Chapter III

Characterizing the functional classes of mosquito hemocytes

1 ScRNA-seq: a new era of cell biology

"*Omnis cellula e cellula*"
– **Rudolf Virchow**

The invention of the microscope revolutionised biological investigations. This new technology allowed Robert Hooke to publish in 1665 *Micrographia*, a collection of his microscopic observations. Among these were the depictions of the microscopic units of cork, classically considered the first description of cells. Indeed, in Latin *cella* means a 'little room with a rigid wall.' And cellular biology was born [325].

It took time however to progress from this basic definition of a cell to modern cell biology. In 1896 E.B. Wilson finally defined the cell as "the basis of life of all organisms." [326] However, the foundations for this conclusion were laid even earlier, in 1861 by Max Schultze, who recognised the importance of a cell not for the rigid wall enclosing it, but rather for what it contained. He set out his vision poetically, defining the cell as a "naked speck of protoplasm with a nucleus" (where protoplasm is now called cytoplasm) [327]. Nuclei had nevertheless been observed before, first by abbot Fontana in 1781, and then by Robert Brown in 1831, who recognised the nucleus as an essential component of cells. Finally, in 1838-9 Jakob Schleiden and Theodor Schwann formulated modern 'cell theory' for the first time, declaring "the elementary parts of all tissues to be formed of cells." [328–331] However, it was only in the 1850s through the work of Remak, Virchow, and Kölliker that cells were shown to form through scission of pre-existing cells, finally disputing the theory of spontaneous generation. Virchow went even further, showing cells not only to be the basic unit of life, but also of human pathology [332, 333].

Finally, as the 19th century came to a close, further technological advances in microscopy led to the discovery of all the major organelles we now know comprise a cell, spearheaded by work of Camillo Golgi [334]. Golgi was also responsible for disproving the theory that nervous tissue formed a completely interconnected syncytium. The development of

the 'black reaction' and the work of Santiago Cajal completely dispelled the syncytium theory and confirmed the neurons as the basic cellular unit of the brain [335, 336].

Single-cell transcriptomic techniques are now becoming just as transformative in morphing our understanding of cells, their identities, origins, and functions. Since Hooke's first observations of a cell now almost four centuries ago, generations of scientists have toiled to catalogue and describe all the different cell types in humans, animals, and plants by looking at morphology and function. Before the advent of scRNA-seq it was thought 210 different cell types existed in the human body [337]. And yet, the diversity within all of these cell types is still bewildering. Even markers traditionally thought to define individual cell types in fact isolate multiple subtler subtypes of cells. Nowadays however we are able to measure the expression level of genes in each individual cell and thus define its circuitry through single cell transcriptomics. But then, what is a cell state, and what is a cell type? When does a transcriptional perturbation define the advent of a new cell? And when is that perturbation a transition point between different cell types, and when the consequence of stochastic processes with no long-term consequences on cellular function? These are still very much active areas of investigations, but at least we now do have for the first time the tools to look anew at the cellular landscape of organisms, with a fresh set of eyes, and yet the same thirst for discovery.

We applied these technologies to mosquitoes. Three hemocyte types have been described in *Anopheles* and *Aedes* based on their morphology[4]. Granulocytes are highly phagocytic cells of about 10-20 μm, while oenocytoids are relatively smaller (8-12 μm), round cells that produce melanin, an insoluble pigment involved in wound healing and pathogen containment by encapsulation. Finally, prohemocytes are small round cells (4-6 μm) with a high nuclear to cytoplasmic ratio, thought to be precursors of the other two cell types. Hemocytes can be circulating or sessile, and alternate between these two states[146, 150]. However, the full functional diversity of mosquito hemocytes and their developmental trajectories have not been established, and it is not clear to what extent morphologically similar hemocytes are functionally equivalent. Here, we use single cell RNA sequencing (scRNA-seq)

to analyse the transcriptional profiles of individual mosquito hemocytes in response to blood feeding or infection with *Plasmodium*. We reveal a previously unknown functional diversity of hemocytes, with different types of granulocytes expressing distinct and evolutionarily conserved subsets of effector genes. And we identify two basic lineages and differentiation pathways in prohemocytes and granulocytes, and we discover new hemocyte populations and markers of immune activation. Finally, a comparison of hemocyte types from *Anopheles* and *Aedes* show that some are shared, while others appear to be unique to each mosquito species.

1.1 Aims

1. To investigate the diversity of the adult *A. gambiae* M-form (*A. coluzzi*) hemocytes in response to *Plasmodium* infection by scRNA-seq.
2. To identify markers of cell types and states and generate RNA-FISH probes and antibodies for functional studies.
3. To learn about cell lineages of hemocyte subtypes and their differentiation to functional effector subtypes.
4. To validate bioinformatic results microscopically in *A. gambiae* M-form (*A. coluzzi*) and *A. gambiae* (G3 NIH strain), and characterise hemocyte types in sections, whole-mounts and isolated hemolymph of the mosquito through RNA-FISH
5. To compare *Anopheles* hemocytes with *Aedes* hemocytes

1.2 Colleagues

Dr. Ana Beatriz Ferreira and the NIH imaging core prepared the single hemocytes RNA-FISH / morphology correlative images, and prepared *Aedes* samples up to fixed cells. Tom Metcalf aided in some of the dissections for bulk RNAseq. Mirjana Efremova calculated correlation between *Aedes* and *Anopheles* hemocytes. All other data presented is a result of my own work unless stated otherwise.

2 Methods

2.1 *A. gambiae* mosquito rearing and *P. berghei* infection

A. gambiae (G3 NIH strain) and *A. gambiae* M-form (*A. coluzzi*) were reared at 28 °C, 80% humidity, 12-hour light/dark cycle with standard laboratory procedures. For infections we utilized GFP-CON transgenic *P. berghei* (259cl2 strain), maintained with serial passage in female 4-8 weeks old BALC/c mice [319]. Parasitemia was assessed by light microscopy following methanol-fixed blood-smears stained with 10% Giemsa and air-dried. Mosquitoes were blood-fed on infected mice at a parasitemia of 3-5%, with 1-2 exflagellations per field. Infected mosquitoes were kept at 21 °C to allow for infection and midgut invasion. To confirm infection 10 mosquito midguts were dissected 5 days post blood-feeding and oocysts counted by fluorescence. *Aedes* mosquitoes were reared and challenged as of Chapter II.2.1-2.4.

2.2 Hemocyte collection, fixation, cell counting

For details of collection apparatus and collection methodology see Chapter II.2.5. Hemocytes were collected by gradually injecting in the thorax of cold-anesthetized mosquitoes 10 μL of anti-coagulant media (2 μL at a time) composed of 60% Schneider's insect media, 30% citrate buffer, 10% heat-inactivated fetal bovine serum, final pH 7.0-7.4, sterilized by 0.22 μm filtration. A total volume of 10 μL was collected per mosquito (8-12 mosquitoes per condition) and transferred with a sterile non-stick pipette tip into 500 μL vivoPHIX at room temperature. Cells were fixed for 2 hours at RT and then stored at 4C until Chromium 10X processing.

2.3 RNA extraction and bulk RNAseq library preparation

For bulk RNAseq hemocytes were collected as described above from 8 mosquitoes, but transferred directly in 500 μL of TRIZOL reagent (Invitrogen). From the same mosquitoes, midguts and carcasses were transferred into separate 1.5 mL Eppendorf tubes containing 150 μL TRIZOL reagent by Tom Metcalf. The samples were well triturated with an electrical

homogenizer and disposable pestles before adding 350 μL more TRIZOL reagent and mixing. Samples were allowed to lyse for 15-30 minutes at room temperature to allow for full dissociation, then stored at 4C overnight and then at -20C until RNA extraction. Non-hemocyte samples were then spun for 12,000 RCF, 10 minutes at 4C to remove all insoluble material. The supernatant, as well as the homogenate of hemocyte samples were transferred to Phase Lock GelHeavy 2 mL tubes that had been pre-spun for 1500 RCF for 1 minute, and allowed to incubate for 5 minutes at room temperature. 100 μL of chloroform (200 μL per 1 mL TRIZOL) was added, the tubes capped, and then vigorously shaken for 15 seconds. Samples were then centrifuged for 12,000 RCF, 10 minutes, 4C. If the clear, aqueous phase was still mixed with TRIZOL matrix then 100 μL more of chloroform was added, and the samples again mixed vigorously and spun as before. The aqueous phase was then transferred to a fresh 1.5 mL Eppendorf tube and the RNA precipitated by adding 0.25 mL of isopropyl alcohol (500 mL per 1 mL TRIZOL reagent used). For midguts and hemocyte samples 20 μL of glycogen (5 mg / mL) were also added to aid in precipitation and pelleting. Samples were mixed by repeated inversion 10 times, incubated at 10 minutes at room temperature, and then spun at 12,000 RCF, 10 minutes, 4C. All the supernatant was removed, and the RNA pellets washed twice with 75% ethanol (minimum 1 mL of ethanol per 1 mL of TRIZOL used). Each time the samples were mixed by vortexing and centrifuged 7,500 RCF, 5 minutes, 4C. At the end, the supernatant was removed and samples air-dried until almost dry, but not completely (still translucent). RNA was resuspended with 20 μL of RNAse free water for hemocyte samples, 30 μL for midgut samples, and 70 μL for carcass samples, pipetting a few times to homogenize and then incubating at 55C for 10 minutes to completely resuspend. Samples were then stored at -20C until library preparation by Bespoke Low-Throughput Team at the Wellcome Sanger institute. Total RNA quantity was assessed on a Bioanalyser and ranged from 300 ng to 39 μg. mRNA was then isolated with the NEBNext Poly(A) mRNA magnetic isolation module. RNA-seq libraries were prepared from mRNA using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) as by manufacturer instructions, except that a proprietary Sanger UDI (Unique Dual Indexes) adapters / primer system was used. Furthermore, Kapa Hifi polymerase rather than NEB Q5 was employed.

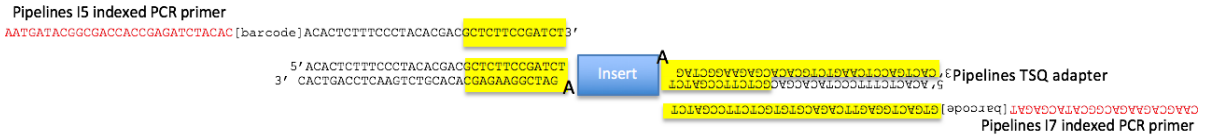

Fig. III.1 Bulk RNAseq proprietary Sanger UDI adapter / primer system. Used with NEBNext Ultra II Directional RNA Library Prep Kit.

2.4 scRNA-seq library preparation

2.4.1 Smart-seq2

See Chapter II.2.8.1 for details. 61 cells passed initial QC after Smart-seq2, as defined by wells containing a majority of sequenced reads mapping onto the *A. gambiae* genome. These cells were processed downstream as ChapterIII.2.8.2, and 48 cells passed stricter QC (>100 features per cell and <30% total reads in mitochondrial genes)

2.4.2 Chromium 10X

Fixed hemocytes were mixed with one volume of pure molecular grade ethanol before centrifugation for 30 minutes at 3k RCF at room temperature. Supernatant was discarded and pellet resuspended in pure molecular grade water before 10X Chromium scRNA-seq library processing. See Chapter II.2.8.2 for details.

2.5 Sequencing

For bulk RNAseq samples HS4000, (using kit version 1) 75PE (RNA): libraries were run on the Illumina HiSeq 4000 instrument with standard protocols using a 150-cycle kit set to a 75bp paired-end configuration. Libraries supplied at 2.8 nM and loaded with a loading concentration of 280 pM. For scRNA-seq Chromium 10X V2 and V3 kits, HS4000 (using kit version 1) 10X V2 and V3 read lengths: libraries were run on the Illumina HiSeq 4000 instrument with standard protocols using a 150-cycle kit set. As recommended by 10x Genomics an elongated reverse read was used during the sequencing run. For V2, the read lengths were as follows: Read 1: 26 bases, index 1: 8 bases, read 2: 98 bases. For V3, read lengths were as follows: Read

1: 28 bases, index 1: 8 bases, read 2: 91 bases. Libraries supplied at 2.8 nM and loaded with a loading concentration of 280 pM. For quality control, lanes passed QC if tags were decoded appropriately, reference matches were as expected either *A. gambiae* or *A. aegypti*, quality metrics met in-house expectations, other run metrics such as error rates were as expected, and yield expectation was met (given the number of cycles run and/or platform expectations). The data was then fit to the sequencing requested and any significant deviation from expected explained and appropriately annotated. For assessment two main pieces of software were used. Sequencing Analysis Viewer (SAV) was used to assess the instruments' performance. The Summary tab gave statistics for the whole run in question whereas the Analysis and Imaging tabs allowed QC to delve deeper and assess if the lanes have performed as expected across all the cycles of the run. NPG pages was used both for staff analysis and annotation, and user's visualisation of data. NPG is an in-house bespoke analysis/software package to include tag analysis, reference matching/mapping details and contamination which is the final point where lanes or tags in the run either passed or failed QC.

