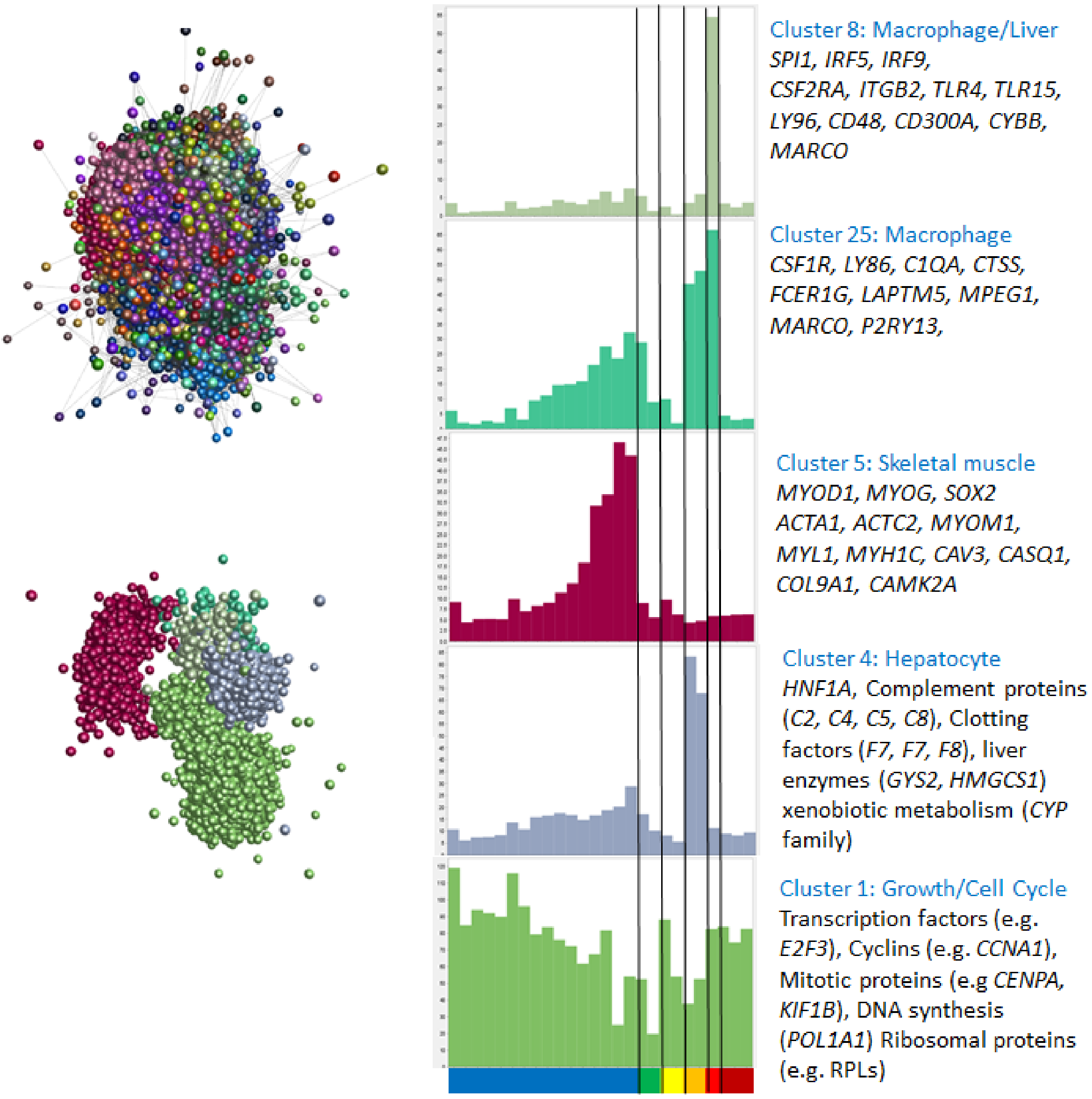

Figure 5. The panel on the left shows the clustered nodes for the main element in the layout graph (upper section). Nodes allocated to the same cluster are the same colour. The panel on the right shows the average expression profiles for five clusters highlighting the different phases of chick embryo development, and key genes for each cluster are shown in the boxes. The layout of these clusters within the main element is shown in the lower part of the left panel. Node colour matches the colour of the bars on the histograms. The X-axis shows the different samples (blue – embryo developmental time course from 1.5 hours to day 20 after fertilisation (HH45); green – extraembryonic tissues; yellow – limb buds; orange – hepatocytes; red – bone marrow derived mesenchymal stem cells; dark

red – aortic smooth muscle cells. Full detail of the samples can be found in Lizio, *et al.* [53]. Y axis shows average TPM for TSS in the cluster for each sample.

