RESEARCH

Open Access

Cross-species single-cell analysis reveals divergence and conservation of peripheral blood mononuclear cells

Siyu Zhang[1,2†], Xiang Fang[3†], Mengyang Chang[4†], Ming Zheng[1,2], Lijin Guo[1,2], Yibin Xu[1,2], Jingting Shu[5*], Qinghua Nie[1,2*] and Zhenhui Li[1,2*]

Abstract

Background Single-cell transcriptome sequencing (scRNA-seq) has revolutionized the study of immune cells by overcoming the limitations of traditional antibody-based identification and isolation methods. This advancement allows us to obtain comprehensive gene expression profiles from a diverse array of vertebrate species, facilitating the identification of various cell types. Comparative immunology across vertebrates presents a promising approach to understanding the evolution of immune cell types. In this study, we conducted a comparative transcriptome analysis of peripheral blood mononuclear cells (PBMCs) at the single-cell level across 12 species.

Results Our findings shed light on the cellular compositional features of PBMCs, spanning from fish to mammals. Notably, we identified genes that exhibit vertebrate universality in characterizing immune cells. Moreover, our investigation revealed that monocytes have maintained a conserved transcriptional regulatory program throughout evolution, emphasizing their pivotal role in orchestrating immune cells to execute immune programs.

Conclusions This comprehensive analysis provides valuable insights into the evolution of immune cells across vertebrates.

Keywords scRNA-seq, Across species, Innate immunity, Adaptive immunity, Evolution

†Siyu Zhang, Xiang Fang and Mengyang Chang contributed equally to this work.

*Correspondence:
Jingting Shu
shujingting@163.com
Qinghua Nie
nqinghua@scau.edu.cn
Zhenhui Li
lizhenhui@scau.edu.cn
¹State Key Laboratory of Swine and Poultry Breeding Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular

Breeding, College of Animal Science, South China Agricultural University, Guangzhou, Guangdong 510642, China
²Key Laboratory of Chicken Genetics, Breeding and Reproduction, Ministry of Agriculture and Rural Affair, South China Agricultural University, Guangzhou, Guangdong 510642, China
³Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong 999077, China
⁴Institute of Aquatic Biotechnology, College of Life Sciences, Qingdao University, Qingdao, Liaoning 266071, China
⁵Key Laboratory for Poultry Genetics and Breeding of Jiangsu Province, Jiangsu Institute of Poultry Science, Yangzhou, Jiangsu 225125, China

Zhang *et al. BMC Genomics*   (2024) 25:1169

Page 2 of 12

Background

Host's response to invasive pathogens is a fundamental physiological reaction observed across all organisms. Even prokaryotes employ restriction enzymes and clustered regularly interspaced palindromic repeats (CRISPRs) to defend against invading foreign pathogens [1]. Unicellular eukaryotic amebae have evolved the ability to phagocytose foreign material as a part of their food uptake mechanism, and this basic phagocyte function is conserved as an immunological function in invertebrates and vertebrates. In invertebrates, various types of phagocytes (such as amebocytes, hemocytes, coelomocytes, granulocytes, monocytes, and macrophages) distinguish between self and non-self and display a wide array of innate immune functions. In primitive vertebrates (agnathans), such as jawless fish, lymphocytes are characterized by the widespread expression of leucine-rich repeat (LRR) sequences, which are receptors that integrate innate and adaptive immunity. The immune system of vertebrates (gnathostomes) is a highly complex structure that coordinates various types of both innate and adaptive immune cells to recognize and initiate a defensive response against potentially lethal pathogens, including bacteria, viruses, fungi, and parasites. The innate immune system consists of various types of cells, such as granulocytes, NK cells, monocytes/macrophages, dendritic cells, and mast cells. Adaptive immune cells are broadly classified into B and T cells that can directly recognize antigens with great specificity.

The exploration of immune systems in evolutionarily diverse vertebrates provides a compelling method to decipher the evolutionary pressures that have shaped immune mechanisms, molecules, specialized cells, and structures over time. The advent of large-scale animal genome sequencing has revealed genomic elements that are deeply conserved across the immune systems of various species. Traditionally, in humans and mice, clusters of differentiation (CD) antigens have been pivotal in identifying and isolating distinct immune cell types for comprehensive phenotypic and functional analyses [2–4]. However, the application of standard antibody-based cell purification across species is still challenging due to epitope and antibody differences. The emergence of scRNA-seq offers a promising solution to these obstacles, enabling the analysis of immune cell heterogeneity in invertebrates such as shrimp, oysters, mosquitoes, and Drosophila [5–8]. This approach has shed light on the evolution and development of immune cells [9–11].

To deepen our understanding of the evolution of vertebrate immunity from nonmammalian to mammalian at the cellular level, single-cell atlases of peripheral blood mononuclear cells (PBMCs) have been developed for 12 distinct species. This initiative has facilitated the analysis of conserved genes within PBMCs across these species, allowing for a comparative examination of conserved and divergent patterns of marker genes, cellular interactions, and genetic regulatory networks across a diverse array of species.

Methods

Cell capture, sequencing and alignment

In this study, a peripheral blood mononuclear cell (PBMC) atlas of six species, namely, *Tachysurus fulvidraco* (yellow catfish), *Sebastes schlegelii* (Jacopever), *Pelodiscus sinensis* (Chinese softshell turtle), *Gallus gallus* (chicken), *Rattus norvegicus* (rat), and *Homo sapiens* (human), was generated. In the next section, *Tachysurus fulvidraco*, *Sebastes schlegelii*, *Pelodiscus sinensis*, *Gallus gallus*, *Rattus norvegicus* and *Homo sapiens* are referred to as catfish, jacopever, Chinese softshell turtle, chicken, rat and human, respectively. Each species contained two samples, and all animals were female, with ages were 6 months (catfish and jacopever), 2 years (Chinese softshell turtle), 2.5 years (chicken), 2 months (rat) and 26 years (human). PBMCs were collected from peripheral blood by density gradient centrifugation. The collected cells were stained with 0.4% trypan blue to estimate cell viability. Next, cells with >85% viability were subjected to further scRNA-seq.

A total of 12 sample pools were loaded into different lanes of BMKMANU chips, which were utilized in conjunction with the BMKMANU DG1000 Library Construction Kits. The BMKMANU DG1000 system (Biomarker) was used to generate cDNA libraries. The libraries were then fragmented and sequenced on an Illumina NovaSeq 6000 (Illumina).

The generated raw reads were then processed and aligned to the respective reference genomes for each species, including GCF_015220745.1 (catfish), GCF_022655615.1 (jacopever), GCF_000230535.1 (Chinese softshell turtle), GCF_016699485.2 (chicken), GCF_015227675.2 (rat), and GCF_000001405.40 (human), using BSCMATRIX (http://www.bmkmanu.com/portfolio/tools) with default parameters. This alignment process enabled the filtering of both cell and unique molecular identifier (UMI) barcodes, resulting in high-quality data for gene expression quantification per individual cell.