2.6 RNA-FISH

2.6.1 Whole mount

Mosquitoes were cold anesthetized, micro-injected with 69 nL of 16% fresh paraformaldehyde (PFA) as of Chapter II.2.2, and after 15 seconds immediately dissected while bathing in freshly prepared 4% PFA. Carcasses and midguts were separated by adding carcasses directly into an Eppendorf containing 4% PFA on ice, while midguts were quickly fixed for one minute in ice-cold fresh 4% PFA and then transferred to fresh 1X PBS where they were carefully opened along their longitudinal axis with two small gauge needles under the dissecting microscope to release the blood meal. Using the surface tension of PBS guts were gently raised up and down the PBS to release all blood from the gut until clean and then fixed in a 1.5 mL Eppendorf tube containing fresh 4% PFA. The samples were fixed overnight at 4C on a gentle rocker to guarantee good mixing and fixation. Non-stick tubes and pipette tips were used to prevent sample adhesion. In all next steps care was shown in removing solutions, as guts especially can stick onto or be sucked into pipette tips, or remain stuck on tube walls. Solutions were always

removed against a source of light to increase contrast and decrease likelihood to remove samples by error. Each wash was performed on a gentle rocker, as samples were fragile and could easily break apart.

The day after collection all PFA was carefully removed and guts and carcasses washed twice with 1mL of PBST (0.1% v/v Tween 20 in 1x PBS). Samples were then incubated for 5 minutes in a 40C rocking water-bath with 300-500 μL of RNAscope Protease Plus. After removing as much solution as possible without disturbing the samples, these were twice washed with 500 μL of probe diluent before following the RNAscope 4-plex Ancillary Kit for Multiplex Fluorescent Reagent Kit v2 technical note protocol. Briefly, the pre-mixed C1, C2, C3, and C4 probes were mixed and then 1 or 2 drops added into each sample tube and incubated for 2 hours at 40C. Samples were washed twice for 5 minutes at room temperature on a gentle rocker with pre-warmed RNAscope 1X Wash Buffer. Wash buffer had been pre-warmed to 40C for 10-20 minutes before being diluted from 50X to 1X with distilled water. Samples were then either stored overnight in 5X SSC buffer at room temperature or immediately prepared for hybridisation. 1-2 drops of RNAscope Multiplex FL v2 Amp1, Amp2, and Amp3 were added in series and incubated for 30 minutes (except Amp3 for 15 minutes) in a rocking 40C water bath. Between each reagent samples were washed twice with RNAscope 1X Wash Buffer for 5 minutes on a gentle rocker. Then Opal fluorophores were prepared at the appropriate dilutions (between 1:750 and 1:3000) and each incubated for 30 minutes in a gently rocking water bath at 40C in the dark. Before adding each Opal, samples were treated with the corresponding RNAscope Multiplex FL v2 HRP-C(1/2/3/4) for 15 minutes in a gently rocking water bath at 40C in the dark. Then, samples were treated with RNAscope Multiplex FL v2 HRP-Blocker for 15 minutes in a gently rocking water bath at 40C in the dark. Between all these steps samples were washed twice with RNAscope 1X Wash Buffer for 5 minutes on a gentle rocker in the dark. Finally, as much wash buffer was removed before adding 1-2 drops of DAPI for 30 seconds. DAPI was then in turn removed and samples added onto a slide with 1 drop of Prolong Gold antifade reagent. The samples were flattened in the Prolong Gold reagent (important: without DAPI or background fluorescence will be high) under a dissecting

microscope to prevent flaps and folding of the tissue. After adding coverslips corners were sealed with transparent nail polish and the samples let dry overnight at room temperature in the dark. The day after nail polish was added all around the slide to seal the samples. These were then stored at 4C in the dark until imaging.

Probes	Channel	Dilution	Amount	Annotation
General				
AGAP009623	C1	1:1500	Std	GAPDH - mosquito + control
AGAP008296	C2	1:3000	1/2	Trypsin - gut
AGAP004203	C2	1:3000	1/2	Vitellogenin - fat body
Hemocytes / Granulocytes T. I and II				
AGAP004017	C4	1:1000	1.5	LRR. All hemocytes' marker
AGAP011974	C4	1:1000	Std	SCRC1. General hemoc. marker
AGAP000790	C3	1:1000	Std	Prohem. / granulocyte marker
AGAP003057	C1	1:1000	Std	Gran. Type II
AGAP011871	C2	1:750	Std	Gran. Type I
Rapidly dividing				
AGAP005363	C3	1:750	n/a	
Fat Body - Baseline				
AGAP007033	C1	1:750	n/a	
AGAP028406	C1	1:750	n/a	APL11C
Oenocytoids				
AGAP004981	C2	1:1500	Std	PPO4
AGAP012851	C1	1:1500	Std	Aldo-keto-reductase
AGAP012000	C3	1:1500	Std	Fibrinogen/fibronectin
Effector				
AGAP007318	C3	1:1000	1.5	Transmembrane
Secretory				
AGAP011239	C1	1:1500	Std	Some also in oenocytoids

Table III.1 RNAscope probe channels and Opal dilution for whole-mounts and sections. See RNAscope 4-plex Ancillary Kit for Multiplex Fluorescent Reagent Kit v2 technical note protocol for details. 'Amount' column indicates the ratio of probes added to hybridization mix compared to standard protocol. 'Std' indicated standard, 0.5 is half of standard. 'n/a' indicates a probe was not successful even with the strongest Opal dilution (1:750) and highest probe amount. Note all dilutions were 1:750 for RNAscope of isolated hemocytes.

2.6.2 Isolated hemocytes

Wells of μ-Slide Angiogenesis Chambers (Cat# 81506 from IBIDI) were coated with 3.5 μg / cm^2 of Cell-Tak Cell and Tissue Adhesive (Corning, 734-1081) by first preparing a fresh 300 μL coating solution with 10 μL Cell-Tak, 285 μL Sodium Bicarbonate pH 8.0 and 5 μL 1N NaOH and immediately coating the glass slides. Wells were incubated at room temperature for least an hour, after which they were washed with sterile water, air-dried and stored at 4C for a maximum of one day.

Hemocytes were collected as of above but directly onto the wells. Eight mosquitoes were processed per sample. Hemocytes were then let to attach onto the coated wells for 15 minutes at 28C in an incubator, before removing all of the media, and fixing cells with 4% PFA for an hour at room temperature before proceeding to RNA-FISH protocol as of Chapter III.2.7.1. The process was made easier by not having to take care of aspirating tissue with the washes, however care was shown not to disperse liquid to strongly, but to always do it gently on the sides of the well to prevent cell detachment. Dr. Ana Beatriz Ferreira performed the isolated *P. berghei* experiments and the correlative experiments.

2.6.3 Sections

Mosquitoes were cold anesthetized, dipped in 100% ethanol to decrease surface tension, and then dipped and fixed in 10% formalin for 18-24 hours overnight at room temperature. Following that the Histology Core of the Sanger Institute processed the samples to make slides. The Sakura Tissue-Tek VIP Tissue processor on Rapid Biopsy programming was used (10 min VIP1 and 10 min VIP2 for each solution except: no VIP2 for 50% and 70% ethanol; first paraffin wax 20 min for both VIP1 and VIP2), with the following solutions in order: 50% ethanol, 70% ethanol, 90% ethanol, 3X 100% ethanol, 3X xylene, and 4x wax. For embedding, two orientations (longitudinal and transverse) were used for each condition (sugar-fed, blood-fed and *P. berghei* infection), before 5 μm sectioning. H&E sections were prepared for every other section, with the mirror section available for RNA-FISH (RNAscope) as of above.

2.7 Imaging

Mosquito sections and whole mounts were imaged with the 3DHISTECH MIDI II automatic digital slide scanner (3DHISTECH, Budapest, Hungary), with 20x and 40x objectives (numerical aperture 0.8 to 0.95), and a bespoke DAPI, Opal 520, 570, 620 and 690 filter sets and a 4.2MP 16-bit camera with wideband LED, or with a 20x bright-field camera for H&E mosquito sections and a 4.2MP 16-bit camera with RGB illumination. Sections and whole-mounts were imaged with extended focus, sequential acquisition, and variable z-steps, mosaic size and integration.

For whole-mount and hemocytes samples images were captured at the National Institute of Health using a Leica TCS SP8 DMI8 confocal microscope (Leica Microsystems, Wetzlar, Germany) with a 20x, 40x and 63x oil immersion objective (using zoom factor of 2, 3 or 4; numerical aperture, 1.25 to 1.4) equipped with a photomultiplier tube/hybrid detector. Samples were visualized with a white light laser and specific emission and excitation range were used depending on the fluorophore used. For these experiments we used the following spectra for excitation/ emission: 488/520, 550/ 570 594/620, and 670/690. DAPI was excited using a 405-nm diode laser. Images were taken using sequential acquisition, and variable z-steps, mosaic size and integration. Image processing was performed using proprietary Leica LAS X and Imaris 9.2.1 (Bitplane, Concord, MA, USA). At the Wellcome Sanger Institute images were captured using a Leica TCS SP8 DMI8 confocal microscope (Leica Microsystems) using a 40×, 63×, or 100× oil immersion objective (using zoom factor of 2, 3 or 4; numerical aperture, 1.25 to 1.4) and equipped with photomultiplier tube/hybrid detectors. Fluorochromes were excited using a 405nm DMOD laser for DAPI, 488-nm CSU laser for Opal 520, a 552-nm CSU laser for Opal 570 and Opal 620, 638-nm CSU laser for Opal 690. Images were taken using sequential acquisition, and variable z-steps, mosaic size and integration. Image processing was performed using proprietary Leica LAS X and Imaris 9.2.1 (Bitplane, Concord, MA, USA).

2.8 Bioinformatics

2.8.1 Bulk RNA-seq

Sequencing reads in CRAM format were fed into a bespoke BASH pipeline to first automatically convert cram files to fastq using biobam's bamtofastq program (Version 0.0.191). Then, forward and reverse fastq reads in paired mode were aligned to the *A. gambiae* AgamP4.3 reference genome using hisat2 (Version 2.0.4) and featureCounts (Version 1.5.1) with recommended settings. Combined counts matrix was then produced by a python script before downstream data processing and analysis within R version 3.5.3 (RStudio version 1.0.153). Downstream normalization, differential expression analysis and visualization were done with DESeq2 R package (Version 1.18.1) [280]. Base factor was defined as the sugar condition, and time 0 (non-infected). One outlier was removed (blood fed hemocyte sample at 48 hours, experiment GR88) after plotting residuals of internal batch correction and visually inspecting a PCA plot. Data was normalized by making a scaling factor for each sample. First the log(e) of all the expression values were taken, then all rows (genes) were averaged (geometric average). Genes with zero counts in one or more samples were filtered out and the average log value from log(counts) for all genes was subtracted. Finally, the median of the ratios calculated as above for each sample was computed and raised to the e to make the scaling factor. Original read counts were divided by the scaling factor for each sample to get normalized counts. Then, the dispersion for each gene was estimated, and a negative binomial generalized linear model fitted. P values for the differential expression analysis were adjusted for multiple testing using the Bonferroni correction. Genes were considered as differentially expressed if they had an adjusted P value < 0.001 (Wald T-test) and a $\log_2$ fold change > 2. All body parts, conditions and timepoints were considered together while running the following model for differential expression analysis focused on body part, with experimental repeats, time, and effects of treatment (*P. berghei*, blood feeding and sugar feeding) as covariates:

```
ddsMat <- DESeqDataSetFromMatrix(countData = countdata, colData = coldata,
                                 design = ~ 0 + experiment + time + treatment + part)
```

2.8.2 scRNA-seq

Droplet-based sequencing data were aligned and quantified using the Cell Ranger Single-Cell Software Suite [246] (version 2.0, 10x Genomics) against the *A. gambiae* PEST, AgamP4.9 reference genome provided by Vectorbase [338]. Cells with fewer than 100 and more than 2500 genes and for which total mitochondrial gene expression exceeded 20% (or 50%) were removed. Genes that were expressed in fewer than three cells were also removed.

Downstream analyses—such as normalization, shared nearest neighbor graph-based clustering, differential expression analysis and visualization—were performed using the R package Seurat (version 2.3.4 or 3.0.2) [256, 277, 339]. The two experimental batches were integrated using canonical correlation analysis, implemented in the Seurat alignment workflow. In the newer Seurat version, batches were integrated with a hybrid CCA / MNN strategy identifying 'anchors' of similar cells between conditions and CCs. Cells for which the expression profile could not be explained by low-dimensional canonical correlation analysis compared to low-dimensional principal component analysis were discarded. Clusters were identified using the community identification algorithm as implemented in the Seurat 'FindClusters' function. For Seurat V2 the shared nearest neighbour graph was constructed using 13 canonical correlation vectors as determined by the dataset variability. The resolution parameter to obtain the resulting number of clusters was fine-tuned so that it produced a number of clusters large enough to capture most of the biological variability. UMAP analysis was performed using the RunUMAP function with default parameters. Differential expression analysis was performed based on the Wilcoxon rank-sum test. The P values were adjusted for multiple testing using the Bonferroni correction. Clusters were annotated using canonical cell-type markers. We remove a blood-fed 24 hours post-feeding sample (experiment GR72) because it formed a technical outlier in the initial PCA-driven quality control and all cells clustered separately without mixing with other samples. Some clusters were further analyzed by partitioning the clusters separately and performing the analysis anew, with the same

alignment and clustering procedure. For example, all hemocytes were subdivided from other non-hemocyte cells and reanalyzed.