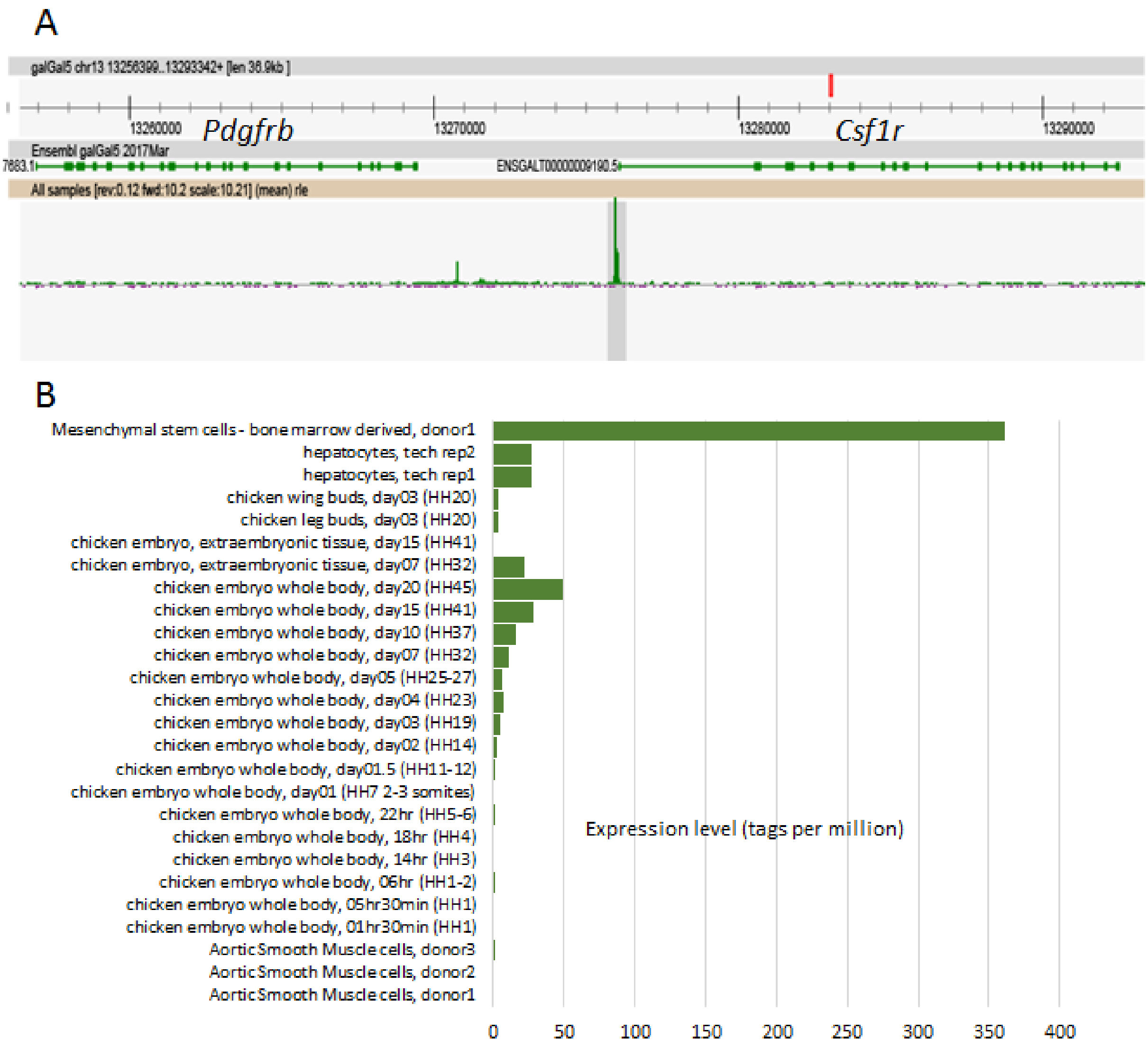

Figure 6. ZENBU (http://fantom.gsc.riken.jp/zenbu/) view of the chicken *CSF1R* locus, identifying the transcription start site downstream of the *PDGFRB* locus (A), and the time course of appearance of *CSF1R* transcripts in the embryo and their expression in isolated cells (B).

