Supplementary information Gioacchino, Koyunlar et al.,

Supplementary Methods

*Single cell transcriptome analysis of the progenitor compartment of WT and gata2b$^{-/-}$ zebrafish allows for unbiased lineage investigation*

Kidney marrow from two 5 mpf female *Tg(CD41:GFP)* WT or *gata2b*$^{-/-}$ zebrafish were pooled and sorted and experiment was performed in replicate (Figure S2A-C). In total 70.000 cells from the gate in panel B were sorted and between 214 and 1607 CD41:GFP$^{low}$ cells were added to this pool in PBS/10%FSC/2%BSA/2% carp serum. This resulted in final CD41:GFP$^{low}$ percentage of 0.45 – 2.73%. The sorting strategy also included the CD41:GFP$^{high}$ expressing cells which were previously identified as thrombocytes [1] (Figure S2C) so for every CD41:GFP$^{low}$ cell, 1.9 CD41:GFP$^{high}$ cells were present in our final population. 7630 cells in WT1, 4033 in WT2, 3675 in *gata2b*$^{-/-}$ 1 and 5229 in *gata2b*$^{-/-}$ 2 were obtained after quality control (Figure S2D-F), with a read depth of approximately 50.000 reads per cell. Gross differences in cell numbers between WT and *gata2b*$^{-/-}$ cells may influence cluster identification when a nearest-neighbor algorithm is used. Therefore, all replicates were randomly down-sampled to 3675 to match each other. First the replicates were aligned using anchor based integration and then the WT and *gata2b*$^{-/-}$ samples were aligned using the same method to correctly identify the clusters, avoiding batch specific differences. We could identify 20 different cell clusters using the R Seurat package[2](Figure S2G).

*Cluster identification*

Using the FindMarkers function in Seurat, differentially expressed genes were identified compared to the other clusters. Subsequently the functions FeaturePlot and VlnPlot were used to analyse gene expression patterns between clusters to test validity and exclusivity of individual clusters. Finally, marker analysis was visualized using the DoHeatmap function. Importantly, known Gata2 target genes like *cebpb* (p = 0,0002) and *alas2* (p = 4.77E-53) were found significantly downregulated in gata2b mutants.

Erythroid cells, thrombocytes, neutrophils, monocytes and their progenitors were identified by comparing our data to known single cell expression analysis and known lineage markers[3-7]. Canonical lineage differentiation markers are generally low expressing transcription factors and are poorly amplified by droplet based single cell sequencing methods, therefore we have presented the lineage differentiation using a combination of canonical lineage differentiation markers and high differentially expressed genes between clusters as identified by the FindMarkers function of Seurat and presented in a heatmap (Figure S2H). In short, cells of the erythroid lineage are known to express hemoglobins like *hbba1. itga2b* was used as a marker for thrombocytes in the form of *Tg(CD41:GFP)*, *mpeg1.1* for monocytes and *lysozyme (lyz)* for neutrophils (Figure S2H). Furthermore, the differentiated populations were devoid of expression of proliferation genes like *myca* and *pcna,* indeed suggesting that these are differentiated cells and not a progenitor compartment (Figure S2H). Interestingly, one population

expressed markers like *CCAAT enhancer binding protein alpha* (*cebpa)* and *granulin1 (grn1)* (Figure 3B) which are expressed in the monocyte lineage. This population also showed very high expression levels of *s100a10b* (Figure S3A). *In situ* hybridization confirmed that macrophages expressed high levels of *s100a10b* (Figure S3A, B).

The sorting strategy for the kidney progenitor population also contains cells of the lymphoid lineage. A lymphoid progenitor population expressing *rag1* was found. Furthermore, this population also expressed the *rag1* homologue *topoisomerase 2a (top2a)* (Figure S2H), probably as an alternative mechanism in V(D)J recombination[8], and high expression of proliferation markers *myca* and *pcna* suggestive of the lymphoid progenitors. T cells were marked by the expression of *tox, il2rb* and *dusp2*, NK cells were marked by *nkl.3, nkl.4* and *ccl33.3* and two distinct populations of B-cells were marked by expression of *CD37, CD79a* and *pax5* (Figure S2H, Figure 3C). Interestingly, the second B-cell population showed high levels of the immunoglobulins *ighv1-4* and *ighz*[7,9] but did not express the proliferative marker *pcna*, indicating this is a more mature or activated population of B-cells (Figure S2H). Analysis of IgM:GFP transgenic zebrafish [10] showed indeed the presence of several B-cell populations (Figure 4L-S, Figure S4G-L).

One cluster was marked by high expression of *pigrl3.5* which encodes a polymeric Ig receptor[11]. The *pigrl3.5* expressing cells are in close association in the UMAP with the NK population suggesting this is a lymphoid population (Figure S2G, H).

A small cluster with expression of *epithelial cell adhesion molecule (epcam)* was detected (Figure S2H). We hypothesize that this may be niche cells, but since this population is very small and our sorting strategy was not meant to obtain the niche cells, we have not further investigated this population.