Diffusion pseudotime [340] implemented in the SCANPY package [257] was applied to find the major non-linear components of variation across cells, using the most highly variable genes. The first diffusion component correlated with oenocytoids identity as defined by know marker genes, whereas the second diffusion component second diffusion component correlated with immune activation and cell division. Genes which changed along the identified trajectories (diffusion components) were identified by performing a likelihood ratio test using the function differentialGeneTest in the monocle 2 package [341]. The Seurat implementation of velocyto [342] was then applied to estimate RNA velocity and infer in which direction cells were changing along the previously inferred trajectories or UMAP. scVelo was used as an additional RNA velocity analysis tool to confirm the results [343].

Lineage tree reconstruction was performed with partition-based graph abstraction (PAGA) as implemented in SCANPY package [344]. The graph abstraction algorithm combines clustering and trajectory inference to elucidate the variability of scRNA-seq through discrete and continuous variables. PAGA takes into consideration a partitioned graph of neighbourhood relations. It quantifies distances between nodes with a random-walk based measure and then it quantifies what connectivity partitions there is. The abstracted graph is anchored on nodes which are the clusters first identified with Seurat. The differentiation tree is a tree-like subgraph which best explains topology. Slingshot was another highly rated lineage tree reconstruction software that we used to validate PAGA results [309]. With a matrix input representing cells in a reduced-dimensional space (UMAP) and a vector of cluster labels the Slingshot algorithms then built a minimum spanning tree (MST) of the clusters to infer the lineage structure. Finally, smooth lineage curves were built and pseudotime inferred for all lineages. We then used the pseudotime values calculated by Slingshot to discover differentially expressed genes between the identified lineages with the tradeSeq package (TRAjectory Differential Expression analysis for SEQuencing data) [345]. TradeSeq uses pre-calculated UMAP coordinates and pseudotime values to fit generalized additive models (GAMs).

To compare the *A. gambiae* with the *Aedes* cell types, a logistic regression with L2-norm regularization and a multinomial learning approach (implemented by the scikit-learn function LogisticRegression) was trained on the anopheles gambiae clusters. The log-transformed normalized data was used. The model was used to predict the probabilities of each *Aedes* cell belonging to each one of the anopheles gambiae clusters (implemented by the predict_log_proba function).

3 Results

Hemocytes were obtained from mosquitoes at different states of immune activation in order to survey their diversity. In the first experiment we collected mosquitoes at both 24- and 27-hours post-infection to potentially gain information about the early hemocyte response to *P. berghei*. The 48- and 72-hours timepoints were chosen to explore hemocyte changes after infection. In the second experiment, the 27 hours timepoint was removed to make space (cost concerns) for a day 7 timepoint, which we hypothesised could give information on hemocyte deactivation. We chose sugar feeding as baseline control. However, we also used blood feeding as control for *P. berghei* infection due to the large changes blood feeding causes in the mosquito.

Experiment 1	**Day 0**	**Day 1 PF**		**Day 2 PF**	**Day 3 PF**
Condition		*24 h*	27 h	48 h	72 h
Cntrl (SF)	SF	X	X	X	Bleed
Cntrl (BF)	BF	Bleed	Bleed	Bleed	↓
P. berghei	BF	↓	↓	↓	↓

Experiment 2 and bulk	**Day 0**	**Day 1 PF**	**Day 2 PF**	**Day 3 PF**	**Day 7 PF**
Condition		*24 h*	48 h	72 h	7 days
Cntrl (SF)	SF	Bleed	Bleed	Bleed	Bleed
Cntrl (BF)	BF	↓	↓	↓	↓
P. berghei	BF	↓	↓	↓	↓

Table III.2 Experimental strategy: bulk and scRNAseq of *Anopheles*. PF = post-feeding; BF = blood-feeding. Experiment 1 refers to scRNA-seq repeat 1. Experiment 2 was the second scRNA-seq repeat and the same scheme was used for the bulk RNAseq samples.

Following hemocyte capture and 10X library preparation and sequencing we then normalized and performed QC on all cells from an experiment together, then batch corrected the experiments, clustered, and investigated differences between clusters, time points, and conditions as of below and method chapter.

3.1 scRNA-seq identifies at least six hemocyte subpopulations

3.1.1 QC of Chromium 10X single cell data

Processed scRNA-seq matrices from each individual sample were loaded onto the R-based Seurat (v2.4 or v3.0) analysis suite. First, cells were filtered based on QC metrics to remove poor quality cells. The total number of genes (or of UMIs) within a cell is traditionally considered a useful marker to distinguish low quality cells or empty droplets from healthy cells. In addition, an excessive gene count can indicate that the original droplet contained a doublet or multiplet and should also be excluded. Cells were thus filtered if they were found to have less than 100 or more than 2500 unique genes. Then, we identified which *Anopheles* genes map to the mitochondrial genome to calculate the percentage of reads mapping to mitochondrial genes. Typically (though not necessarily always) damaged, dying, and low-quality cells will show a high ratio of mitochondrial reads to total reads. In our data-set we initially excluded all cells that had more than 20% of total reads mapping onto the mitochondrial genome. We repeated this process for both our scRNA-seq experiments, plotting data both with violin plots and scatter plots to identify outlier cells. We discarded outlier samples: blood-fed 24 hours (experiment 1 and 2), sugar-fed 48 hours (experiment 2).

Filtering appeared successful in removing all outliers, with each parameter showing a compact distribution in both experiments [Fig. III.2]. The first experiment had a total of 7762 cells before QC, with means of 85 genes and 221 UMIs per cell. After QC we were left with 2081 cells (mean of 180 genes per cell, and 575 UMI per cell). In the second experiment before QC we had a total of 3883 cells, with a mean of 380 genes per cell and 1422 UMIs per cell. After QC 3162 cells remained, with a mean of 441 genes per cell and 1516 UMIs per cell. Statistics showed the first experiment had lower data quality than the second. Of note, samples from the first experiment had been stored for about a month at 4C while the second experiment was processed within a week of collection.

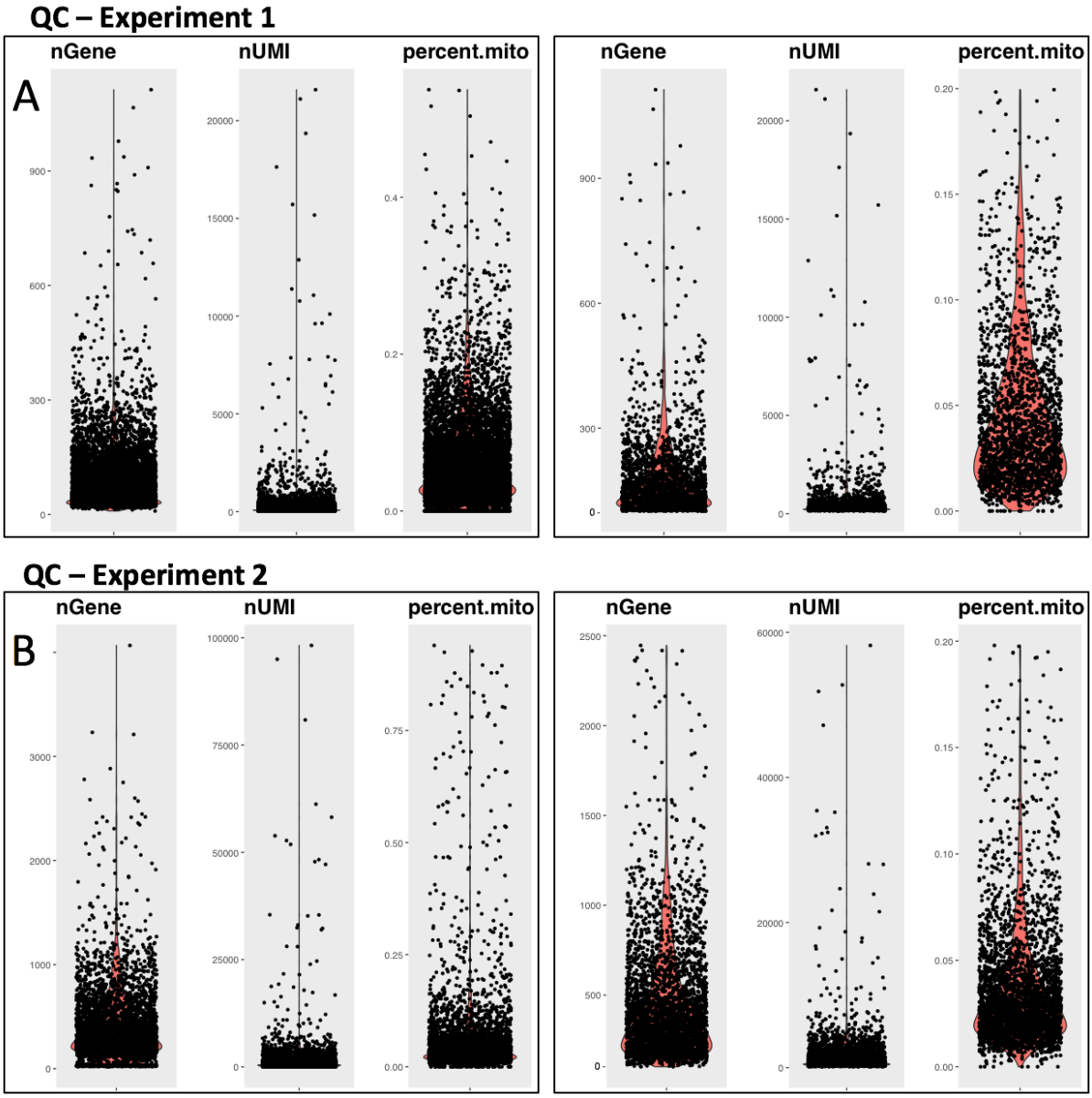

Fig. III.2 Seurat scRNAseq QC. (A) QC metrics for the first experiment. To the left metrics before QC, to the right after QC. **(B)** QC metrics for the second experiment. To the left metrics before QC, to the right after QC. nGene = total number of genes detected per cell. nUMI = total number of UMIs detected per cell. percent.mito = the proportion of total reads mapping to mitochondrial genes.

3.1.2 Normalisation, scaling, identification of variable genes, and PCA

Data was then normalized using the Seurat global-scaling normalization method, which normalizes gene expression data of our cells by total expression, multiplies it by a scale factor of 10,000, and then takes the natural logarithm of the resulting number. Highly variable genes (focus of downstream analyses) were calculated with a variance stabilizing transformation (VST) [277, 339]. We identified 2000 variable genes in each experiment. We then linearly transformed the data ('scaling') to pre-process data for dimensionality reduction techniques such as PCA, the first step of an integrated analysis. Scaling reduced the importance of highly expressed genes. This step shifted gene expression so that the mean across cells is zero, and scaled expression so that variance across cells is 1. Many of these highly variable genes were common among the two experiments. For instance, AGAP011294, AGAP01002, or AGAP011230 were identified as top variable genes in both [Fig. III.3].