Preprocessing and quality control

The gene expression matrix of all samples was loaded in R (version 4.2.2) using Seurat (version 4.3.0), and the analysis was performed with a standard workflow (SCTransform, RunPCA, RunUMAP, FindNeighbors and FindClusters) [12]. After preprocessing, DoubletFinder (version 2.0.3) [13] was used to evaluate doublets. Cells with fewer than 300 detected genes were defined as low-quality cells in 12 species. For mitochondrial gene

content thresholds, 20% for chicken based on their distributions; 20% for pig and cattle as reference [14, 15]; 10% for mouse and rat based on their distributions; 20% for human based on their distributions and reference (Figure S1A-E) [16]. For catfish, jacopever and Chinese softshell turtle, filtering was not applicable due to the absence of annotated mitochondrial chromosomes and genes in their genomes. In Egyptian fruit bat, Rhesus macaque, and chimpanzee datasets, mitochondrial genes were excluded from the matrix data [17]. Hemoglobin-expressing cells were defined as erythrocytes. Doublets, low-quality cells and erythrocytes were removed before subsequent analysis and visualization. After preprocessing and quality filtering, samples from the same species were merged. We conducted a comprehensive evaluation using the scIB framework [18] to assess the performance of 12 single-cell data integration tools: MNN [19], scVI [20], scANVI [21], Scanorama [22], BBKNN [23], SAUCIE [24], Harmony [25], ComBat [26], DESC [27], trVAE [28], trVAEP [28] and scGen [29]. This rigorous benchmarking process revealed that Harmony consistently achieved the highest overall integration score across multiple metrics (Table S1). Based on these results, we proceeded to correct for batch effects across different samples using Harmony (version 1.0), employing its default parameters. We estimated the cell cycle distribution with CellCycleScoring (SCTransform) while regressing the cell cycle. After RunHarmony or RunPCA, dimensions accounting for 95% of the total variance were used to generate uniform manifold approximation (RunUMAP) and SNN graphs (FindNeighbors). Leiden clustering (FindCluster) was then performed on the output graphs with default resolution.

Differentially expressed genes

Differentially expressed genes (DEGs) were identified using the Wilcoxon rank sum test with FindAllMarkers in Seurat [12]. Genes were considered to be differentially expressed between groups if they met the following criteria: an absolute average log fold change (|avg_log2FC|) greater than 0.25 and an adjusted p-value (p_val_adj) less than 0.05, as calculated by Benjamini-Hochberg correction for multiple testing. To reduce noise from genes with low expression, only genes detected in at least 1% of cells in one of the groups being compared were included in the differential expression analysis.

Cell type annotation

For humans and mice, singleR (version 2.0.0) [30] and scType [31] were employed to automatically annotate cell types. After automatic annotation, the cell types were manually corrected based on the marker genes from CellMarker 2.0 [32]. For other species, conserved orthologous marker genes were used to define the cell types

(Table S2). The annotation was verified by examining the Gene Ontology (GO) biological process terms enriched by the upregulated DEGs within each cell cluster and the expression of reported conserved marker genes.

Orthologous conversion

To facilitate cross-species scRNA analysis, the orthologous genes were uniformly converted to human gene symbols. Orthologous pairs between human and Chinese softshell turtle, chicken, Egyptian fruit bat, pig, cattle, mice, rat, Rhesus macaque and chimpanzee were downloaded from Ensembl 109 by BioMarkt. Other orthologous pairs were predicted by OrthoFinder (version 2.5.5) with protein files as input [33]. The protein files were downloaded from the National Center for Biotechnology Information (NCBI). Finally, only one-to-one orthologous pairs were used for further analysis.

Conserved human-mouse PBMC signature

To identify conserved human and mouse PBMC markers, we scored DEGs of different cell types (obtained from FindAllMarkers) in human and mouse PBMC atlases using COSG (version 0.9.0) [34]. We then identified cell type-specific markers according to the COSG score and expression fraction. For example, we identified monocyte markers for which the COSG scores of monocytes were greater than the mean COSG score; these markers were expressed in more than 50% of monocytes and less than 30% of cells of other cell types [35]. Next, we selected filtered human markers that were highly variable genes in the mouse PBMC atlas and mouse markers that were highly variable genes in the human PBMC atlas. We finally merged human and mouse markers to construct a human-mouse PBMC signature.

We created heatmaps (Figure S4A-B) using the top 10 cell-type markers of conserved human-mouse signatures in human and mouse atlases, respectively. We quantified the conserved human-mouse signature expression level as CPM (counts per million) in each atlas (Figure S5). To minimize the interference of dropout, we used UCell (version 2.2.0) [36] to score cell-type markers of conserved human-mouse signatures in each cell of all atlases (apart from humans and mice).

Comparison of cross-species cell atlases

To reduce the impact of data sparsity in low-coverage sequencing datasets, we used SuperCell (version 1.0) to coarsely grain single-cell atlases into metacells [37]. The high variable genes and the PCA accounting for 95% of the total variance were used to form metacell. The metacells with purity 100% (all from one cell type) were used for next step of analysis. Notably, metacells not only decrease the computational cost but also increase the expression imputation and clustering consistency (Figure

Zhang *et al. BMC Genomics*   (2024) 25:1169

Page 4 of 12

S6). To systematically assess the transcriptional similarity between cell types across species, we performed unsupervised MetaNeighbor analysis [38]. Finally, the mean AUROC (area under the receiver operating characteristic curve) was used to quantify the similarity of cell-type pairs.

Gene regulatory network inference analysis

We inferred the gene regulatory network (GRN) of a given species using SCENIC [39]. To overcome the impact of the lack of reference databases for nonmodel species and enhance cross-species comparability, we used the human cisTarget database as a reference. We used cell raw count data to run the coexpression algorithm GRBboost2 implemented in pySCENIC (version 0.12.1) [40] in Python 3.8.16 and subsequently inferred the GRN with pySCENIC using default parameters. The TF regulons were defined with "hg38__refseq-r80__10kb_up_and_down_tss.mc9nr.genes_vs_motifs.rankings.feather" and " motifs-v10nr_clust-nr.hgnc-m0.001-o0.0.tbl". Then the AUCell algorithm was used to score the activity of each TF regulon in each cell.

To quantify the cell-type specificity of regulons across human cells, we calculated the Regulon Specificity Score (RSS) and scaled it to facilitate cross-species comparisons [41]. For systematic identification of cell type-specific and conserved transcription factors (TFs), we employed the human TF regulatory network as a reference. Human TF modules were identified using the Connection Specificity Index (CSI) [41]. In brief, we calculated the CSI for each TF pair based on binarized regulon AUCell scores. The resulting CSI matrix was then subjected to k-means clustering, with the optimal number of clusters determined by minimizing the total within-cluster sum of squares. To extend this analysis across vertebrates, we leveraged orthologous relationships between human TFs and those of other species. For visualization, we constructed Sankey diagrams to illustrate the relationships between TF modules and cell types. In these diagrams, a connection between a TF module and a cell type was established if the binarized Regulon Activity Score (RAS) of the TF in that cell type was 1, indicating significant activity.

GO enrichment analysis

GO enrichment analysis was performed using Metascape (http://metascape.org) [42]. Terms with a p-value<0.01, a minimum count of 3, and an enrichment factor>1.5 were collected. The most statistically significant term within a cluster was chosen to represent the cluster. The top 10 enriched terms were selected for visualizing (Fig. 5). Module terms were clustered by membership similarities, with inter-term relationships visualized in a network plot (similarity threshold>0.3) (Figure S11).

Results

PBMC atlases of twelve species

Six PBMC atlases were obtained from public databases, including pig [14], cattle [15], Egyptian fruit bat [17], Rhesus macaque [17], chimpanzee [17], and mouse (10× Genomics). This study generated six additional PBMC cell atlases covering catfish, jacopever, Chinese softshell turtles, chickens, rats, and humans. To facilitate cross-species comparisons across these 12 species, the raw unique molecular identifier (UMI) data were reanalyzed via a standardized and uniform approach (see Materials and Methods). Following the exclusion of doublets, low-quality cells and erythrocytes, a total of 168,716 cells (catfish: 18431, jacopever: 15,617, Chinese softshell turtle: 9,379, chicken: 16,782, Egyptian fruit bat: 17,389, pig: 13,984, cattle: 17,144, mouse: 14,726, rat: 11,647, Rhesus macaque: 13,115, chimpanzee: 4,833, human: 15,669) were obtained (Table S3). Various strategies were employed in combination to annotate the cell types of PBMCs from different species (see Materials and Methods). Marker genes were utilized for cell annotation (Figure S2, Table S2, Table S4). In total, ten distinct cell types were identified across these species, including B cells, T cells, NK cells, monocytes, myeloid dendritic cells (mDCs), plasmacytoid dendritic cells (pDCs), dendritic cells (DCs), neutrophils, platelets, and hematopoietic stem cells (HSCs) (Fig. 1 and Figure S3). The PBMCs of these 12 species commonly contain B cells, T cells, monocytes, and DCs (Table S5). The successful annotation of PBMCs from these 12 species provides us with an opportunity for cross-species comparisons and evolutionary studies.