*HSPC identification*

The most immature population would be an hematopoietic stem or progenitor cell (HSPC). These cell types are marked by little expression of lineage markers and expression of proliferation markers. In zebrafish *fli1a* and *meis1b* are typical HSC markers[3-5]. The HSPC1 and HSPC2 populations meet these criteria by expressing low levels of lineage markers (Figure 3B-D). These populations showed high expression levels of proliferation markers like *myca, pcna* and *mki67* (Figure 3J, Figure S2H) and some cells with high levels of the stem cell markers *fli1a* and *meis1b* (Figure 3F, G). Putative HSC, represented by CD41:GFP$^{low}$ sorted cells, are present in both the HSPC1 and HSPC2 clusters (Figure 3I). As expected, the GFP high expressing cells were thrombocytes. *gata2b* gene expression was enriched in the HSPC1 cluster compared to the other clusters (Figure 3H), indicating that this population would be most affected by the loss of Gata2b. Also, highest expression of *fli1a* and *meis1b* were found in HSPC1 (Figure 3F, G) indicating that this cluster contains some HSCs.

**Lineage trajectory analysis**

After defining the clusters we performed trajectory and pseudotime analysis using Monocle 3 on the integrated data set and HSPC1 cluster, both carrying Seurat embeddings[12]. Monocle 3 uses an algorithm to learn the differentiation trajectory according to the gene expression of each cell. Once the trajectory graph was learned (Figure 3K and 5H), we used get_earliest_principal_node function and chose the "root" to produce the pseudotime graph. For whole data set HSPC1 cluster was chosen as a starting point and for HSPC1 cluster quiescent subcluster was chosen to generate the pseudotime graphs (Figure 3L, 5I). We used plot_genes_in_pseudotime function to identify the gene expression in pseudotime and found that *gata2b* expression is the highest in the most immature HSPCs and decreases with differentiation (Figure 3N). We also showed that when HSPC1 cluster is chosen as a starting point, the expression of lineage specific genes such as *CD37*, *grn1* and *hbba1* were increased in pseudotime (Figure 3O-Q) which shows that HSPC1 cluster is indeed the most immature cluster within the whole dataset. Similarly, in HSPC1 cluster, when quiescent subcluster was chosen as a starting point, we found that the expression of stem cell marker *meis1b*, also *gata2b* and *CD41:GFP* were decreased in pseudotime, confirming the differentiation trajectory for these cells (Figure 5J-L).

**Table S1.**

| | **WT** | ***gata2b*$^{+/-}$** | ***gata2b*$^{-/-}$** |
|---|---|---|---|
| **EHT events** | | | |
| AGM (32 to 40hpf) | 3.3 ± 0.4 (n=18) | 2.9 ± 0.3 (n=33) | 2.2 ± 0.4 (n=18) |
| **CD41:GFP$^{+}$Flt1:RFP$^{+}$ cells** | | | |
| CHT (52-54hpf) | 8.5 ± 1.7 (n=8) | 9.0 ± 1.5 (n=20) | 11.9 ± 1.0 (n=7) |
| CHT (58-60hpf) | 25.2 ± 3.2 (n=16) | 28.7 ± 1.9 (n=28) | 23.3 ± 2.9 (n=12) |
| CHT (76-58hpf) | 55.5 ± 3.8 (n=13) | 43.4 ±5.6 (n=12) | 36.5 ± 3.0 (n=14)** |
| ***cmyb* expression intensity** | | | |
| AGM (24 hpf) | 33.0 ± 6.3 (n=14) | 28.5 ± 4.5 (n=39) | 30.6 ± 2.6 (n=13) |
| AGM (33 hpf) | 32.9 ± 10.9 (n=16) | 30.9 ± 7.9 (n=31) | 28.5 ± 6.5(n=28)** |
| CHT (56 hpf) | 10.5 ± 2.5 (n=19) | 12.8 ± 0.5 (n=41) | 4.2 ± 1.1 (n=14)** |
| CHT (76 hpf) | 21.3 ± 7.4 (n=19) | 20.8 ± 1.125 (n=43)** | 16.142 ± 1.950 (n=15)* |
| ***Runx1* expression intensity** | | | |
| AGM (26 hpf) | 24.1 ± 9.6 (n= 10) | 23.7 ± 11.6 (n= 11) | 25.4 ± 9.3 (n=11) |
| AGM (36 hpf) | 30.0 ± 7.1 (n=18) | 23.1 ± 8.1 (n=20) | 25.3 ± 6.8 (n=11) |
| **mpeg:GFP$^{+}$ cells** | | | |
| CHT (54hpf) | 513.9±23.67 (n=18) | 564.9±23.51 (n=26) | 545.9±33.27 (n=14) |
| **mpx:GFP$^{+}$ cells** | | | |
| CHT (75hpf) | 227.5±14.40 (n=11) | 221.1±8.63 (n=26) | 235.0±15.49 (n=6) |
| **LCK:GFP$^{+}$ pixel number** | | | |
| Thymus (5dpf) | 2093±186.2 (n=9) | 2449±165.5 (n=18) | 2354±206.8 (n=17) |

The data are mean ± SEM. n= number of zebrafish embryos used in analysis. AGM; aorta-gonad-mesonephros region, CHT; caudal hematopoietic tissue. *P< 0.05, **P< 0.01. If data were normally distributed we used One-way ANOVA with Tukey post-test. If data were not normally distributed we used Kruskal-Wallis with Dunn's post-test.