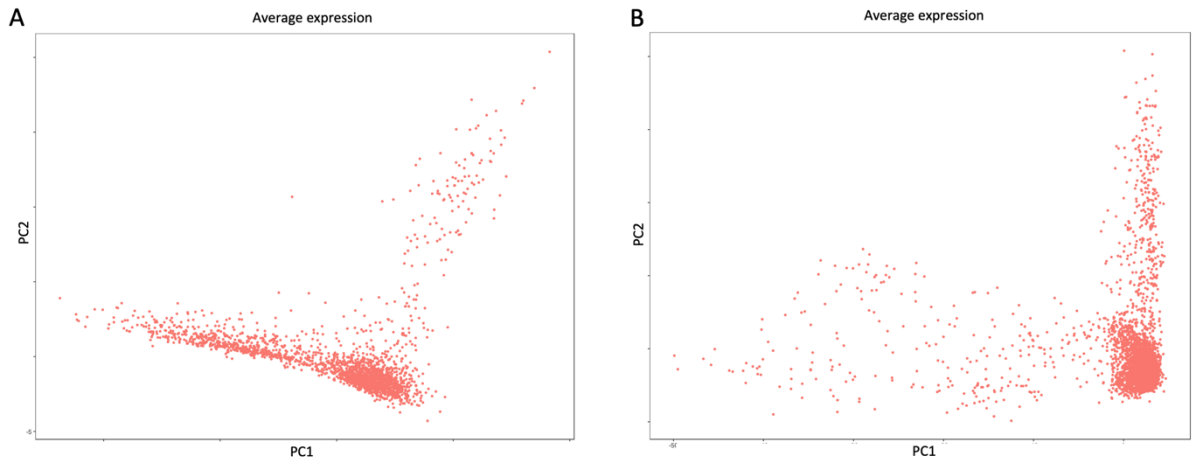

Fig. III.3 PCA profiles are similar between the two experiments (A) PCA showing the first two principal components for first experiment **(B)** and PCA of the two first principal components for the second experiment

3.1.3 Clustering reveals 9 separate cell types

In Seurat 3.0, dataset aggregation was drastically improved by using mutual nearest neighbours (MNN) – 'cell anchors' – in addition to canonical correlation. Different QC parameters returned the same results and so we lowered stringency of mitochondrial gene filtering to 50% (see discussion). After aggregating the two experiments we had a total cell count of 5383 hemocytes after QC, with a mean of 335 genes per cell, and 1142 UMI per cell. We classified *Anopheles* cell types in the hemolymph to identify nine major clusters. Most clusters could be further subdivided into smaller clusters by increasing the resolution of the clustering algorithm. However, increasing resolution typically identifies cell states rather than cell types and initial clustering therefore needs to be more conservative.

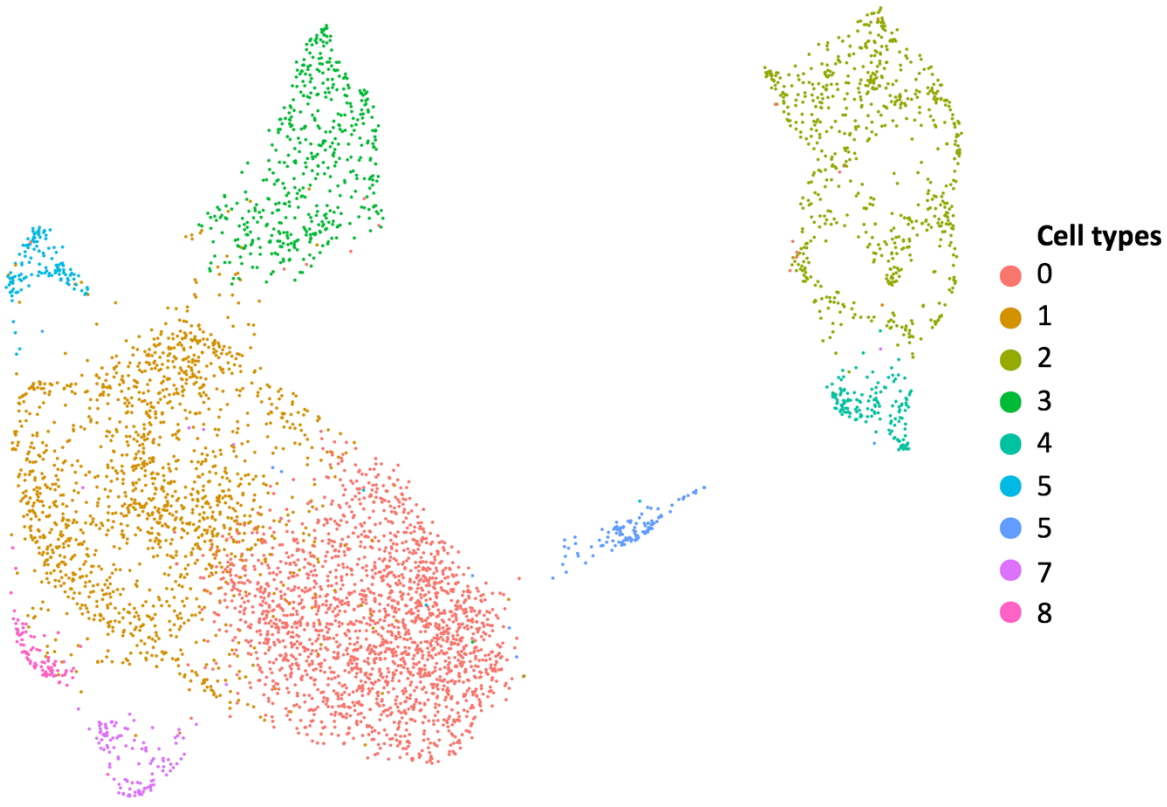

Fig. III.4 Clustering solution of *A. gambiae* hemocytes. UMAP dimensionality reduction separates clusters of cells by overall transcriptomic similarity. Each dot represents a cell, whereas different colors identify clusters of similar cells.

3.1.4 Varying QC parameters does not alter clustering solution

Compared to simple CCA integration of Seurat v2.4 the v3.0 clustering solution was well mixed with regards to both experimental batches as well as individual samples [Fig. III.5A-B].

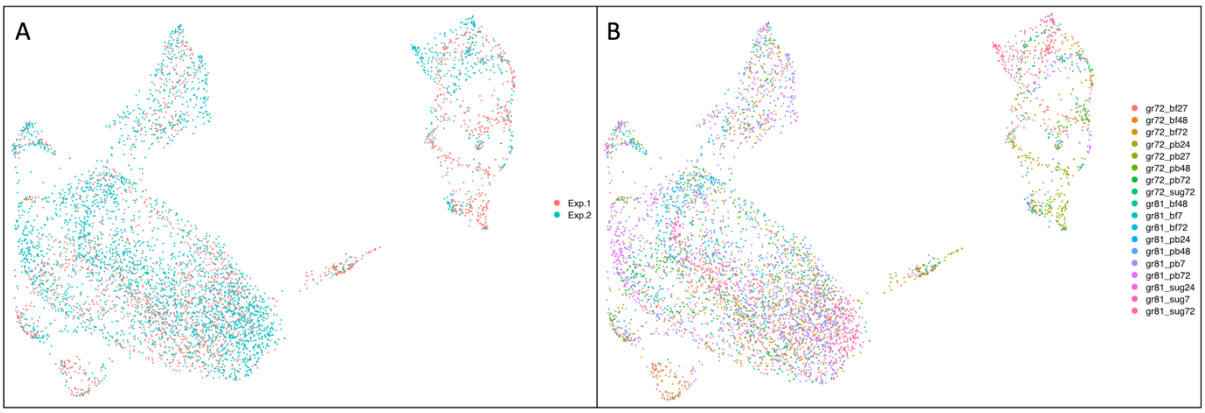

Fig. III.5 Samples and experiments are well-mixed. (A) Both between the two experiments, as well as **(B)** between samples (separate 10X lanes and chips)

The new clustering strategy is robust to a wide spectrum of parameters and is more unsupervised, lowering the risk of bias due to parameter selection. We nevertheless manually checked whether results were reasonable by raising the minimum number of genes per cell to 150 and then to 200, without changes to cluster numbers, structure or markers genes [Fig. III.6].

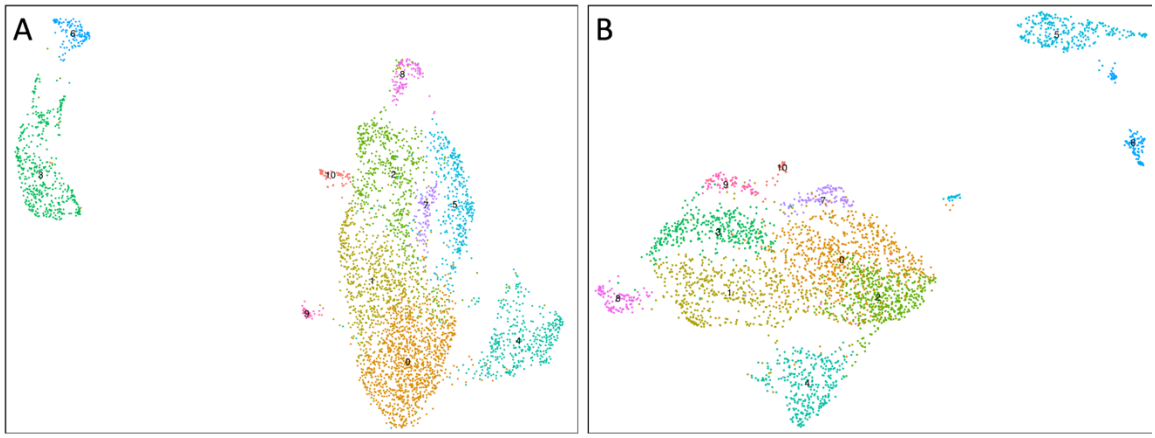

Fig. III.6 Clustering solutions are robust to gene thresholding. Manual QC iteration: increasing minimum gene per cell parameter stringency does not alter computer clusters. **(A)** Minimum 150 genes per cell **(B)** Minimum 200 genes per cell.

Functional classes of mosquito hemocytes

We then removed mitochondrial genes thresholding. Few cells were added and no changes in clustering were detected [Fig. III.7A]. Finally, we compared cells (droplets with more than 100 genes) and background (droplets with less than 50 genes) with principal component analysis. Without calculating a UMAP, already the first two principal components cells and debris clearly separate. Combined, the QC tests demonstrate our thresholds are reasonable for this dataset [Fig. III.7B].

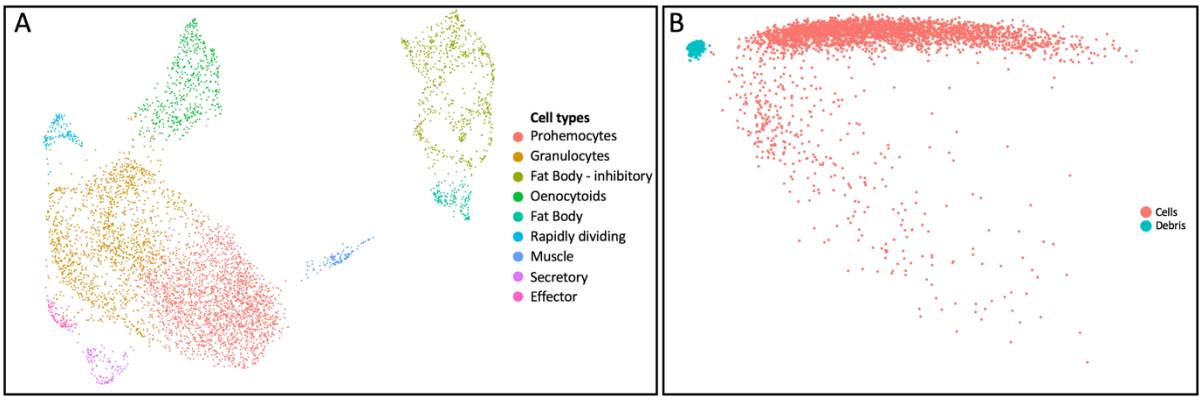

Fig. III.7 Clustering solutions is robust to more stringent mitochondrial filtering. Debris and cells are clearly identifiable. Clustering done as above, except threshold was set with **(A)** maximum 100% of reads mapping to mitochondrial genes, showing no changes **(B)** Principal component analysis of debris (blue, droplets with less than 50 genes per droplet) and cells (red, droplets with more than 100 genes per droplet) shows cells separate clearly from debris (PC1 vs PC2).

3.1.5 Differential expression analysis identifies conserved marker genes for each cell cluster, and suggest cellular identity

Though the *Anopheles* genome is poorly annotated we utilised gene ontology annotations from g:Profiler [346], as well as manual curation of *Anopheles* genes [347], to understand the identity of each cell cluster. The table below shows the top 10 genes for each cluster, annotated, while the full list can be found in the Appendix.