Cross-species conserved genes of immune cells

The marker genes for mouse and human PBMCs have undergone extensive validation. To assess the evolutionary conservation of these marker genes between humans and mice, we identified the marker genes for PBMCs in both species within the respective atlases and compiled the cumulative genes for the same cell type (see Materials and Methods). According to the profiles of both humans and mice, these genes exhibited great differences between different cell types (Figure S4A-B). The PBMCs shared conserved orthologs across the 12 species (Figure S4C, Table S6). To quantify gene set enrichment in single-cell data, we employed the UCell score method. This approach compares the expression of genes in a target set to a randomly selected background set of control genes. The resulting score is normalized to account for variations in gene expression levels and the number of detected genes per cell, ensuring comparability across diverse cell types and conditions, regardless of differences in gene detection rates. We subsequently analyzed the expression of conserved marker genes and calculated
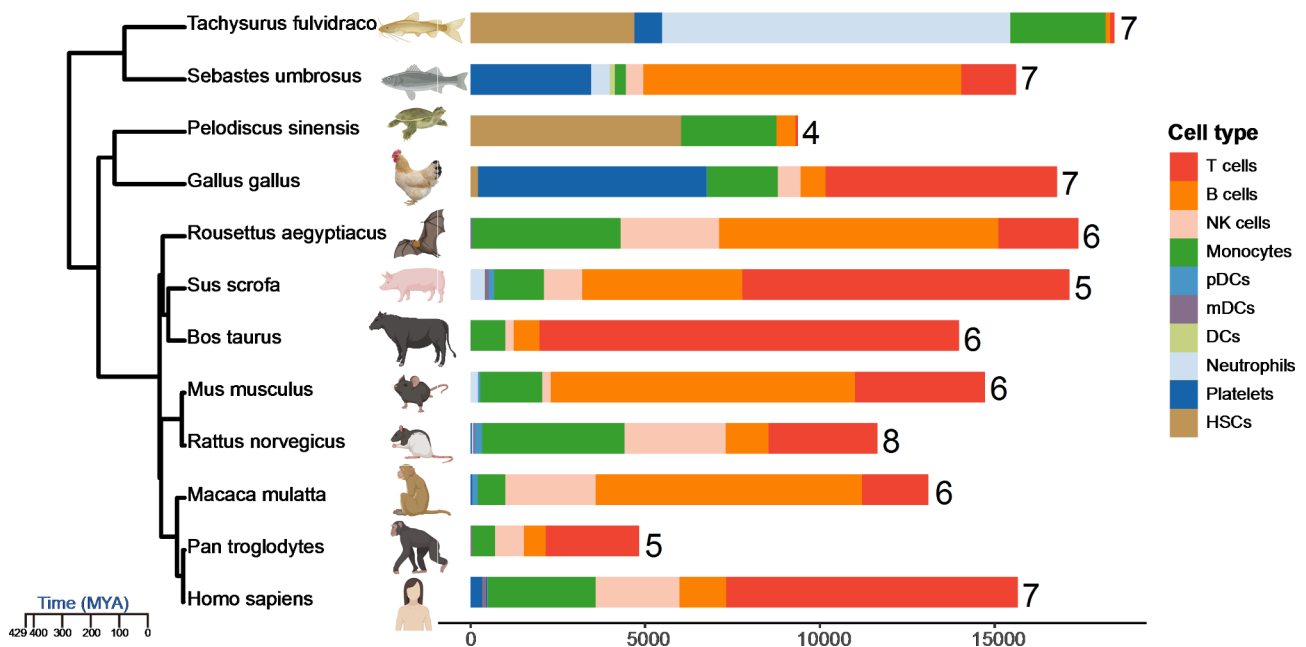
Fig. 1 Immune cell composition in the PBMCs across vertebrates. PBMC atlases from catfish (*Tachysurus fulvidraco*), jacopever (*Sebastes schlegelii*), Chinese softshell turtle (*Pelodiscus sinensis*), chicken (*Gallus gallus*), Egyptian fruit bat (*Rousettus aegyptiacus*), pig (*Sus scrofa*), cattle (*Bos taurus*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), Rhesus macaque (*Macaca mulatta*), chimpanzee (*Pan troglodytes*) and human (*Homo sapiens*). The numbers at the end of the bars indicate the number of cell types identified in each species

UCell scores in non-mouse and non-human species (Fig. 2, Figure S5). Our findings demonstrate that human-mouse conserved marker genes for B cells, T cells, NK cells, monocytes, neutrophils, and platelets effectively identify corresponding cell populations across all species, based on both gene expression levels and UCell scores. The human-mouse conserved marker genes for pDCs showed limited efficacy in distinguishing cattle and chicken pDCs from other PBMCs, with the exception of those in rats. Similarly, the conserved marker genes for mDCs exhibited variable conservation across species.

Analysis of immune cell similarity across species

MetaNeighbor analysis [38], which utilizes orthologous gene expression, is robust for identifying the similarities and heterogeneities of different cell types among mammalian species [43] and is sensitive for revealing cell-type relationships [44]. To determine the cell types in a given species, pairwise unsupervised MetaNeighbor analyses were performed to quantify the similarity between cell-type pairs. Prior to MetaNeighbor analysis, the meta-cell strategy [37] was employed to increase the gene number, improve clustering consistency (Figure S6), and reduce computation time. Notably, the AUROC for the same cell type was greater than that for different cell types (Figure S7A), a trend that was consistently observed in cross-species cell-type comparisons and was unaffected by intraspecies cell interference (Figure S7B-C).

The cell-type heatmap arrangement, based on AUROC scores between cell types, underwent hierarchical clustering (Fig. 3). High AUROC values led to the clustering of monocytes from various species into a distinct module, which also included dendritic cells (DCs) from cattle, Rhesus macaques, chimpanzees, and humans (Fig. 3). This clustering suggests a similarity in gene expression patterns between mammalian DCs and monocytes. The observed coclustering highlights a potential shared transcriptional profile between these immune cell types across different species.

Transcription factor regulatory programs underlying cell-type identity

The transcriptional state of a cell reflects its transient condition and is governed by an underlying gene regulatory network (GRN) regulated by transcription factors (TFs). Our MetaNeighbor analysis revealed that cells of the same cell type exhibit similar gene expression patterns across species, suggesting a conserved regulatory mechanism. Using SCENIC, a method for constructing regulatory networks and predicting cell-specific TFs from single-cell gene expression data, we explored the conservation of TF repertoires across species [39, 41, 44]. To investigate whether conserved cell types are accompanied by conserved TF repertoires, we used human TFs as a reference for other species.