**Table S2. Cell numbers within the different populations in single cell analysis**

| | HSPC1 | HSPC2 | Myeloid progenitors 1 | Myeloid progenitors 2 | Myeloid progenitors 3 | Monocyte progenitor | Monocytes | Neutrophils | Trombocytes | Ery progenitors | hbba1 + klf1 | Erythrocytes | Lymphoid progenitors | B-cells | Phagocytic B-cells | Nk cells | T cells | pigrl3.5 | epcam | doublets |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WT 2 | 588 | 188 | 1110 | 300 | 146 | 33 | 1 | 216 | 39 | 331 | 6 | 301 | 105 | 44 | 58 | 164 | 21 | 5 | 3 | 16 |
| WT 1 | 498 | 110 | 822 | 31 | 109 | 59 | 30 | 142 | 55 | 112 | 168 | 1008 | 88 | 65 | 149 | 109 | 34 | 20 | 21 | 45 |
| *gata2b*$^{-/-}$2 | 846 | 102 | 410 | 61 | 71 | 119 | 41 | 55 | 19 | 202 | 9 | 148 | 267 | 201 | 238 | 99 | 60 | 20 | 9 | 14 |
| *gata2b*$^{-/-}$1 | 572 | 570 | 77 | 34 | 44 | 110 | 49 | 144 | 60 | 127 | 20 | 361 | 332 | 248 | 362 | 185 | 84 | 64 | 38 | 20 |
| Total | 2504 | 970 | 2419 | 426 | 370 | 121 | 321 | 557 | 173 | 772 | 203 | 1818 | 792 | 558 | 807 | 557 | 199 | 109 | 71 | 95 |

**Supplementary Figure 1. Maternal contribution of *gata2b* does not affect *gata2b*$^{-/-}$ survival.**
A) qRT-PCR of *gata2b* references against *elfa* on pooled WT and *gata2b*$^{-/-}$ embryos at 30 hpf (n = 3) indicating that *gata2b* expression levels are significantly reduced in mutant embryos (p = 0.0218). B) qRT-PCR of *gata2a* references against *elfa* on pooled WT and *gata2b*$^{-/-}$ embryos at 30 hpf (n = 3) shows reduced *gata2a* expression levels are in mutant embryos (p = 0.0565). C) Genotype distribution of matings between *gata2b*$^{+/-}$ and *gata2b*$^{+/-}$ zebrafish. D) Genotype distribution of matings between *gata2b*$^{+/-}$ female and *gata2b*$^{-/-}$ zebrafish males or *gata2b*$^{+/-}$ male and *gata2b*$^{-/-}$ zebrafish females. *; P<0.05.

**Supplementary Figure 2. Single cell RNA sequencing reveals several progenitor populations**
A-C) Experimental strategy to obtain single cells for RNA sequencing. B) FACS plot indicating the progenitor population which was sorted and supplemented with the remaining C) CD41:GFP$^{low}$ expressing cells from the kidney marrow pool of cells. D-~~E~~-F) Quality control parameters for each sample G) UMAP showing cluster analysis on aggregated data set of both WT and *gata2b*$^{-/-}$ cells indicating 20 different clusters with different colors. H) Heatmap showing top10 marker genes for each cluster calculated in an unbiased way. I) Genotype distribution of each of the clusters, area of the bars indicate the cell numbers in each cluster, white = WT, black = *gata2b*$^{-/-}$. Each replicate is depicted in a separate bar. J) Differentiation trajectory and pseudotime calculated only for WT cells. K) Differentiation trajectory and pseudotime calculated only for *gata2b*$^{-/-}$ cells.

**Supplementary Figure 3. *s100a10b* is expressed in the neutrophil lineage.** A) Feature analysis with gradual gene expression in shades of blue of *s100a10b.* B) WT representative images of *s100a10b in situ*

hybridization on zebrafish kidney marrow smears. A banded neutrophil is indicated by the arrow and shown enlarged in the corner. Scale bar indicates 10 µm. UMAP = Uniform manifold approximation and Projection. C) Feature analysis showing *mpx* expression in the neutrophil cluster. D) Feature analysis showing high *CD41:*GFP expression in the thrombocytes cluster and low *CD41:*GFP expression in the HSPC1 and HSPC2 clusters. E) Gating strategy for *Tg(mpx:GFP)* in WT and *Gata2b*$^{-/-}$ KM.