Cluster 0

Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP012100	RpS26	5.21E-87	0.325	0.97	0.98	40S ribosomal protein S26
AGAP002464	-	9.33E-75	0.471	0.95	0.90	secreted ferritin G subunit
AGAP011828	Cp1	1.00E-71	0.498	0.83	0.70	cathepsin L
AGAP010163	RpL38	2.29E-68	0.322	0.95	0.96	60S ribosomal protein L38
AGAP000305	-	6.01E-58	0.383	0.88	0.70	SPARC
AGAP004936	-	5.04E-50	0.428	0.79	0.62	None
AGAP007740	RpLP1	4.04E-45	0.258	0.96	0.97	60S ribosomal protein LP1
AGAP002422	CLIPD1	2.74E-41	0.656	0.61	0.54	CLIP-domain serine protease
AGAP011119	-	1.73E-40	0.421	0.74	0.62	None
AGAP002465	-	1.54E-36	0.421	0.82	0.77	ferritin heavy chain

Cluster 1

Gene	Name	Pval_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP011228	-	2.12E-189	0.746	0.99	0.75	None
AGAP007312	-	7.96E-162	0.799	0.77	0.35	None
AGAP004936	-	1.16E-142	0.596	0.92	0.59	None
AGAP006278	-	3.23E-137	0.666	0.86	0.53	None
AGAP000651	actin5c	2.72E-136	0.713	0.78	0.39	Actin-5C
AGAP004017	-	8.90E-129	0.590	0.82	0.41	None
AGAP004164	GSTD1	1.58E-125	0.704	0.44	0.13	glutathione S-transf del. c1
AGAP028028	lrim16a	1.70E-121	0.593	0.82	0.44	leucine-rich immune prot
AGAP004016	-	2.29E-119	0.557	0.69	0.29	None
AGAP006367	-	2.62E-118	0.869	0.33	0.08	None

Functional classes of mosquito hemocytes

Cluster 2

Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP010968	CLIPA9	0	2.460	0.48	0.04	CLIP-domain serine protease
AGAP013060	-	0	1.976	0.66	0.09	None
AGAP012571	-	0	1.943	0.78	0.17	None
AGAP008011	-	0	1.902	0.48	0.04	None
AGAP003473	-	2.70E-303	3.031	0.85	0.27	None
AGAP003474	-	1.54E-298	2.450	0.99	0.95	None
AGAP005888	-	1.20E-295	1.828	0.96	0.53	None
AGAP008004	-	7.26E-291	2.367	0.89	0.37	None
AGAP004674	-	1.01E-278	2.010	0.38	0.02	Phenoloxidase inhibitor prot
AGAP009527	-	2.92E-272	2.043	0.61	0.10	None

Cluster 3

Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP004978	PPO9	0	4.469	0.81	0.12	prophenoloxidase 9
AGAP011223	-	0	4.448	0.84	0.11	None
AGAP006258	PPO2	0	4.364	0.79	0.13	prophenoloxidase 2
AGAP004977	PPO6	0	4.055	0.98	0.34	prophenoloxidase 6
AGAP012616	PPO5	0	3.961	0.83	0.08	prophenoloxidase 5
AGAP012851	-	0	3.829	0.74	0.02	Aldo-keto reduct fam 1,C3
AGAP006570	-	0	3.669	0.73	0.11	myo-inositol-1(4)-monoph
AGAP006743	-	0	3.489	0.63	0.03	None
AGAP000162	-	0	3.471	0.80	0.06	Cystathionine beta-synth
AGAP000679	-	0	3.159	0.98	0.36	Aminoacylase

Cluster 4

Gene	Name	Pval_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP004203	Vg	2.94E-162	2.998	0.78	0.10	vitellogenin
AGAP007940	-	9.56E-127	2.767	0.72	0.11	Reticulon-like protein
AGAP006548	-	1.20E-126	2.565	0.91	0.21	glycine cleavage sys H
AGAP002593	-	6.61E-114	2.098	0.43	0.04	outer membr lipopr Blc
AGAP001065	-	8.30E-105	2.551	0.76	0.15	glycine hydromethyltran
AGAP004700	-	3.30E-100	2.239	0.38	0.03	None
AGAP010046	-	4.33E-88	2.512	0.29	0.02	None
AGAP009173	Fbp	7.86E-83	2.189	0.38	0.04	fructose-1,6-bisphosph I
AGAP001116	-	1.29E-81	1.946	0.44	0.05	D-amino-acid oxidase
AGAP002198	Gnmt	2.09E-76	2.051	0.46	0.06	glycine N-methyltransf

Cluster 5

Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP005363	-	0	1.729	0.45	0.003	None
AGAP004962	-	0	1.526	0.41	0.004	cyclin B
AGAP007855	-	4.72E-295	1.583	0.43	0.007	aurora kinase, other
AGAP013736	-	8.53E-285	1.075	0.31	0.002	None
AGAP005019	-	2.01E-274	2.028	0.56	0.018	None
AGAP003550	-	3.62E-271	1.302	0.32	0.003	None
AGAP006671	-	1.30E-267	1.117	0.30	0.002	None
AGAP006105	-	5.29E-230	1.018	0.28	0.003	None
AGAP004963	-	7.99E-223	0.989	0.25	0.002	cyclin B
AGAP004239	-	1.13E-212	1.284	0.28	0.003	polo-like kinase 1

Cluster 6

Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP009526	-	1.7E-104	2.864	0.74	0.12	None
AGAP006181	-	1.12E-97	2.621	0.58	0.07	troponin C
AGAP003939	-	5.44E-83	2.674	0.56	0.08	None
AGAP001622	-	2.17E-72	2.640	0.76	0.19	myosin light chain 5
AGAP003778	-	1.13E-70	2.417	0.50	0.07	None
AGAP001569	-	6.19E-66	2.279	0.48	0.07	myosin alkali light chain 1
AGAP004161	-	8.04E-64	2.322	0.74	0.20	myofilin variant C
AGAP002358	-	3.84E-58	2.334	0.45	0.07	ADP,ATP carrier protein 2
AGAP008311	-	2.87E-50	2.092	0.27	0.03	acylphosphatase
AGAP004790	-	5.28E-46	1.918	0.91	0.50	Up skl mscl growth 5 hom

Cluster 7

Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP007347	Lysc1	7.3E-217	4.377	0.91	0.08	C-type lysoz
AGAP005848	-	6.2E-105	2.455	0.39	0.03	Fic A
AGAP011294	DEF1	2.59E-69	1.857	0.28	0.02	defensin anti-micr
AGAP000694	CEC3	2.91E-63	2.455	0.27	0.02	cecropin anti-micr
AGAP000376	Tsf1	1.50E-51	2.139	0.76	0.24	-
AGAP011197	-	1.33E-40	1.779	0.78	0.29	-
AGAP005888	-	2.24E-37	2.573	0.93	0.58	-
AGAP000693	CEC1	1.49E-32	2.855	0.49	0.13	cecropin anti-microb
AGAP005612	-	8.23E-23	2.085	0.32	0.07	-
AGAP010816	TEP3	1.11E-17	1.344	0.34	0.09	thioester-contain prot 3

Cluster 8 Gene	Name	Pval adj	Avg logFC	Pct.1	Pct.2	Annotation
AGAP007318	-	0	3.648	0.79	0.02	None
AGAP009053	LL3	7.0E-212	3.014	0.54	0.02	LITAF-l3
AGAP028208	-	4.0E-195	2.728	0.34	0.01	cuticular prot CPLCP22
AGAP009051	LL1	1.6E-177	1.972	0.37	0.01	LITAF-l1
AGAP007320	-	4.3E-175	1.529	0.29	0.01	None
AGAP001002	-	2.3E-129	3.812	0.42	0.02	Toll
AGAP001652	-	9.6E-107	2.219	0.61	0.05	lipase
AGAP003319	-	6.01E-95	2.147	0.49	0.04	None
AGAP011226	-	1.25E-92	1.941	0.42	0.03	None
AGAP005209	-	1.06E-73	1.817	0.47	0.04	Uridine kinase

Table III.3 Marker genes for each cell cluster. P_val_adj = P value adjusted for multiple testing. Avg_logFC = average log fold change for the gene between cluster of interest and other clusters. Pct.1 = percentage of cells in cluster of interest where gene is detectable. Pct.2 = percentage of cells in other clusters where gene is detectable. Annotation = electronic annotation of gene.

We then assigned putative cell type names based on their gene markers. We molecularly confirmed known cell types such as granulocytes, expressing SPARC, collagens, laminins, scavenger receptors, LRIMs, Nimrod, LRR8 (leucine-rich-repeats), CLIPs [202, 348]. Putative oenocytoids also expressed well known markers such as PPOs (2, 4, 5, 6, 9), fibrinogens, and fibronectins. Potential prohemocytes shared many of the granulocyte markers, including collagens, LRR (leucine-rich-repeats), SPARC, CLIPD1, but also ferritin and ribosomal genes. Of note, expression of granulocyte markers in prohemocytes is not fully abrogated, but rather of lower intensity, suggesting granulocytes and prohemocytes might be different cell states, and not cell types.

We also characterised previously unknown hemocytes classes. For instance, 120 cells baptised ‘secreting hemocytes’ specifically expressed proteins with N-terminal signal peptides for secretion, such as e.g. LYSC1, TEP3, ficolins, cecropins, and defensins. A cluster of 131 ‘Rapidly dividing granulocytes’ was enriched in cell cycle and spliceosome markers such as aurora kinase, Cyclin Bs (G2/Mitotic specific), polo-kinase 1, inhibitor of apoptosis 5, Barrier-

to-autointegration factor B. Finally, 85 ‘effector hemocytes’ were characterised by high expression of LITAF (LPS-Induced TNF-alpha transcription factor) 3 and LITAF 1, AGAP007318 (an uncharacterised membrane protein upregulated in *P. berghei* infection [349]), Toll proteins, NFkappaB essential modulator, CLIPB8. Full table in Appendix.

Interestingly, fat body cells divided into two major cell states, correlated with activation. A baseline fat body state of 701 cells expressed many immune-related and regulatory genes such as CLIPs (CLIPA1, 7, 8, 9, 14), LRIMs (LRIM 1, 4A, 8A, 8B, 9, 17), lectins (CTL 4, MA2), APL1C, SRPN2, TEP1, and phenoloxidase inhibitor protein. Conversely, activated fat body cells (149 cells) highly expressed a canonical marker of fat body after feeding: vitellogenin. Finally, 121 cells have been classified as muscle cells due to the expression of markers such as troponin C, myosin light chain 5, myosin alkali light chain 1, myofilin variant C, and numerous transcripts related to energy production. A heatmap of the top 10 marker genes for each subtype follows below [Fig. III.9].

We also quantified each cell type cluster, looking at both number of cells and total UMI per cell in each cluster to reinforce our hypotheses regarding putative cellular identities. Putative cells types were then identified and quantified. Prohemocytes were the most common cell type with 2034 cells, followed by granulocytes (1553). Baseline fat body cells followed with 701, oenocytoids with 489, and fat body with 149. Rare cells included dividing granulocytes (131), muscle (121), secretory cells (120), and effector cells (85). We classified cell types by taking into consideration both the RNA content of cells - using the number of UMIs per cell as a proxy - as well as the analysis of the differentially expressed genes between each cell cluster. Putative prohemocytes were characterised by a low number of UMIs (yet distinct from background as shown by Fig. III. 8B), consistent with a high nuclear-cytoplasmic ratio and small overall size [Fig. III.8]. Conversely, granulocytes are transcriptionally active, have large diameters, and have high UMIs, similarly to oenocytoids.

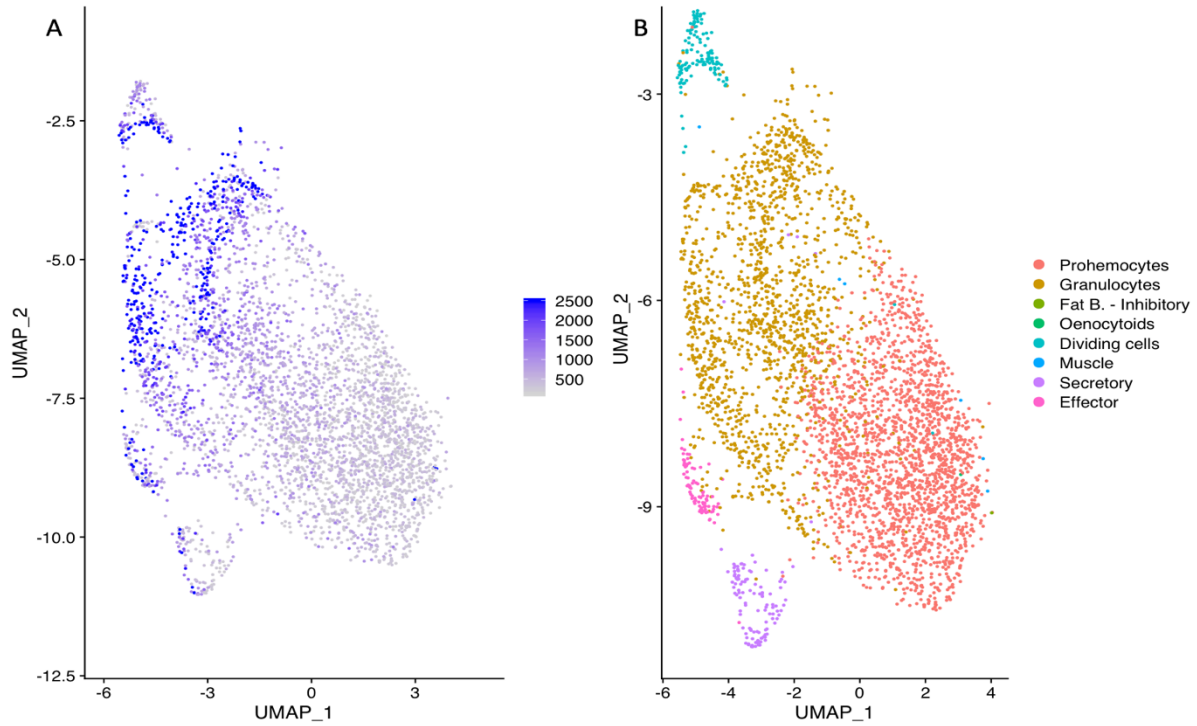

Fig. III.8 UMI count as proxy for size suggests prohemocyte-granulocyte split. Clustering done as above, data split to remove oenocytoids, fat body, and muscle cells **(A)** number of UMIs per cell plotted onto the UMAP visualisation of selected cells, capped at 2500 UMIs to aid visualisation **(B)** clustering solution mapped onto UMAP as above.