Using SCENIC, we evaluated TF regulons in all 12 species (Figure S8). By matching with TF regulons in
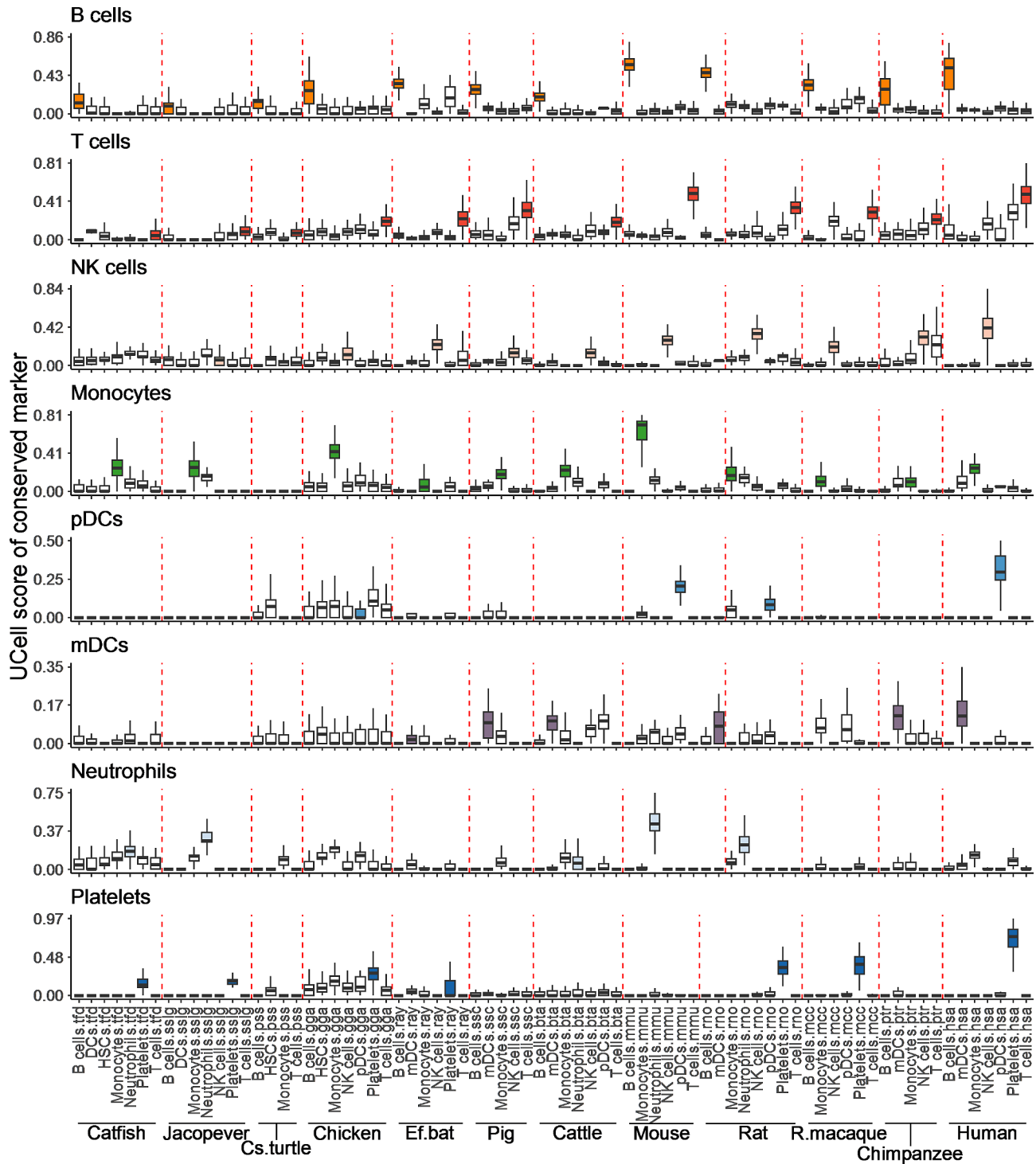
Zhang *et al. BMC Genomics*   (2024) 25:1169

Page 6 of 12

Fig. 2 Conservation marker genes of PBMCs. UCell scores of different PBMC cell type human-mouse signature genes in B cells, T cells, NK cells, Monocytes, pDCs, mDCs, Neutrophils and Platelets. Red lines demarcate individual species. Boxes corresponding to the same cell type as the human-mouse signature genes are highlighted in color. Abbreviations: tfd: catfish, sslg: jacopever, pss: Chinese softshell turtle, gga: chicken, mmu: mouse, ray: Egyptian fruit bat, ssc: pig, bta: cattle, rno: rat, mcc: Rhesus macaque, ptr: chimpanzee, has: human

humans, we found that other species shared 44 to 97 of the same TF regulons as humans, and these regulons account for between 42.98% and 54.04% of those identified by SCENIC (Figure S8). We identified B-cell-specific TF BCL11A, T-cell-specific TF TCF7, NK cell-specific TF EOMES, monocyte-specific TF RXRA, mDC-specific TF SPI1, and pDC-specific TF IRF7 in human (Figure S9) and these TFs are essential for the maintenance of
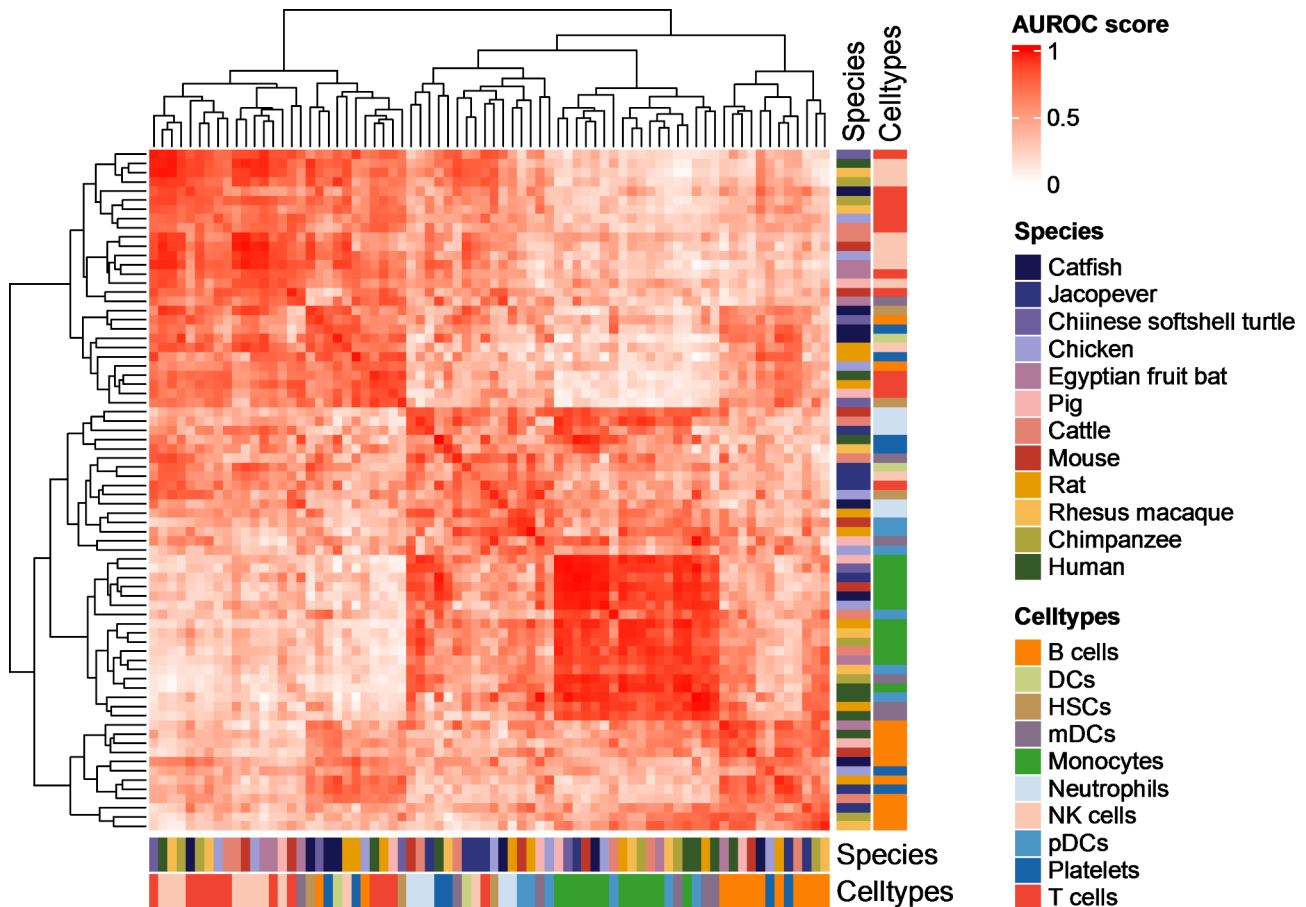
Fig. 3 Cross-species similarity comparison of cell types. AUROC scores of cell-type pairs similarity in all PBMC cell types from 12 vertebrates

cellular identity [41, 45–50]. Moreover, the identification of cell type-specific TFs based on regulon specificity scores (RSSs) in *Danio rerio, Caenorhabditis elegans, Ciona intestinalis, Schmidtea mediterranea, Nematostella vectensis* was feasible [44].