**Supplementary Figure 4. Differentiation markers are not altered in *gata2b*$^{-/-}$ embryos.**
A) Representative picture of *Tg(mpeg:GFP)* embryo at 54 hpf. B) Quantitation of *mpeg*:GFP$^{+}$ cells in WT, *gata2b*$^{+/-}$ and *gata2b*$^{-/-}$ embryos at 54 hpf. C) Representative picture of *Tg(mpx:GFP)* embryos at 75 hpf. D) Quantitation of *mpx*:GFP$^{+}$ cells in WT, *gata2b*$^{+/-}$ and *gata2b*$^{-/-}$ embryos. E) Representative picture of *Tg(lck:GFP)* embryos at 5 dpf F) Lck:GFP$^{+}$ area represented as pixel number in WT, *gata2b*$^{+/-}$ and *gata2b*$^{-/-}$ embryos. Each dot represent one embryo, see Table S1 for exact cell numbers and numbers of embryos analysed. G) Negative GFP control. H) WT *Tg(IgM:GFP)* zebrafish KM with 4 GFP populations gated with clear differences in size (FSC) or GFP positivity, I) Similar gating strategy for *Gata2b*$^{-/-}$ *Tg(IgM:GFP)* KM. J-L) Quantitation of ~~these~~ total IgM:GFP$^{+}$, IgM:GFP2$^{+}$ and IgM:GFP4$^{+}$ populations for WT and *Gata2b*$^{-/-}$ KM cells as percentage of single viable cells. M) Gating strategy for *Tg(lck:GFP)* WT and *Gata2b*$^{-/-}$ KM. N) Quantitation of GFP$^{+}$ cells in *Tg(lck:GFP)* WT and *Gata2b*$^{-/-}$ KM as percentage of single viable cells. O-P) Gating strategy of *Tg(mpeg1.1:GFP)* WT and *Gata2b*$^{-/-}$ KM. Q) Quantitation of GFP$^{+}$ cells in *Tg(mpeg1.1:GFP)* WT, *gata2b*$^{+/-}$ and *Gata2b*$^{-/-}$ KM as percentage of single viable cells. Error bars represent SEM. ~~*; P<0.05.~~

**Supplementary Figure 5. The HSPC1 cluster is composed of multiple HSPC subtypes.** A) Heatmap showing marker genes for each HSPC1 subcluster calculated in an unbiased way. B) Differentiation trajectory and pseudotime calculated only for WT HSPC1 cells. C) Differentiation trajectory and pseudotime calculated only for *gata2b*$^{-/-}$ HSPC1 cells.

**Supplementary Figure 6. HSPC2 shows reduced myeloid differentiation and increased in lymphoid differentiation.** A) Selection of HSPC2 for further analysis in B) subclusters (0-4) of *gata2b*$^{-/-}$ cells on the left and WT cells on the right. C) Proportion analysis of the different subclusters, indicating unequal distribution of WT and *gata2b*$^{-/-}$ cells in the individual subclusters. D) Pointrange plot showing the difference between proportion of WT and *gata2b*$^{-/-}$ cells for each HSPC2 subcluster calculated by permutation test. If FDR < 0.05, point is colored in pink and if not in grey. E) Volcanoplot showing differential gene expression analysis between WT and *gata2b*$^{-/-}$ cells showing robust downregulation of myeloid genes in *gata2b*$^{-/-}$ cells. F-J) violin plots of individual lymphoid gene expression within subclusters of HSPCs with WT cells in green and *gata2b*$^{-/-}$ cells in pink of F) *ikzf2*, G) *fcer1gl*, H) *ighv1-4*, I) *ccr9a* and J) *xpb1.*

**Supplementary references**


1. Ma D, Zhang J, Lin HF, Italiano J, Handin RI. The identification and characterization of zebrafish hematopoietic stem cells. *Blood*. 2011;118(2):289-297.
2. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411-420.
3. Macaulay IC, Svensson V, Labalette C, et al. Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell Rep*. 2016;14(4):966-977.
4. Athanasiadis EI, Botthof JG, Andres H, Ferreira L, Lio P, Cvejic A. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat Commun*. 2017;8(1):2045.
5. Tang Q, Iyer S, Lobbardi R, et al. Dissecting hematopoietic and renal cell heterogeneity in adult zebrafish at single-cell resolution using RNA sequencing. *J Exp Med*. 2017;214(10):2875-2887.
6. Carmona SJ, Teichmann SA, Ferreira L, et al. Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. *Genome Res*. 2017;27(3):451-461.
7. Moore JC, Tang Q, Yordan NT, et al. Single-cell imaging of normal and malignant cell engraftment into optically clear prkdc-null SCID zebrafish. *J Exp Med*. 2016;213(12):2575-2589.
8. Riou JF, Grondard L, Petitgenet O, Abitbol M, Lavelle F. Altered topoisomerase I activity and recombination activating gene expression in a human leukemia cell line resistant to doxorubicin. *Biochem Pharmacol*. 1993;46(5):851-861.
9. Danilova N, Bussmann J, Jekosch K, Steiner LA. The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z. *Nat Immunol*. 2005;6(3):295-302.
10. Page DM, Wittamer V, Bertrand JY, et al. An evolutionarily conserved program of B-cell development and activation in zebrafish. *Blood*. 2013;122(8):e1-11.
11. Kortum AN, Rodriguez-Nunez I, Yang J, et al. Differential expression and ligand binding indicate alternative functions for zebrafish polymeric immunoglobulin receptor (pIgR) and a family of pIgR-like (PIGRL) proteins. *Immunogenetics*. 2014;66(4):267-279.
12. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019;566(7745):496-502.