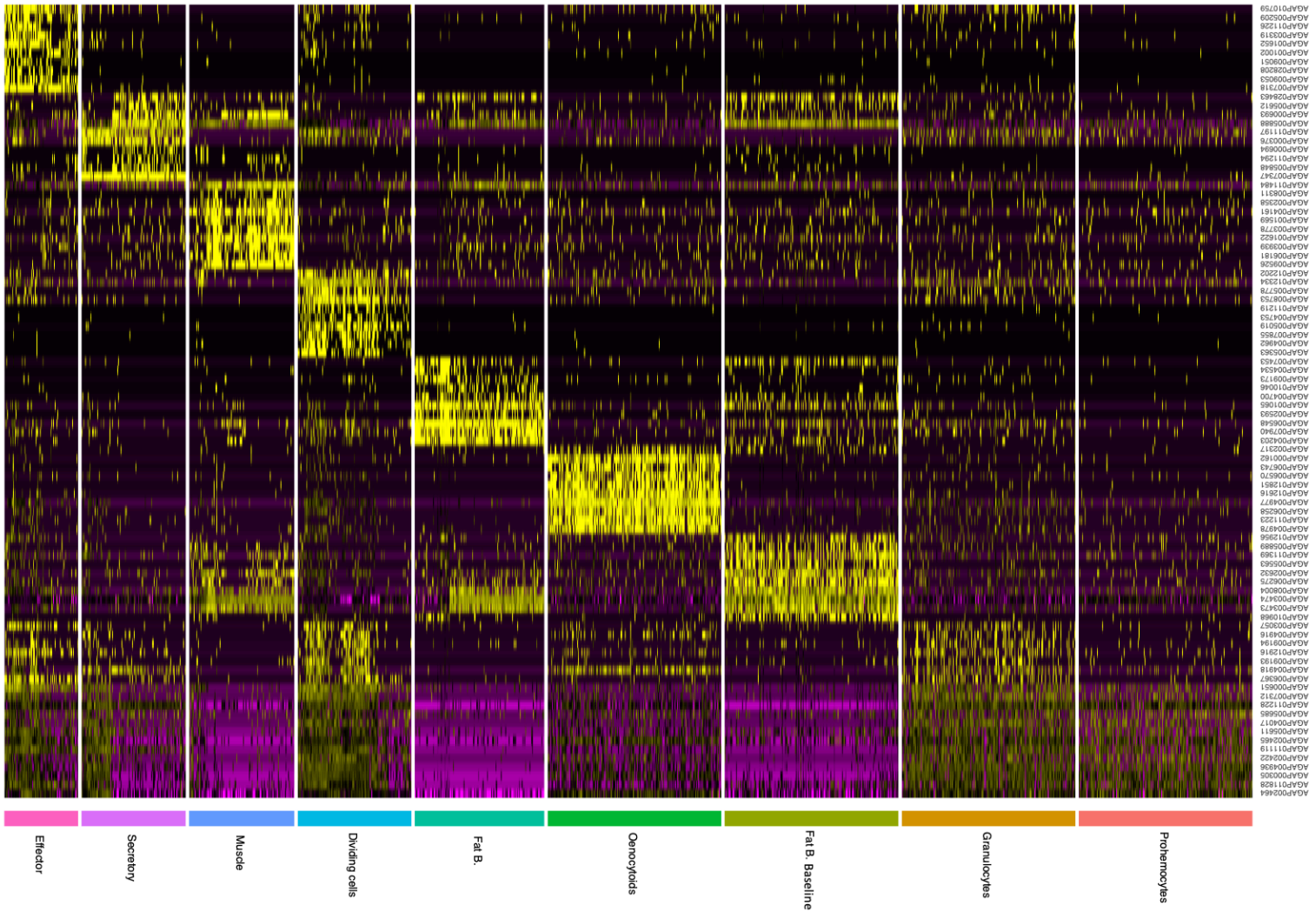

Fig. III.9 Heatmap of the top ten gene markers for each cell type identified. DE genes were identified with the Wilcoxon rank-sum test. P values were adjusted for multiple testing using the Bonferroni correction. All P-adjusted values < 0.001, ordered by average log fold change between cluster of interest and all other cells. Down-sampled to 300 cells per cluster for clarity.

3.1.6 Specific hemocyte markers for RNA-FISH validation identified by combining scRNA-seq and bulk RNA-seq results

We then set out to validate our cell types. The first step was to confirm the exclusive expression of cell type markers in hemocytes, excluding those also expressed in the mosquito midgut or the rest of the body (carcass). Bulk RNAseq of *Anopheles* hemocytes, guts, and carcasses was performed with the same time-points and conditions of the scRNAseq experiments: 1,3 and 7 days after sugar-feeding, blood-feeding, or mosquito infection with *P. berghei*. Between 8-12 mosquitoes per group were used for each condition, with three biological replicates to increase statistical power. After alignment, quantification, and normalisation (see methods) a PCA of the samples showed all biological replicates clustering together. Rather, samples correctly split by body part. Differences between carcass samples in red, gut samples in green, and hemocyte samples in green were the main drivers of sample diversity [Fig. III.10A]. Furthermore, sample-to-sample distances were plotted on a distance matrix to obtain a qualitative appreciation of similarities between samples. The correlation matrix once again demonstrates clear differences between three sample groups: guts, carcasses, and hemocytes.

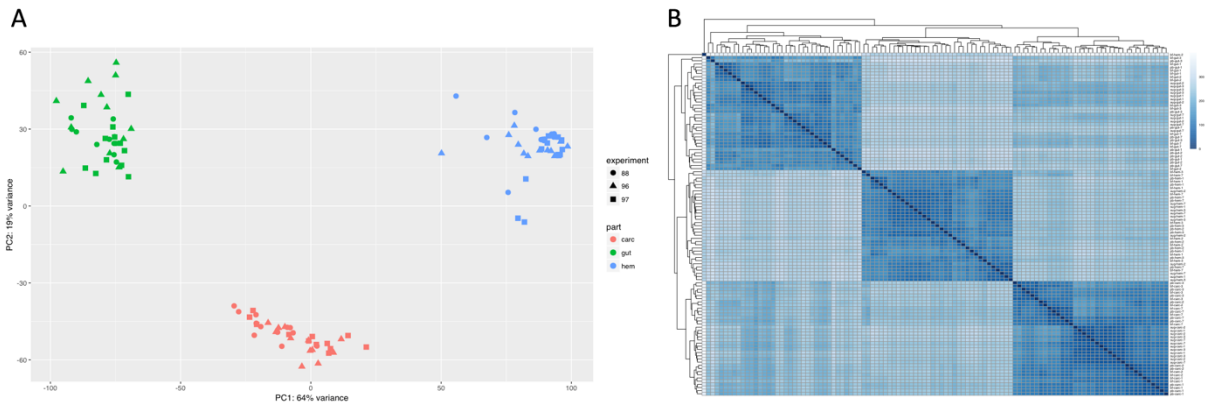

Fig. III.10 Bulk RNA-seq dataset QC. (A) PCA analysis and clustering of samples based on overall transcriptional similarity divides samples into three main groups: carcasses in red, guts in green, and hemocytes in blue **(B)** Distance matrix correlating the overall similarity and hierarchical clustering of each sample. Three large groups (gut top left block, hemocytes in the centre, and carcass at the bottom right)

After QC, normalisation, and fitting of a generalised linear model as of methods we performed a differential expression analysis with DESeq2 on hemocyte samples against the average expression of carcass and gut samples. We filtered for an adjusted p-value after Wald significance testing of P<0.001 and an absolute log2 fold change larger than 2 and identified 5126 differentially expressed genes, of which 1587 were upregulated in hemocytes and 3539 downregulated. Running separate DE analyses of hemocytes vs guts' samples and hemocytes vs carcasses returned similar results. Among the top upregulated genes in hemocytes we found well characterised genes associated either with hemocytes or with immune function, such as PPO2,3,5,6,9, fibrinogen and fibronectin, CLIPs, SPARC, laminins, collagens, scavenger receptors, toll proteins, LRIMs, TEP4, PPO activating factor, CD63, antimicrobial peptides, and REL1.

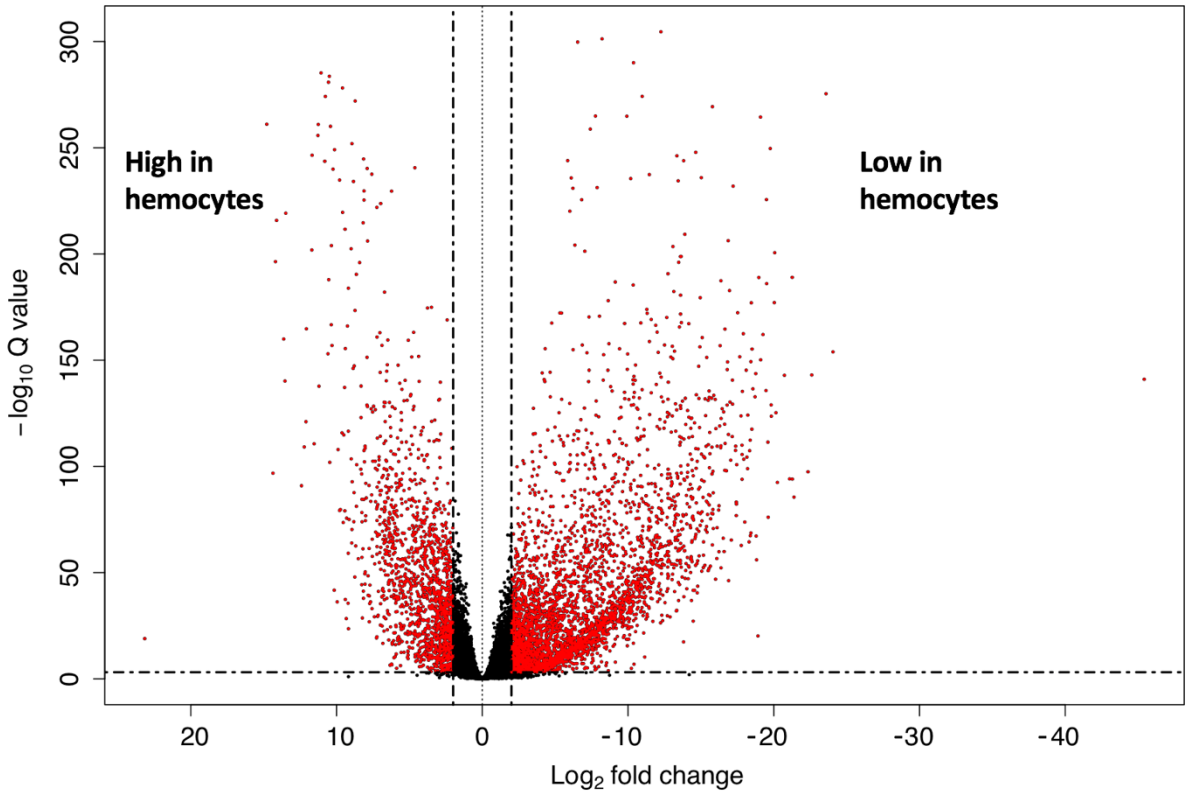

Fig. III.11 Differential expression analysis - hemocytes vs carcasses and guts. DEseq2 DE analysis of hemocytes vs averaged gut and carcass expression, filtered for $\log_2$ fold change >2 and Wald significance testing Q <0.001.

Functional classes of mosquito hemocytes

There was a strong correlation between markers identified by bulk RNAseq and biomarkers of scRNA-seq cell clusters. Especially so for common cells such as prohemocytes (91.2% of sc-RNAseq markers also present in the list of positively upregulated genes in bulk RNAseq hemocytes' samples) and granulocytes (71.3%). Less markers were identified for rare cell types such as secretory cells (only 28.1%) or muscle cells (25.9%), and intermediate levels for cell types such as dividing cells (44.3%) and effector cells (46.5%). Non-hemocyte contaminants such as fat body cells, are also well represented (86.6% and 50.0% for baseline fat body and activated fat body respectively). These cells are large and feature substantial amounts of RNA.

Cluster	Total markers - scRNAseq	Pos. in bulk RNAseq	Percentage
Prohemocytes	34	31	91.2
Granulocytes	178	127	71.3
Fat B. - Baseline	112	97	86.6
Oenocytoids	52	39	75.0
Fat Body	118	59	50.0
Dividing cells	221	98	44.3
Secretory	32	9	28.1
Muscle	58	15	25.9
Effector	99	46	46.5

Table III.4 Correlation of scRNA-seq markers with positively upregulated bulk RNAseq markers in hemocyte samples. First, scRNA-seq marker genes were filtered to select those with Wilcoxon test p adjusted value <0.05. The resulting table was then merged with DE markers in bulk RNAseq hemocyte samples as above, filtered for log2 fold change >2 and Wald significance testing of Q <0.001.