TABLES

Table 1. Genes in cluster 14 with known function.

Gene symbol	Gene name	Protein function	References
AMZ1	archaemetzincin 1	metalloprotease, possibly involved in tissue remodelling to form feather follicles	[140, 141]
ANKK1 (PKK2)	ankyrin repeat and kinase domain containing 1	interacts with keratin filaments	[142]
AREG	amphiregulin	epithelial growth factor	[143]
CLDN9	claudin 9	tight junction membrane protein found in all epithelia	[144]
DLX4	homeobox protein DLX4	homeobox protein that regulates epithelial-mesenchymal interactions	[145]
EDMTF4 EDMTFH	epidermal differentiation protein starting with MTF motif 4 epidermal differentiation protein starting with MTF and rich in histidine	markers of the feather barbule and members of the epidermal differentiation complex; this has a role in integumentary development, including feather pigmentation	[146-148]
FK21	feather keratin 21	feather keratins	

FK27	feather keratin 27		
PNPLA4	patatin-like phospholipase domain-containing protein 4	enzyme with a role in retinol metabolism (retinol and related compounds regulate epithelial cell growth and differentiation)	[149, 150]
RAB38	Ras-related protein RAB38	GTPase involved in melanosome biogenesis and epithelial pigmentation	[151, 152]
RASSF10	Ras association domain family member 10	tumour suppressor that mediates the epithelial-mesenchymal transition	[153]
SDR16C5 (RDH-E2)	epidermal retinol dehydrogenase 2	overexpressed in psoriatic human skin	[154]
XG	Xg blood group	blood group antigen	[155]

SUPPLEMENTAL MATERIAL

Dataset S1. Expression level estimates generated after randomly down-sampling the BMDM (+/- LPS) datasets to 10 million reads 100 times.

Table S1. Data sources for creating an RNA-seq meta-atlas.

Table S2. Independent datasets sequencing the same tissue/cell type.

Table S3. Exponents of the log-log plots after plotting the reverse cumulative distribution of TPM per gene on a log-log scale.

Table S4. Number of genes with detectable expression, per tissue, after the first iteration of Kallisto.

Table S5. Transcripts not detectably expressed (at > 1 TPM) in any tissue, after the first iteration of Kallisto.

Table S6. Chicken RNA-seq meta-dataset, after the second (and final) iteration of Kallisto.

Table S7. Proportion of RNA-seq reads retained by down-sampling the LPS-stimulated BMDM datasets.

Table S8. Number of detectably expressed genes after randomly down-sampling the LPS-stimulated BMDM datasets.

Table S9. Range of expression estimates, and absolute difference between largest and smallest estimate, after randomly down-sampling the LPS-stimulated BMDM datasets.

Table S10. GO term enrichment for those subsets of genes whose highest PEM is for a given tissue.

Table S11. All-against-all correlation matrix for each tissue in the meta-dataset.

Table S12. Tissues whose expression vectors are most strongly correlated with each other.

Table S13. Clusters of co-expressed genes (obtained via network analysis of the RNA-seq meta-dataset), including candidate gene names for unannotated GalGal5 protein-coding genes.

Table S14. Proportion of genes in each co-expression cluster whose highest PEM is for a given tissue.

Table S15. GO term enrichment for co-expression clusters containing >= 100 genes.

Table S16. Correlation of expression profiles for genes with a known protein-protein interaction.

Table S17. Clusters of co-expressed CAGE tags, obtained via network analysis of the Lizio, *et al.* dataset [53].

Table S18. Comparison of co-expression clusters between the RNA-seq atlas and the Lizio, *et al.* CAGE dataset [53].