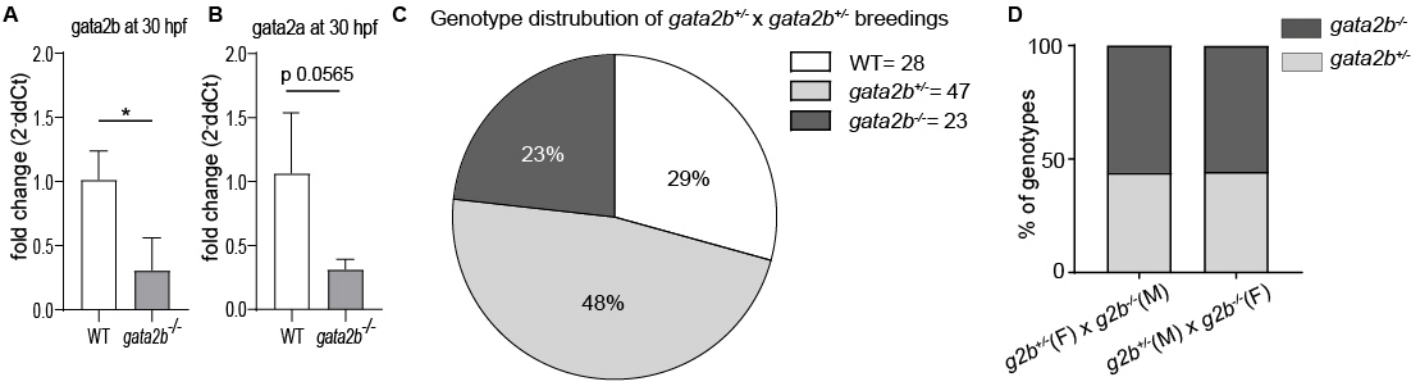



**Figure supplementary 1, related to Figure 1.**

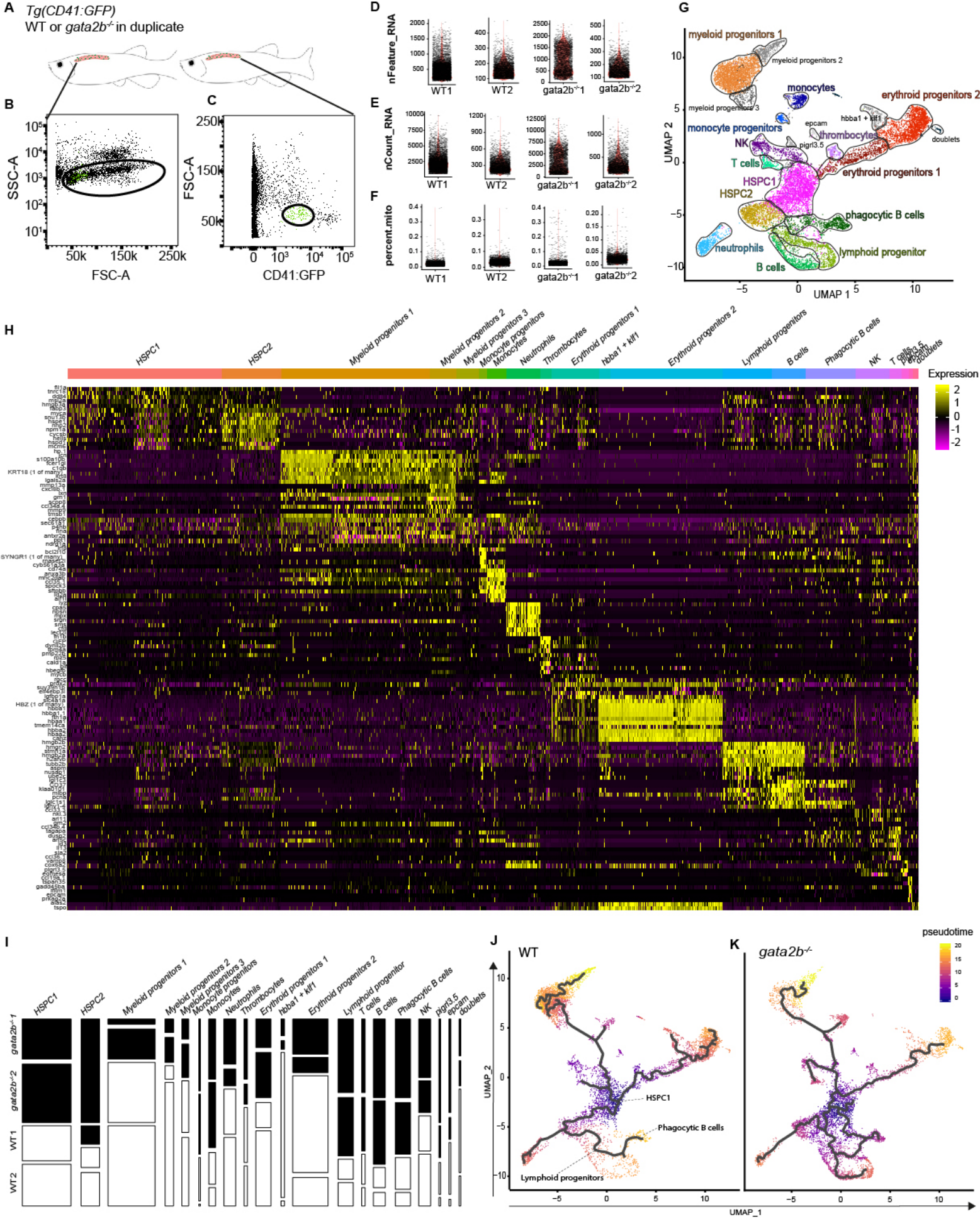