However, the TFs with the highest RSS values in human PBMCs were not always conserved in the PBMCs of other species. For example, TCF7 is a TF with a high RSS value only in humans, cattle, and Egyptian fruit bats (Figure S10). To identify the most conserved TFs in the PBMCs across species, we scaled the RSS values of the PBMCs of all 12 species and clustered the TFs based on the RSSZs [41]. The orthologous TFs of FOSL2, FOS, TCF7L2, SPI1, RXRA, and CEBPD were found to be activated in monocytes conservatively (Fig. 4). Among them, FOSL2, FOS, TCF7L2, and SPI1 are activated in all 12 species (Fig. 4, Table S7). These four TFs were activated in human monocytes (Figure S11A). The regulatory relationships of four TFs and their target genes with GENIE3 > 1 were visualized in an interaction network (Figure S11B). These four TFs shared fewer co-target genes, revealing the relative independence of the target genes among these four TFs (Figure S11B). We performed Gene Ontology (GO)

analysis to characterize the enriched functions of target genes among these four TFs, revealing their regulation of 'hemopoiesis', 'response to hormone', 'endocytosis', and 'regulation of cell activation' of monocytes among 12 species (Figure S11C).

To systematically characterize the combinational patterns of expressed TFs, we compared the atlas-wide (human) similarity of binarized regulon activity scores (RAS) of every regulon pair based on the connection specificity index (CSI) [51] (see Materials and Methods). After hierarchical clustering, 190 regulons were organized into seven major TF modules (Fig. 5A). For each module, we selected several representative TFs and cell types through their RAS (Fig. 5A). Modules 1 and 7 are associated with all cell types (Fig. 5A, Figure S12). Module 2 is mainly associated with monocytes and mDCs (Fig. 5A, Figure S12). Module 3 has a low level of cellular activation of regulators that are not representative of a particular cell type (Fig. 5A, Figure S12). Module 4 contains a series of monocyte- and mDC-specific regulators, such as RXRA [52] and IRF7 [53]. The regulators of Module 5 are expressed in T cells and NK cells (Figure S12). Module 6 contains regulators that are specifically
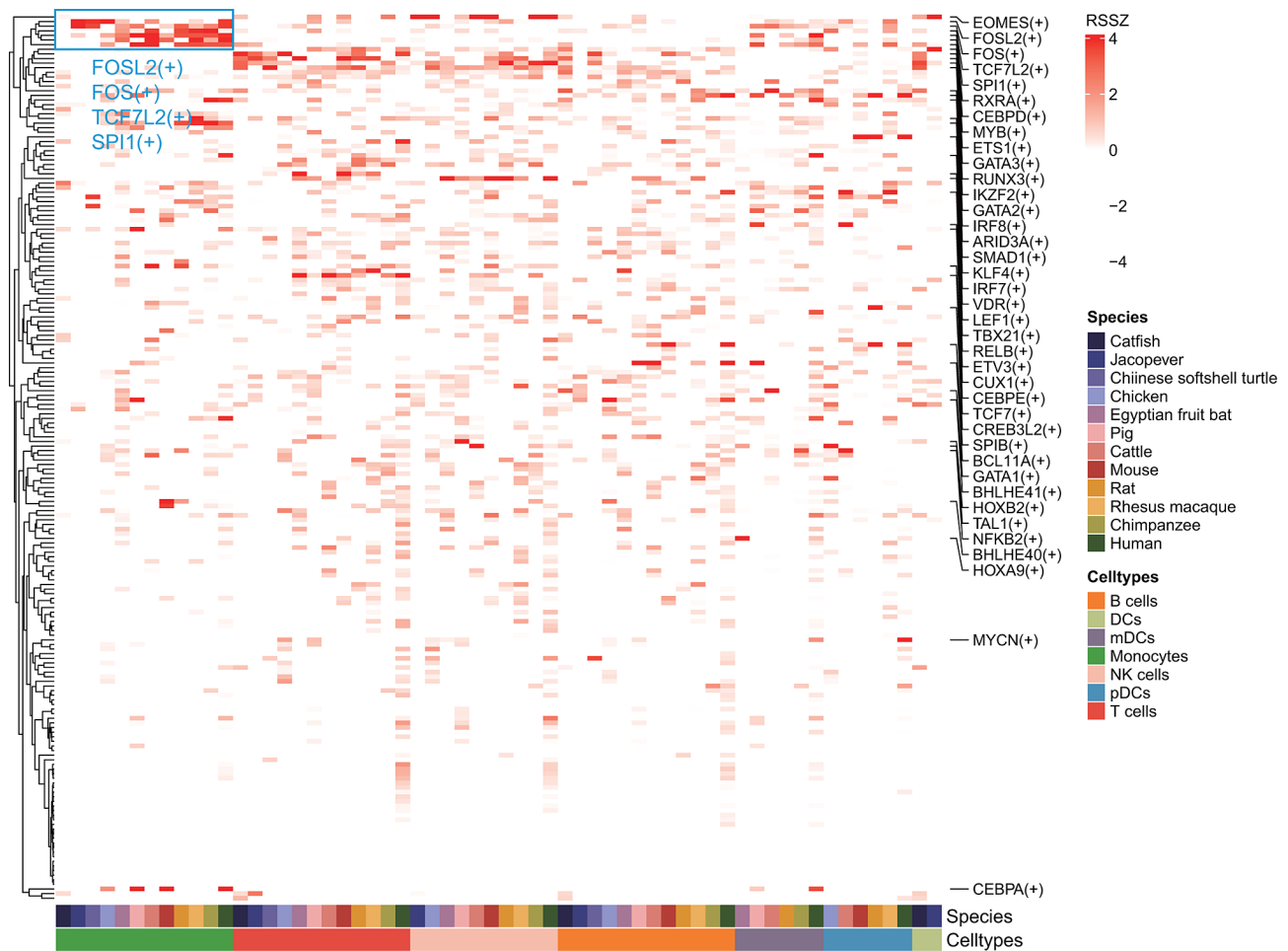
Fig. 4 Conservation regulons in PBMCs across vertebrates. Clustering analysis of 12 vertebrates PBMC cell type specific TF regulons. The regulons are clustered according to regulon specificity score Z-score (RSSZ). The regulons with RSSZ greater than 4 are annotated on the right side of the heatmap

expressed in B cells, such as BCL11A [45] and SPIB [54]. The TFs in Modules 1, 4 and 7 have high CSI. The TFs in Module 1 and 7 are enriched in "embryonic morphogenesis" and "tube morphogenesis", which may be the common character of all cell types in PBMC. The TFs in Module 4 are enriched in "hemopoiesis" and "Adipogenesis". Interestingly, monocyte TFs (FOSL2, FOS, TCF7L2, and SPI1) conserved across species were found in Module 4 (Table S8), reflecting their important roles in synergistically regulating the biological co-functions of monocytes.