Once DE genes between hemocytes and mosquito midguts and carcasses were identified we cross-references the top ten marker genes for each cluster to the bulk RNAseq gene list to identify the best marker of each cellular subtype for RNA - FISH validation. Markers were selected according to the following criteria:

1) Highest and most specific expression of markers in each scRNA-seq cell type cluster
2) Highest and most specific expression of markers in bulk RNAseq data of hemocytes

Markers were selected using the clustering solution identified with Seurat v2.4. The following table summarises our findings. All markers previously identified and then validated via RNA-FISH were also found to be valid cellular markers in the new Seurat v3 analysis.

Markers	scRNA - specificity	scRNA - expression	Bulk vs gut - log2 fold	Bulk vs body - log2 fold	Description
General					
AGAP009623	n/a	n/a	n/a	n/a	GAPDH – pos. control
AGAP008296	n/a	n/a	-13.2	-7.6	Trypsin - gut
AGAP004203	+++	+++	4.1	-2.5	Vitellogenin - fat body
Hemocytes / Granulocytes					
AGAP004017	n/a	+++	7.3	4.8	LRR. All hemocytes
AGAP011974	n/a	++	5.6	4.2	SCRC1. General hemos
AGAP000790	n/a	+	6.6	4.7	Prohem. / granulocytes
AGAP003057	+	+	4.7	1.8	Active granulocytes
AGAP011871	-	+	2.6	1.2	Granulocytes
Rapidly dividing					
AGAP005363	+++	++	1.2	0.4	
Fat B. - Baseline					
AGAP007033	+	+	6.8	1.2	
AGAP028406	++	++	5.7	3.2	APL1C
Oenocytoids					
AGAP004981	++	++	10.4	4.8	PPO4
AGAP012851	+++	+++	6.9	4.7	Aldo-keto-reductase
AGAP012000	++	++	8.1	5.5	Fibrinogen/fibronectin
Effector					
AGAP007318	+++	++	5.3	2.8	TM7318
Secretory					
AGAP011239	++	++	4.0	2.9	Some also in oenos

Table III.5 RNA-FISH markers chosen by total expression and expression specificity in scRNA-seq and bulk RNAseq samples. scRNA-seq markers were cross-checked with gene tables of DE genes in bulk RNAseq (hemocytes vs guts and hemocytes vs bodies, separately). The most specific and highly expressed genes (qualitative assessment) were chosen.

3.1.7 RNA-FISH validation of putative cell types

We then validated our cell types via imaging. Dr. Ana Barletta Ferreira recovered hemocytes from mosquitoes that were sugar-fed, blood-fed or infected with *P. berghei*, spun the hemocytes onto slides coated with the adhesive Cel-Tek, then fixed them in paraformaldehyde. The cellular morphology was first captured by staining cells with actin and imaging them with confocal microscopy, and then RNAscope commercial RNA-FISH was then performed with the probes of Table III.5, and correlative fluorescent / FISH microscopy was performed by imaging the same area of the slide with confocal microscopy [Fig. III.12].

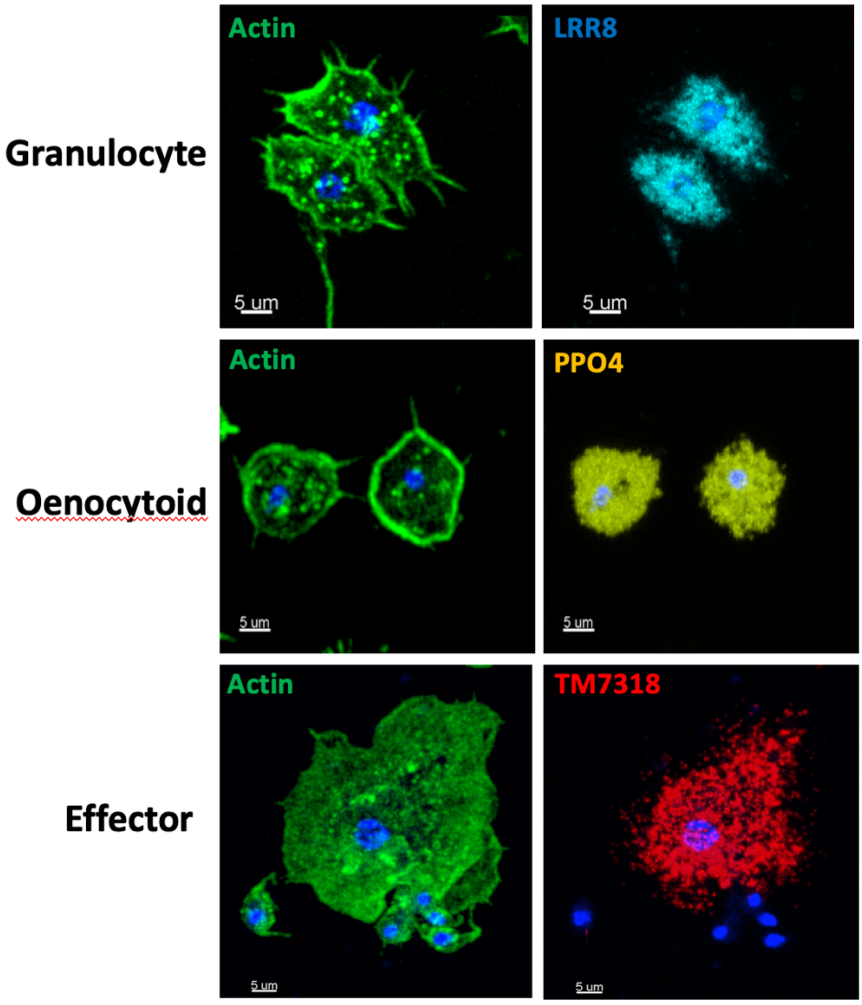

Fig. III.12 Correlation of hemocyte morphology with RNA-FISH markers. Main cell types were confirmed by matching to the left cellular morphology (actin), and to the right gene markers by RN-FISH. Blue is DAPI nuclear stain. Representative images from over 3200 cells.

Granulocytes were identified because of their larger size (10-20 µm) as compared to oenocytoids (8-12 µm) and prohemocytes (4-6 µm). In addition, granulocytes featured an increased number of pseudopodia. Oenocytoids also had pseudopodia, but they were shorter, and less prominent, and cells were rounder. Furthermore, the nuclear size in granulocytes was larger than in oenocytoids [Fig. III.12]. LRR8 mostly identified granulocytes and prohemocytes, whereas PPO_4 identified for the most part oenocytoids. Some cells were double-positive, but typically $LRR8_{high}$ cells would be $PPO4_{neg}$ or $PPO4_{low}$, and conversely $PPO4_{high}$ cells would be $LRR8_{neg}$ or $LRR8_{low}$ [Fig. III.13]

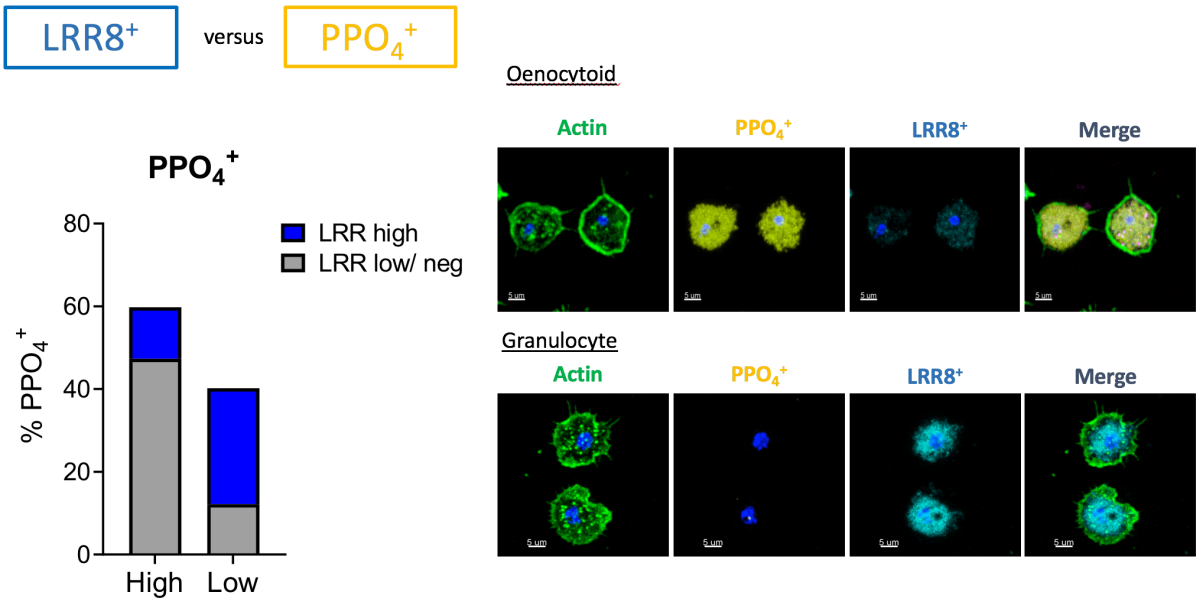

Fig. III.13 Granulocytes vs oenocytoids: morphology and RNA-FISH markers. LRR8+ cells could be split into LRR_8 high and low. PPO4+ cells (oenocytoids) were more likely to be LRR8 negative or low. The opposite for $PPO4_{low}$ cells. Representative images from 435 cells.

We then explored the spatial localisation of hemocytes in the *Anopheles* mosquitoes. Mosquitoes were then sugar-fed, blood-fed or infected with *P. berghei*, then fixed in paraformaldehyde, before paraffin embedding and sectioning. We performed RNA-FISH with the commercial technology RNAscope on the sections per RNAscope protocol and then imaged samples on an automated slide scanner or with confocal microscopy. We alternated one slide for haemotoxylin and eosin (H&E) staining and one slide for RNAscope. H&E staining was

useful to orient ourselves and identify the anatomical features of mosquitoes. In Fig. III.14 we can observe an H&E stain and mirrored RNA-FISH section of the mosquito. From the left to the right we can observe the compound eye, brain, thorax and wing muscles, abdomen and foregut, midgut, and fat body, as well as the ovaries.

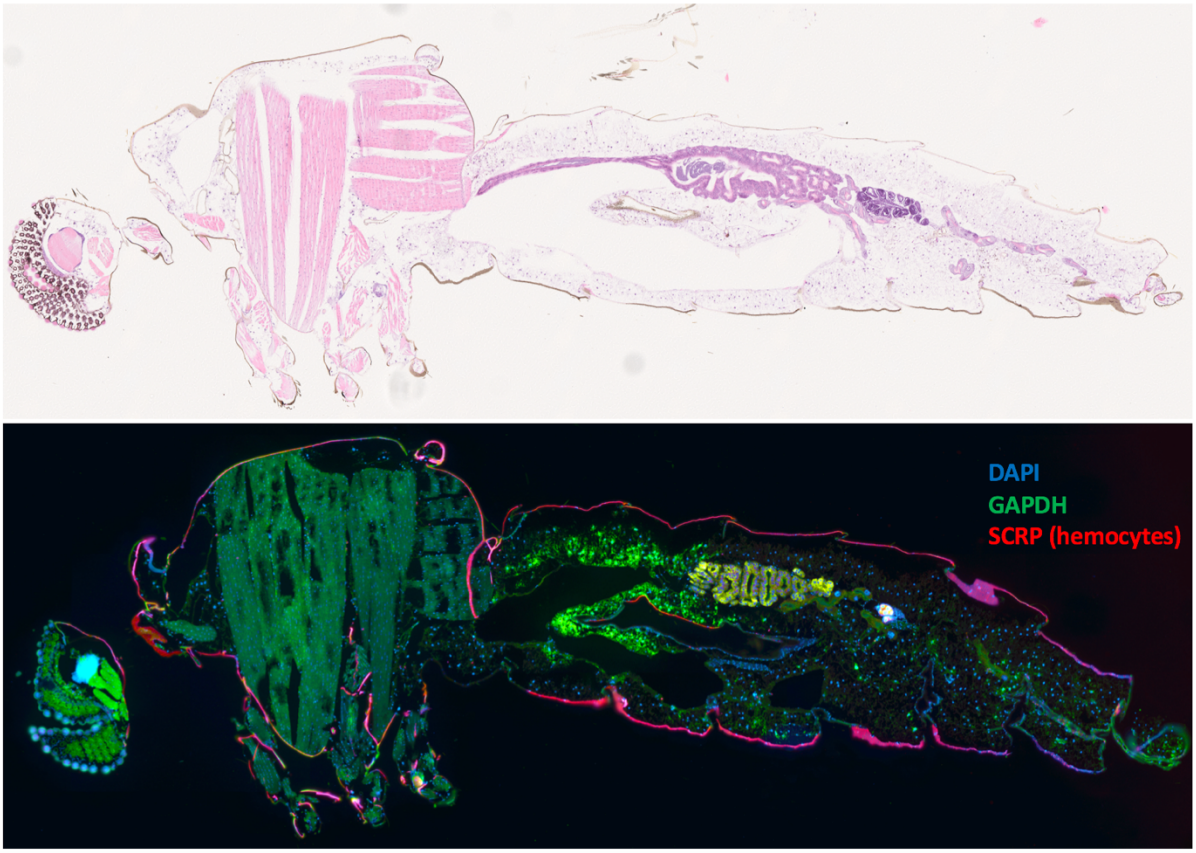

Fig. III.14 Overall view of the *A. gambiae* body with H&E and RNA-FISH. At the bottom, RNA-FISH of hemocytes (red, SCRC1 probe), cellular nuclei (blue DAPI counter-stain), and all mosquito cells (green, GAPDH positive control mosquito probe) on a longitudinal section of an *Anopheles* mosquito. At the top, mirrored H&E section. Both imaged with slide scanner.