**Figure supplementary 2**

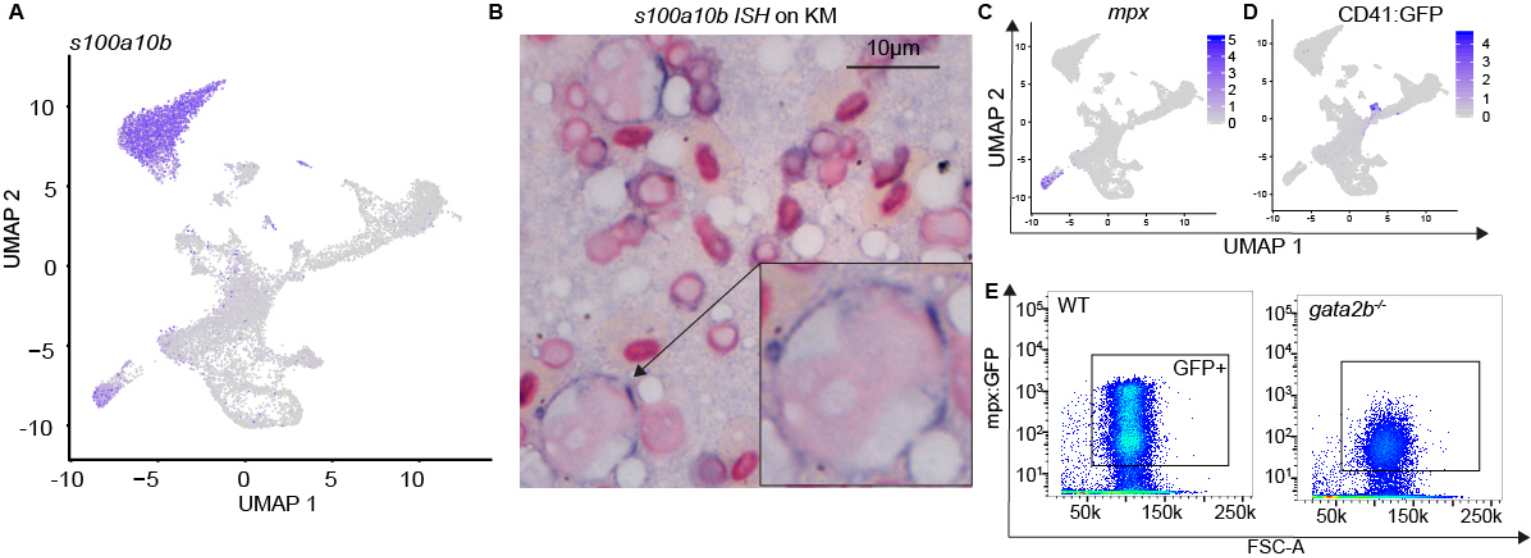



**Figure supplementary 3**