To compare the conservation of these seven modules, we mapped the RASs of the regulons to cell types (see Materials and Methods). As visualized in the Sankey plot (Figure S13), Module 4 shared conserved connections with monocytes across species. In all the given species, an average of 72.29% of the regulons were shared by humans in Module 4, which was greater than that of the other modules (Figure S14A). Additionally, the most conserved TFs identified in the monocytes of the PBMCs of

the given species were the regulons in Module 4 (Figure S14B).

Discussion

The advent of full-genome sequencing in various vertebrate species has greatly facilitated the exploration of immune system evolution by leveraging homologous relationships with established mammalian immune cell markers. However, progress in comparing immune cell types at the cellular and molecular levels has been hindered by the lack of suitable antibodies to label distinct immune cell populations in lower vertebrates. Advances in scRNA-seq technology have played a pivotal role in overcoming this limitation by providing unbiased cellular resolution gene expression atlases. In this study, we applied scRNA-seq to sequence PBMCs from a diverse range of vertebrate species, including catfish (*Tachysurus fulvidraco*), jacopever (*Sebastes schlegelii*), Chinese softshell turtles (*Pelodiscus sinensis*), chickens (*Gallus gallus*), rats (*Rattus norvegicus*) and humans (*Homo sapiens*). Additionally, we incorporated publicly available
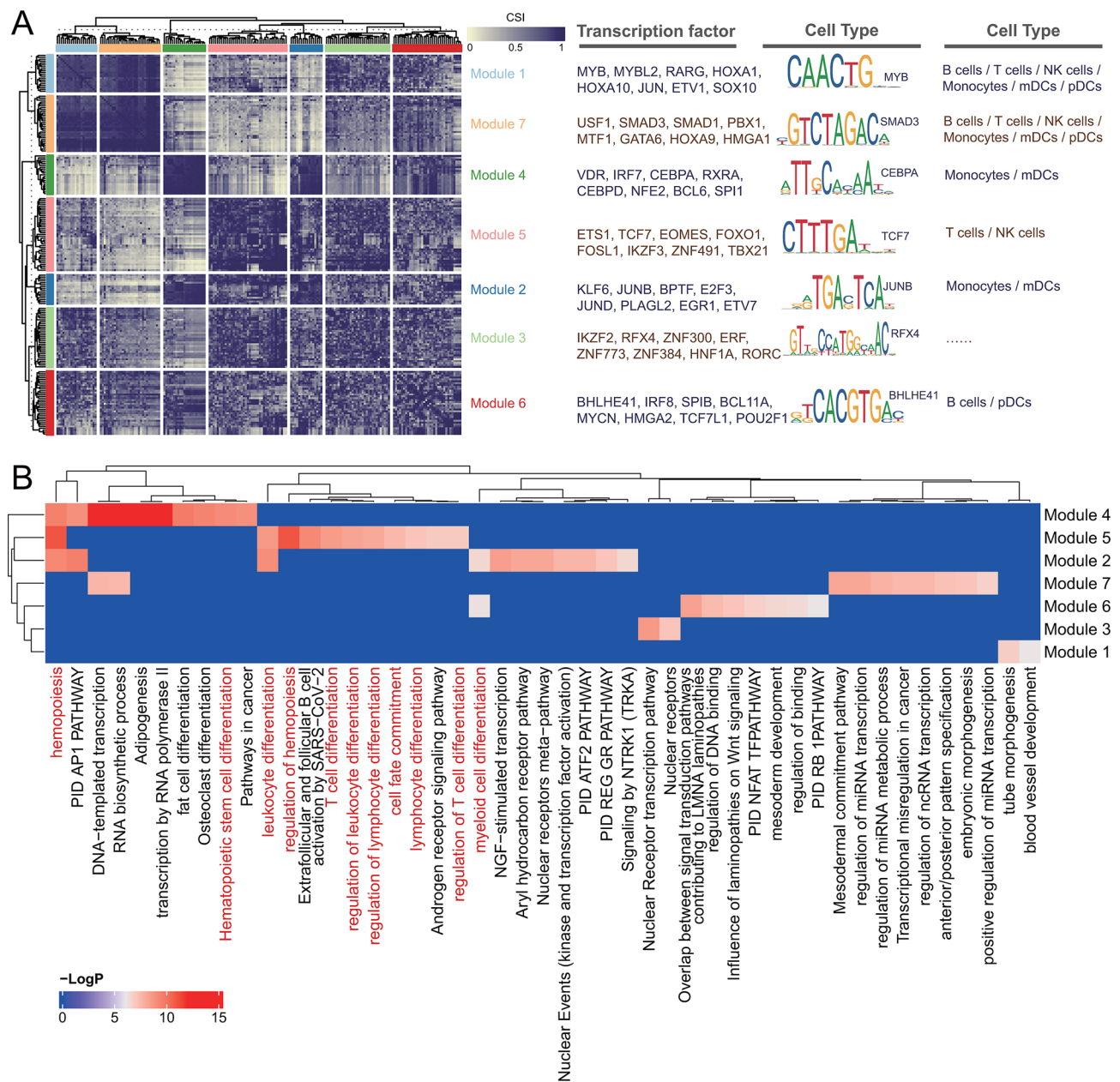
Fig. 5 Transcription factor programs in vertebrate PBMC evolution. (**A**) Identification of 7 human TF modules (modules 1–7) based on the regulon connection specificity index (CSI) of the human PBMC atlas, along with representative transcription factors, corresponding binding motifs, and associated cell types; (**B**) GO biological processes enrichment analysis of seven module regulons. The terms with -LogP > 6 and top 10 of modules were plotted

PBMC scRNA atlases of Egyptian fruit bats (*Rousettus aegyptiacus*), pigs (*Sus scrofa*), cattle (*Bos taurus*), Rhesus macaques (*Macaca mulatta*) and chimpanzees (*Pan troglodytes*). Integrated analyses of PBMC scRNA atlases from these 12 vertebrate species offered a novel perspective for cross-species immune cell comparisons.

The adaptive immune system, a key component of immune responses in jawed vertebrates, has recently gained attention through scRNA-seq studies focusing on zebrafish bone marrow, thymus, spleen and kidney, identifying B cells, T cells and NK cells [9, 55–57]. Expanding

on these findings, our study characterized lymphocytes in PBMCs from catfish and jacopever, providing insights into the main cell types, including B cells, T cells, NK cells, monocytes, and dendritic cells (DCs), from fish to mammals.

While scRNA-seq has generated large datasets in humans and mice, particularly in PBMCs [58–60], stable and reliable markers for annotating PBMCs of nonmodel animals remain scarce. Existing studies have identified markers for different cell types across various tissues of humans and mice, with several useful websites offering

a plethora of valid markers for cell annotation [32, 61]. However, these markers are not readily applicable to nonmodel animals. The expression of genes with one-to-one orthologs can be used to annotate cell types [35, 62]. Leveraging the broad validation of human and mouse markers, we screened for conserved markers across species for peripheral blood mononuclear cell (PBMC) annotation. The identified conserved human-mouse markers demonstrated excellent performance in identifying B cells, T cells, monocytes, neutrophils, and platelets. The NK cell markers effectively identified NK cells in species such as chickens, Egyptian fruit bats, pigs, cows, mice, rats, Rhesus macaques, chimpanzees, and humans. However, they were significantly less effective for identifying NK cells in jacopever. Similarly, while the pDC and mDC markers performed well in identifying these cells in mice, rats, Rhesus macaques, chimpanzees, and humans, they posed challenges in species like catfish, jacopever, chickens, Egyptian fruit bats, pigs, and cattle. Despite these limitations, the markers provide a valuable reference for annotating cell subpopulations using scRNA-seq in non-model animals.