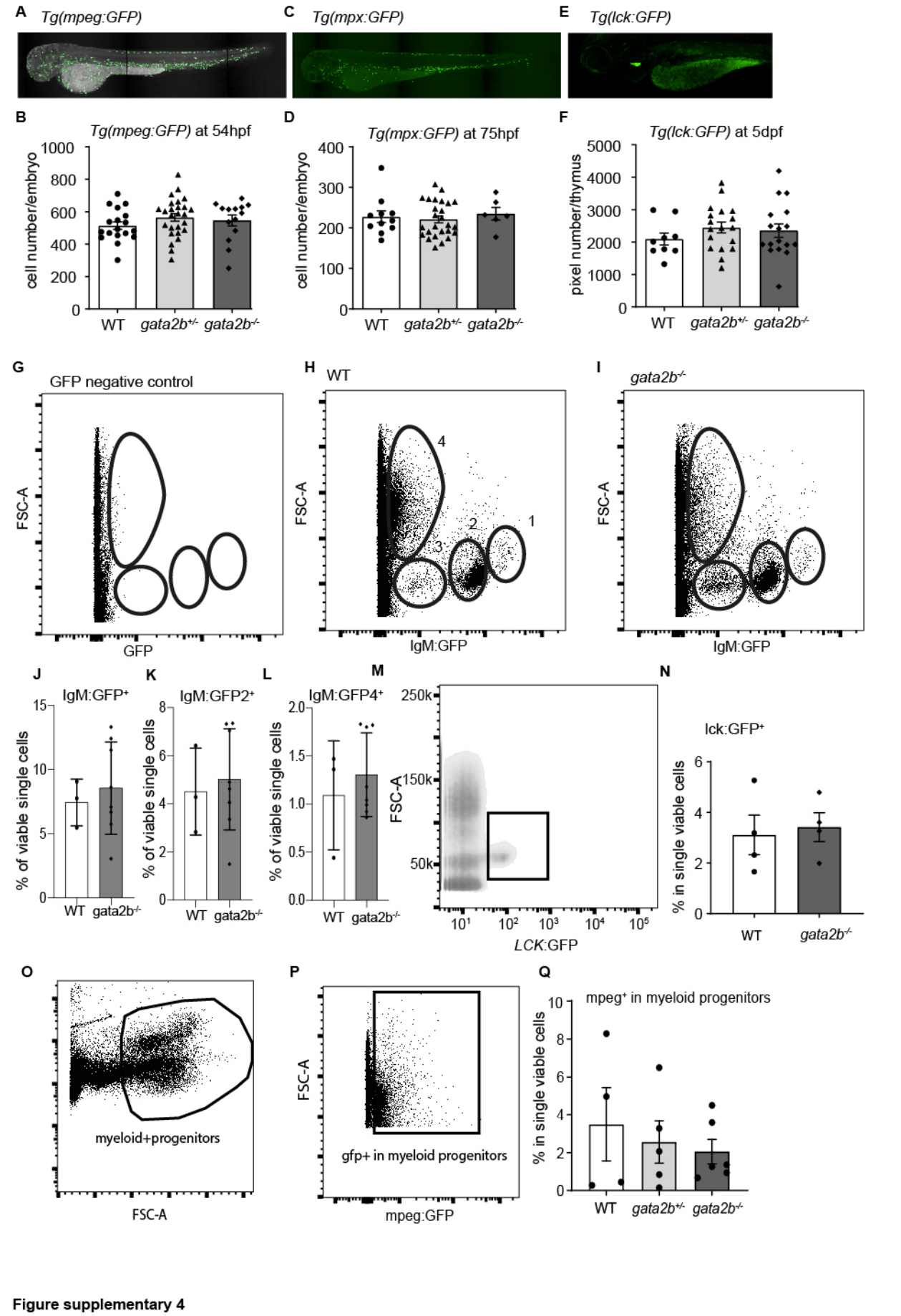



**Figure supplementary 4**

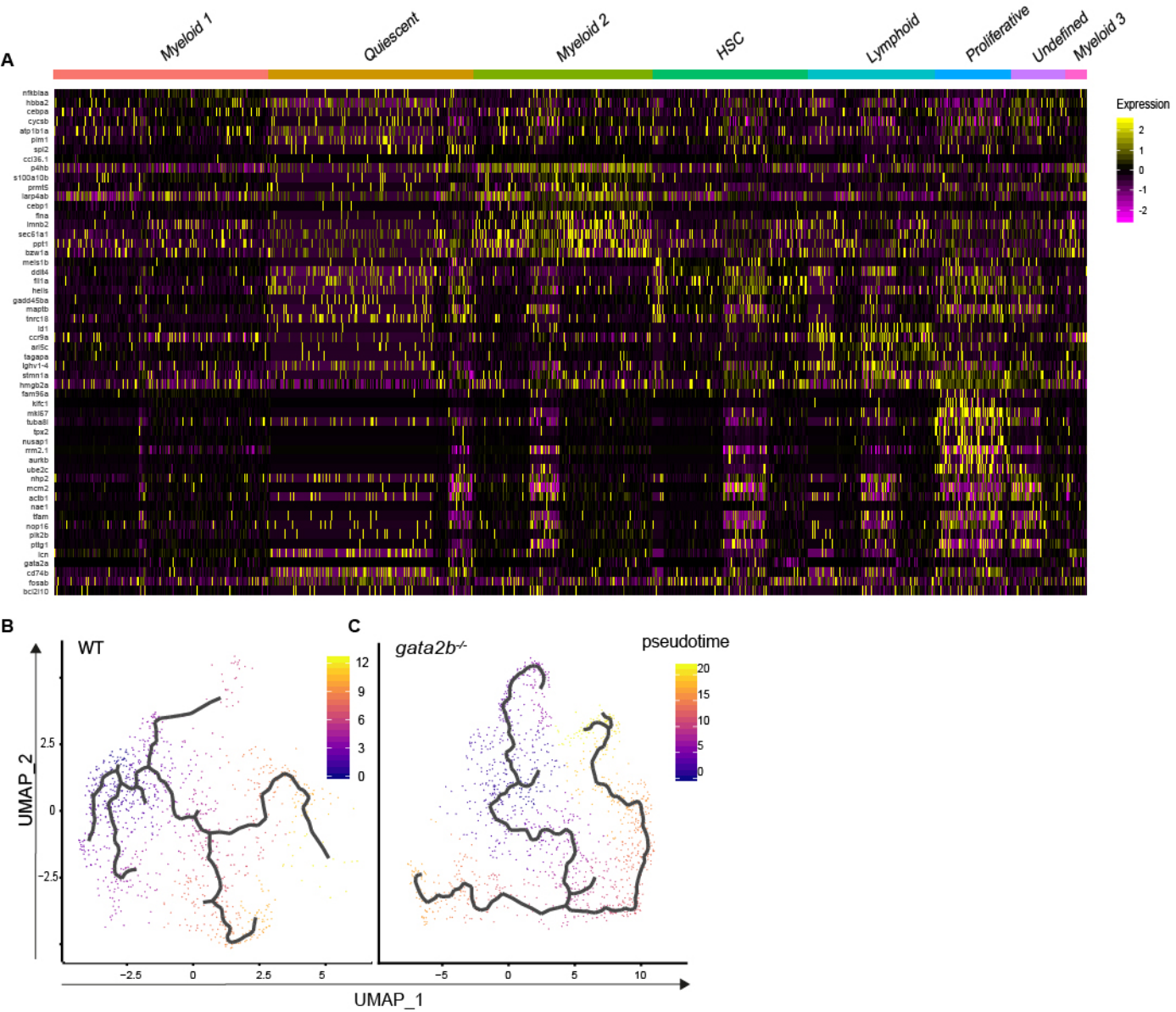
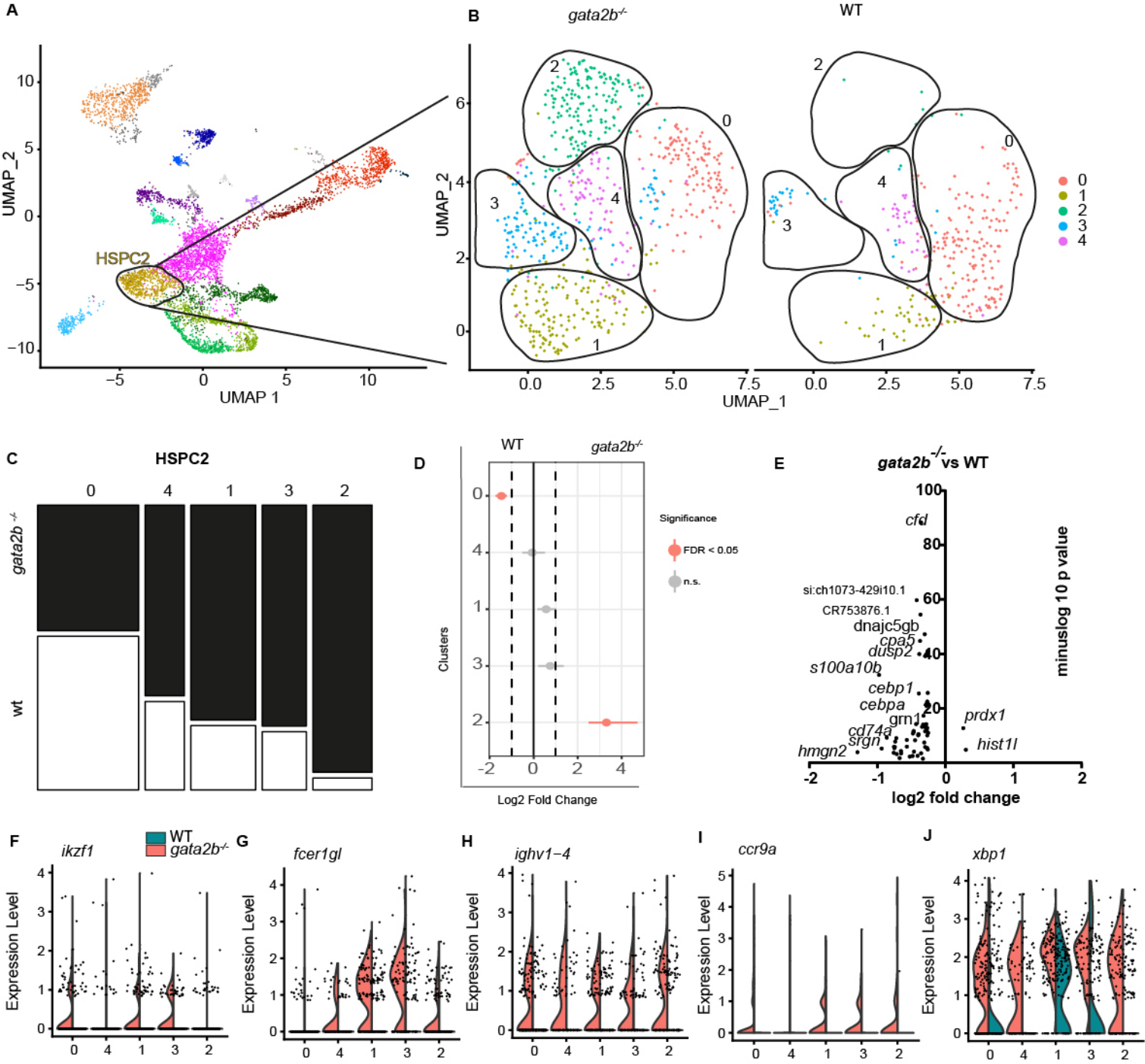



**Figure supplementary 5**




**Supplementary Figure 6, related to figure 5.**