The scRNA-seq data typically exhibit high sparsity, meaning that for each individual cell, only a subset of genes has detectable expression, while the majority of genes have expression values of zero. To address this limitation and improve the robustness of our Meta-Neighbor analysis, we employed the Metacells approach to reduce data sparsity. This strategy aggregates similar cells into metacells, effectively increasing gene detection and improving clustering consistency. While the use of orthologous genes is crucial for cross-species scRNA-seq analysis, it also has some limitations [63–66]. we acknowledge that this approach may overlook species-specific genes or those that have undergone duplication or divergence. Nevertheless, previous studies have demonstrated that cell type similarity in orthologous gene expression often overrides species differences when applying MetaNeighbor analysis [43, 44].

The application of MetaNeighbor analysis in our study demonstrated its reliability in assessing the similarity and heterogeneity of different cell types across diverse species. The same cell type is more similar between species than different cell types within the same species (Figure S7A). This revealed that the key features of different cell types have appeared during early evolution. Monocytes from all species formed a well-clustered cluster, showing their remarkable conservation. Interestingly, within the monocyte clusters, DCs were also present, suggesting a close similarity between DCs and monocytes, likely attributed to shared functions such as phagocytosis and antigen presentation.

Considering that cell-type similarities in gene expression may stem from convergent or concerted evolution [67], we employed TF regulatory programs to validate the proposed homologies. However, the lack of comprehensive cis-regulatory databases for non-model species is a challenge. Nevertheless, the application of SCENIC to a range of vertebrates, including pig [68], cow [69], and zebrafish [70], as well as to invertebrates such as schmidtea and nematostella [44], provides valuable insights into these regulatory networks. We utilized SCENIC to identify cell type-specific TFs by assuming a degree of conservation in TF binding sequences. Conserved TFs, including FOSL2, FOS, TCF7L2 and SPI1, were identified in monocytes across species. These TFs exhibited target genes involved in essential processes such as 'hemopoiesis', 'response to hormone', 'endocytosis' and 'regulation of cell activation'. Furthermore, these TF regulons act cooperatively to play regulatory roles in human monocytes. It is important to note that using the human cisTarget database to identify activated TFs in PBMCs of different species may overlook TFs active in non-human species, particularly non-mammals. Instead, we focus on conserved transcription factors based on one-to-one homologous genes, which enhances the feasibility and accuracy of our analysis.

Conclusions
In summary, our work extends the analysis of immune cell types and their evolution to lower vertebrates, providing insights into adaptive immune cells across fish and mammalian species. Through scRNA-seq analyses and cross-species comparisons, we identified conserved markers, characterized functional similarities, and explored the regulatory programs of immune cells. The observed conservation in gene expression patterns and regulatory programs in monocytes highlights the conservatism of monocytes during evolution.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-11030-6.

Supplementary Material 1

Supplementary Table S1: Scores of scIB metrics.

Supplementary Table S2: Cell type annotation marker genes for PBMC atlases of 12 species.

Supplementary Table S3: Cell number statistics of 12 species

Supplementary Table S4: Differential expression genes of different cell types in 12 species.

Supplementary Table S5: The number and proportion of different cell types in 12 species.

Supplementary Table S6: PBMCs conserved orthologous across 12 species.

Supplementary Table S7: The conserved transcription factors across 12 species.

Supplementary Table S8: The regulons of seven modules in 12 species.

Zhang *et al. BMC Genomics*   (2024) 25:1169

Page 11 of 12

Author contributions
Siyu Zhang, Xiang Fang, Mengyang Chang contributed equally to this work.
Authors and AffiliationsState Key Laboratory of Swine and Poultry Breeding
Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and
Molecular Breeding, College of Animal Science, South China Agricultural
University, Guangzhou 510642, Guangdong, China.Key Laboratory of Chicken
Genetics, Breeding and Reproduction, Ministry of Agriculture and Rural
Affair, South China Agricultural University, Guangzhou 510642, Guangdong,
China.Institute of Aquatic Biotechnology, College of Life Sciences, Qingdao
University, Qingdao, 266071, ChinaKey Laboratory for Poultry Genetics and
Breeding of Jiangsu Province, Jiangsu Institute of Poultry Science, Yangzhou
225125, Jiangsu, China.ContributionsS.Z., J.S., Q.N. and Z.L. conceived the study
and wrote the first draft of manuscript. Q.N. and Z.L. provided funding for the
project. S.Z., X.F. and M.C. generated, collected and/or analyzed the single-cell
data. M.Z. offered the analysis platform. L.G. and Y.X. provided intellectual input
to shape the study design. All coauthors commented on and approved the
manuscript. Corresponding authorCorrespondence to Jingting Shu, Qinghua
Nie and Zhenhui Li.

Data availability
For Mus musculus (mouse), feature matrixes of C57BL/6 and BALB/c mice
were downloaded from datasets released by 10× Genomics. For Rousettus
aegyptiacus (Egyptian fruit bat), Macaca mulatta (Rhesus macaque), Pan
troglodytes (chimpanzee), Sus scrofa (pig) and Bos taurus (cattle), feature
matrixes were downloaded from the Gene Expression Omnibus (GEO) with
accession numbers GSE218199, GSE218199, GSE218200, GSE193975 and
GSE166245, respectively. The raw sequence data generated in this study
have been deposited into the CNGB Sequence Archive (CNSA) of the China
National GenBank DataBase (CNGBdb) with accession number CNP0004344.

Declarations

Ethics approval and consent to participate
The animal samples used in this study were approved by the Ethics
Committee of Qingdao University. The human samples were approved by the
Ethics Committee of the Medical College of Qingdao University, and written
informed consent was obtained from the donors in accordance with the
Declaration of Helsinki.

Consent for publication
All authors consented to the publication of this study.

Competing interests
The authors declare no competing interests.

References
1. Barrangou R, Marraffini LA. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. Mol Cell. 2014;54:234–44.
2. Allard B, Longhi MS, Robson SC, Stagg J. The ectonucleotidases CD39 and CD73: novel checkpoint inhibitor targets. Immunol Rev. 2017;276:121–44.
3. Mora-Velandia LM, Castro-Escamilla O, Méndez AG, Aguilar-Flores C, Velázquez-Avila M, Tussié-Luna MI et al. A human Lin – CD123 + CD127low
Population endowed with ILC features and migratory capabilities contributes to Immunopathological Hallmarks of Psoriasis. Front Immunol. 2017;8.
4. Castiglioni A, Yang Y, Williams K, Gogineni A, Lane RS, Wang AW, et al. Combined PD-L1/TGFβ blockade allows expansion and differentiation of stem cell-like CD8 T cells in immune excluded tumors. Nat Commun. 2023;14:4703.
5. Cattenoz PB, Monticelli S, Pavlidaki A, Giangrande A. Toward a Consensus in the repertoire of Hemocytes identified in Drosophila. Front Cell Dev Biol. 2021;9.
6. Kwon H, Mohammed M, Franzén O, Ankarklev J, Smith RC. Single-cell analysis of mosquito hemocytes identifies signatures of immune cell subtypes and cell differentiation. eLife. 2021;10:e66192.
7. Meng J, Zhang G, Wang W-X. Functional heterogeneity of immune defenses in molluscan oysters Crassostrea hongkongensis revealed by high-throughput single-cell transcriptome. Fish Shellfish Immunol. 2022;120:202–13.
8. Yang P, Chen Y, Huang Z, Xia H, Cheng L, Wu H, et al. Single-cell RNA sequencing analysis of shrimp immune cells identifies macrophage-like phagocytes. eLife. 2022;11:e80127.
9. Carmona SJ, Teichmann SA, Ferreira L, Macaulay IC, Stubbington MJT, Cvejic A, et al. Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. Genome Res. 2017;27:451–61.
10. Geirsdottir L, David E, Keren-Shaul H, Weiner A, Bohlen SC, Neuber J, et al. Cross-species single-cell analysis reveals divergence of the Primate Microglia Program. Cell. 2019;179:1609–e162216.
11. Li Z, Sun C, Wang F, Wang X, Zhu J, Luo L et al. Molecular mechanisms governing circulating immune cell heterogeneity across different species revealed by single-cell sequencing. Clin Transl Med. 2022;12.
12. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573–e358729.
13. McGinnis CS, Murrow LM, Gartner ZJ, DoubletFinder. Doublet Detection in single-cell RNA sequencing data using Artificial Nearest neighbors. Cell Syst. 2019;8:329–e3374.
14. Vaure C, Grégoire-Barou V, Courtois V, Chautard E, Dégletagne C, Liu Y. Göttingen Minipigs as a model to evaluate longevity, functionality, and memory of Immune Response Induced by Pertussis vaccines. Front Immunol. 2021;12.
15. Gao Y, Li J, Cai G, Wang Y, Yang W, Li Y, et al. Single-cell transcriptomic and chromatin accessibility analyses of dairy cattle peripheral blood mononuclear cells and their responses to lipopolysaccharide. BMC Genomics. 2022;23:338.
16. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martínez-Colón GJ, McKechnie JL, et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat Med. 2020;26:1070–6.
17. Aso H, Ito J, Ozaki H, Kashima Y, Suzuki Y, Koyanagi Y, et al. Single-cell transcriptome analysis illuminating the characteristics of species-specific innate immune responses against viral infections. GigaScience. 2023;12:giad086.
18. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19:41–50.
19. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36:421–7.
20. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
21. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17:e9620.
22. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019;37:685–91.
23. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36:964–5.
24. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods. 2019;16:1139–45.
25. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
26. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
27. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11:2338.

28. Lotfollahi M, Naghipourfar M, Theis FJ, Wolf FA. Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics. 2020;36(Supplement2):i610–7.

29. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16:715–21.

30. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72.

31. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13:1246.

32. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 2023;51:D870–6.

33. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.

34. Dai M, Pei X, Wang X-J. Accurate and fast cell marker gene identification with COSG. Brief Bioinform. 2022;23:bbab579.

35. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell. 2022;185:379–e39638.

36. Andreatta M, Carmona SJ, UCell. Robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796–8.

37. Bilous M, Tran L, Cianciaruso C, Gabriel A, Michel H, Carmona SJ, et al. Metacells untangle large and complex single-cell transcriptome networks. BMC Bioinformatics. 2022;23:336.

38. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. Nat Commun. 2018;9:884.

39. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.

40. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35:2159–61.

41. Suo S, Zhu Q, Saadatpour A, Fei L, Guo G, Yuan G-C. Revealing the critical regulators of cell identity in the mouse cell Atlas. Cell Rep. 2018;25:1436–e14453.

42. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10:1523.

43. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581:303–9.

44. Wang J, Sun H, Jiang M, Li J, Zhang P, Chen H, et al. Tracing cell-type evolution by cross-species comparison of cell atlases. Cell Rep. 2021;34:108803.

45. Liu P, Keller JR, Ortiz M, Tessarollo L, Rachel RA, Nakamura T, et al. Bcl11a is essential for normal lymphoid development. Nat Immunol. 2003;4:525–32.

46. Honda K, Yanai H, Mizutani T, Negishi H, Shimada N, Suzuki N, et al. Role of a transductional-transcriptional processor complex involving MyD88 and IRF-7 in toll-like receptor signaling. Proc Natl Acad Sci USA. 2004;101:15416–21.

47. Zhu J, Nasr R, Pérès L, Riaucoux-Lormière F, Honoré N, Berthier C, et al. RXR is an essential component of the oncogenic PML/RARA complex in vivo. Cancer Cell. 2007;12:23–35.

48. Utzschneider DT, Charmoy M, Chennupati V, Pousse L, Ferreira DP, Calderon-Copete S, et al. T cell factor 1-Expressing memory-like CD8+T cells sustain the Immune response to chronic viral infections. Immunity. 2016;45:415–27.

49. Zhang J, Marotel M, Fauteux-Daniel S, Mathieu A-L, Viel S, Marçais A, et al. T-bet and eomes govern differentiation and function of mouse and human NK cells and ILC1. Eur J Immunol. 2018;48:738–50.

50. Chopin M, Lun AT, Zhan Y, Schreuder J, Coughlan H, D'Amico A, et al. Transcription factor PU.1 promotes conventional dendritic cell identity and function via induction of Transcriptional Regulator DC-SCRIPT. Immunity. 2019;50:77–e905.

51. Fuxman Bass JI, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJM. Using networks to measure similarity between genes: association index selection. Nat Methods. 2013;10:1169–76.

52. Rőszer T, Menéndez-Gutiérrez MP, Cedenilla M, Ricote M. Retinoid X receptors in macrophage biology. Trends Endocrinol Metab. 2013;24:460–8.

53. Gabriele L, Ozato K. The role of the interferon regulatory factor (IRF) family in dendritic cell development and function. Cytokine Growth Factor Rev. 2007;18:503–10.

54. Willis SN, Tellier J, Liao Y, Trezise S, Light A, O'Donnell K, et al. Environmental sensing by mature B cells is controlled by the transcription factors PU.1 and SpiB. Nat Commun. 2017;8:1426.

55. Tang Q, Iyer S, Lobbardi R, Moore JC, Chen H, Lareau C, et al. Dissecting hematopoietic and renal cell heterogeneity in adult zebrafish at single-cell resolution using RNA sequencing. J Exp Med. 2017;214:2875–87.

56. Rubin SA, Baron CS, Pessoa Rodrigues C, Duran M, Corbin AF, Yang SP, et al. Single-cell analyses reveal early thymic progenitors and pre-B cells in zebrafish. J Exp Med. 2022;219:e20220038.

57. Jiao A, Zhang C, Wang X, Sun L, Liu H, Su Y, et al. Single-cell sequencing reveals the evolution of immune molecules across multiple vertebrate species. J Adv Res. 2024;55:73–87.

58. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T, JingleBells. A repository of Immune-related single-cell RNA–Sequencing datasets. J Immunol. 2017;198:3375–9.

59. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8:2032.

60. Ren X, Wen W, Fan X, Hou W, Su B, Cai P, et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. Cell. 2021;184:1895–e191319.

61. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019:baz046.

62. Sachkova M, Burkhardt P. Exciting times to study the identity and evolution of cell types. Development. 2019;146:dev178996.

63. Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. eLife. 2021;10:e66747.

64. Song Y, Miao Z, Brazma A, Papatheodorou I. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. Nat Commun. 2023;14:6495.

65. Meyer A, Ku C, Hatleberg WL, Telmer CA, Hinman V. New hypotheses of cell type diversity and novelty from orthology-driven comparative single cell and nuclei transcriptomics in echinoderms. eLife. 2023;12:e80090.

66. Wang R, Zhang P, Wang J, Ma L, E W, Suo S, et al. Construction of a cross-species cell landscape at single-cell level. Nucleic Acids Res. 2023;51:501–16.

67. Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, et al. The origin and evolution of cell types. Nat Rev Genet. 2016;17:744–57.

68. Cai S, Hu B, Wang X, Liu T, Lin Z, Tong X, et al. Integrative single-cell RNA-seq and ATAC-seq analysis of myogenic differentiation in pig. BMC Biol. 2023;21:19.

69. Gao Y, Fang L, Baldwin RL, Connor EE, Cole JB, Van Tassell CP, et al. Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning. Genomics. 2021;113:2045–55.

70. Huang Y, Liu X, Wang H-Y, Chen J-Y, Zhang X, Li Y, et al. Single-cell transcriptome landscape of zebrafish liver reveals hepatocytes and immune cell interactions in understanding nonalcoholic fatty liver disease. Fish Shellfish Immunol. 2024;146:109428.

Publisher's note