Hemocytes can be seen patrolling all areas of the mosquito body, including the thorax - between flight muscles - and the abdomen, both in the fat body or attached to the gut. Hemocytes are found everywhere (except within the gut lumen or the central nervous system) but they particularly line areas of the body in potential contact with pathogens, such as the salivary glands, the proboscis, the gut lining, the rectal area, and the spermathecal vestibule of female mosquitoes. Hemocytes do not normally form clumps but appear as isolated cells,

although in these sections we mainly used the SCRC1 probe for our survey. SCRC1 is more specific for granulocytes and prohemocytes than oenocytoids, secretory, or effector cells.

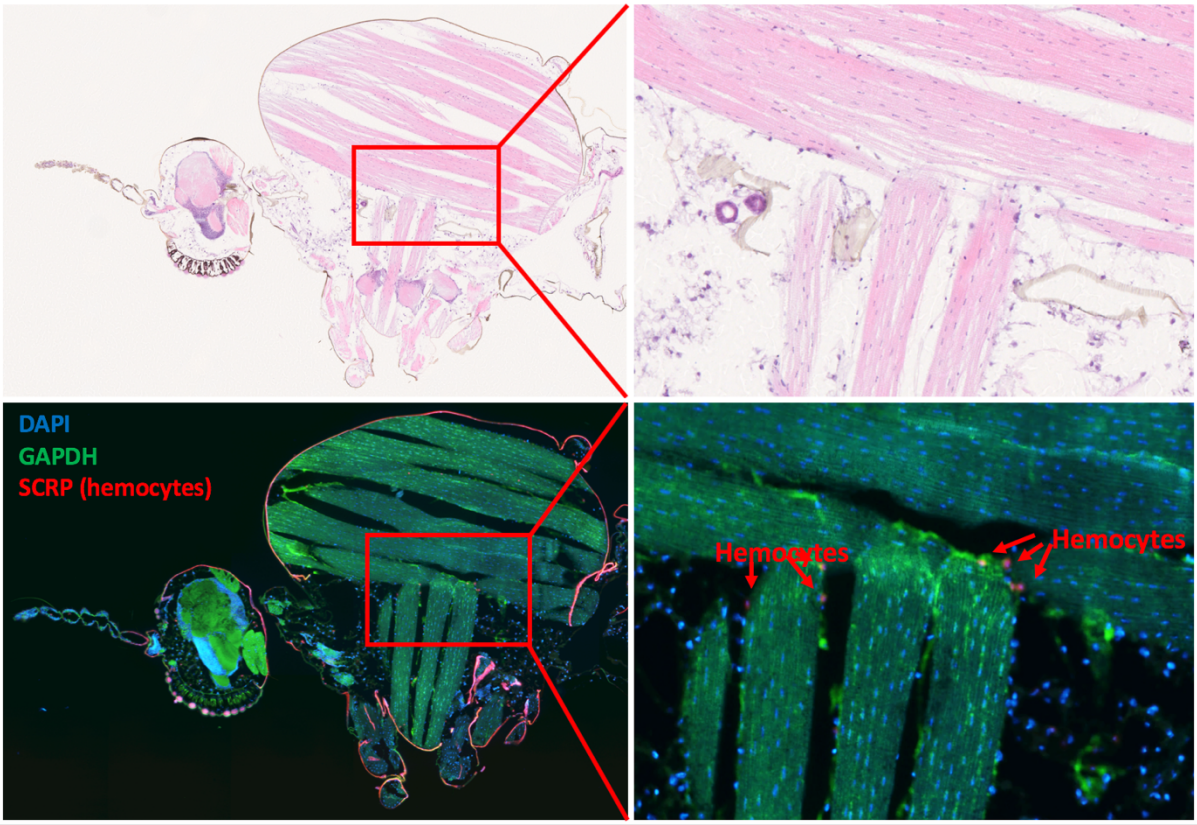

Fig. III.15 Hemocytes patrolling the thorax of *A. gambiae*. At the bottom, RNA-FISH of hemocytes (red, SCRC1 probe), cellular nuclei (blue DAPI counter-stain), and general mosquito cells (green, GAPDH positive mosquito control probe) on longitudinal section of *Anopheles* mosquito. At the top, mirrored H&E section. Both imaged with slide scanner.

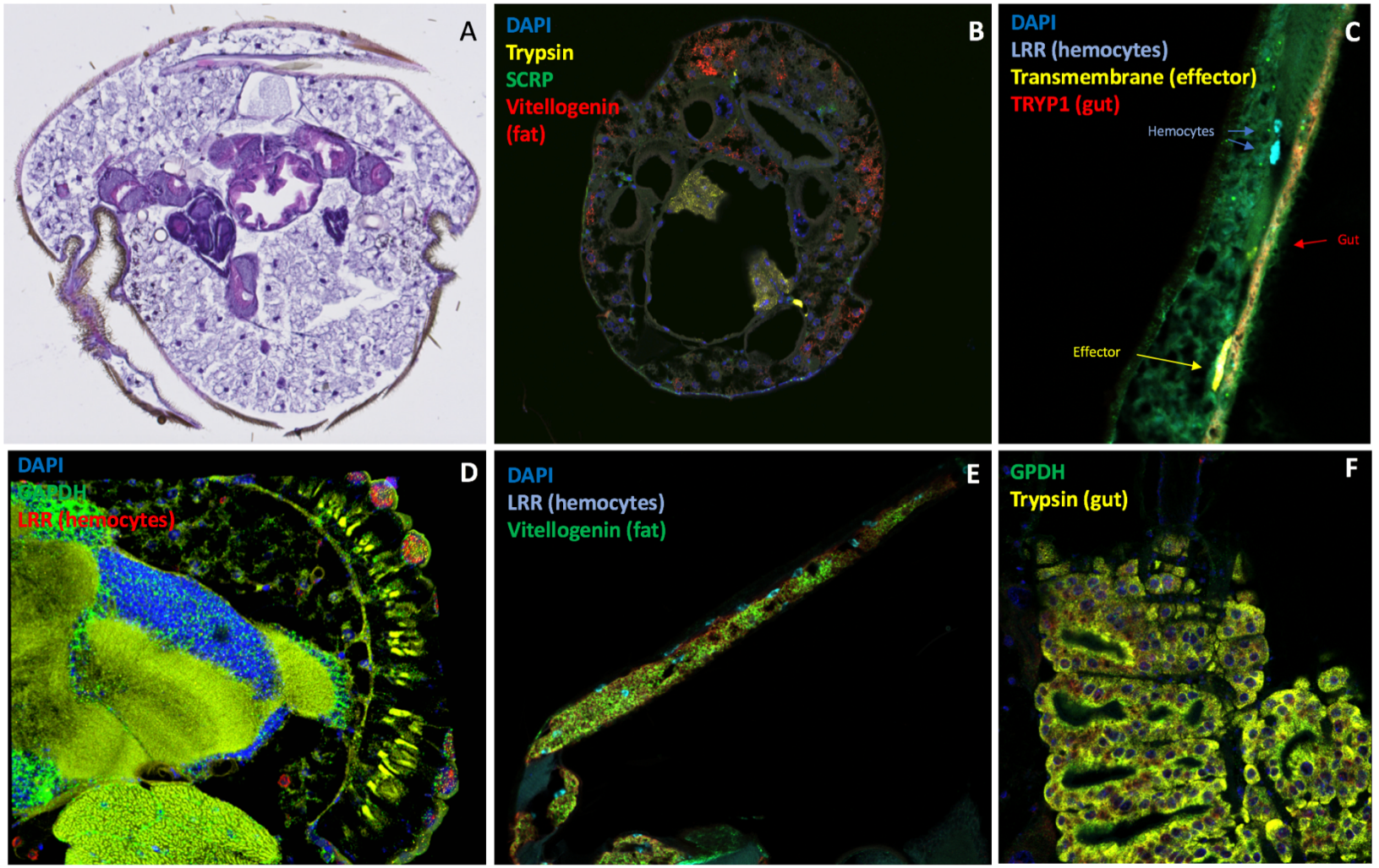

Fig. III.16 Hemocytes patrolling the *A. gambiae* body (A) Vertical H&E section of mosquito abdomen and **(B)** mirrored RNA-FiSH section. From **C to F** RNA-FISH of: gut lining in abdomen, CNS, proboscis, and gut. Imaged with slide scanner **(A-C, E-F)** and confocal microscopy **(D)**. RNA-FISH probes indicated in each separate panel.

Functional classes of mosquito hemocytes

Hemocytes can be both sessile and motile. Imaging requirements for each are different. To capture sessile hemocytes we injected paraformaldehyde inside the mosquito cavity before dissecting the mosquito midgut and the mosquito body wall (carcass). Then, whole-mount RNA-FISH of the whole organs were done with a modified RNAscope protocol (see methods). All hemocyte cell types for which we have probes were identified with the exception of the rapidly diving cellular subtype, for which we have yet to develop an appropriate probe. We observed the general hemocyte population, as well as specific oenocytoids, granulocytes, effector hemocytes, and secretory hemocytes. Body walls were especially rich in immune cells, with control blood fed body walls having 286 (±76 CI) hemocytes. Blood-fed control guts showed fewer numbers of cells, with a total of 23 (±6.6 CI) hemocytes. We also observed pericardial cells, staining positively with the AGAP007318 and AGAP011239 probes (effector and secretory probes). These cells could be recognised both by virtue of their characteristic arrangement along the dorsal wall as well as their larger size.

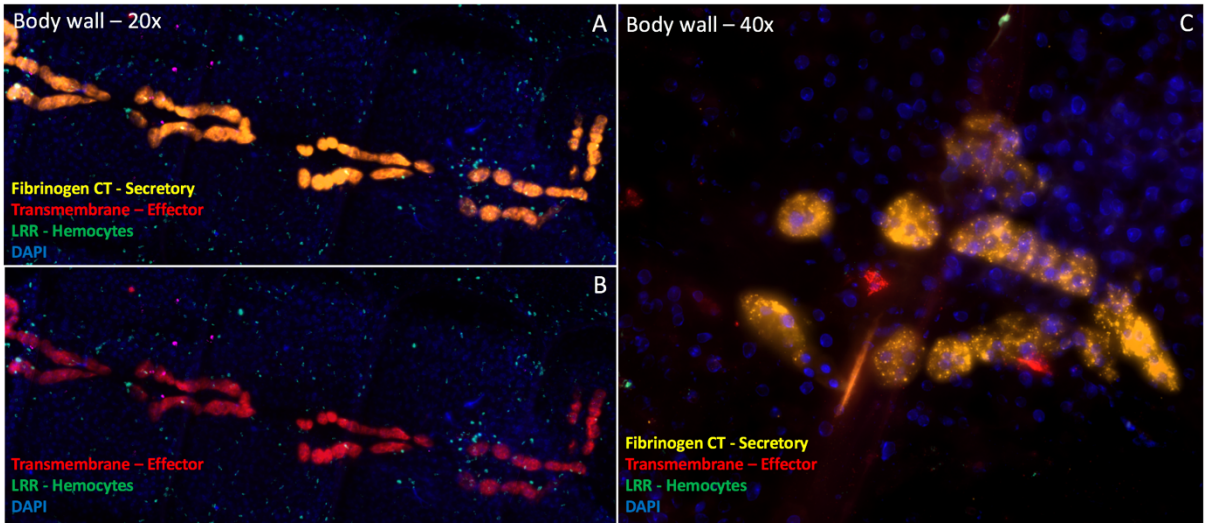

Fig. III.17 Pericardial cells along the *Anopheles* body wall (A) 20x whole-mount RNA-FISH shows AGAP007318 and AGAP011239 positive pericardial cells, in addition to immune cells **(B)** Same as above but without the Fibrinogen-CT probe to show positive staining for Transmembrane (Effector) probe **(C)** 40x whole-mount RNA-FISH of a separate mosquito wall. Two effector hemocytes can be seen in close proximity to the pericardial cells complex.

Functional classes of mosquito hemocytes

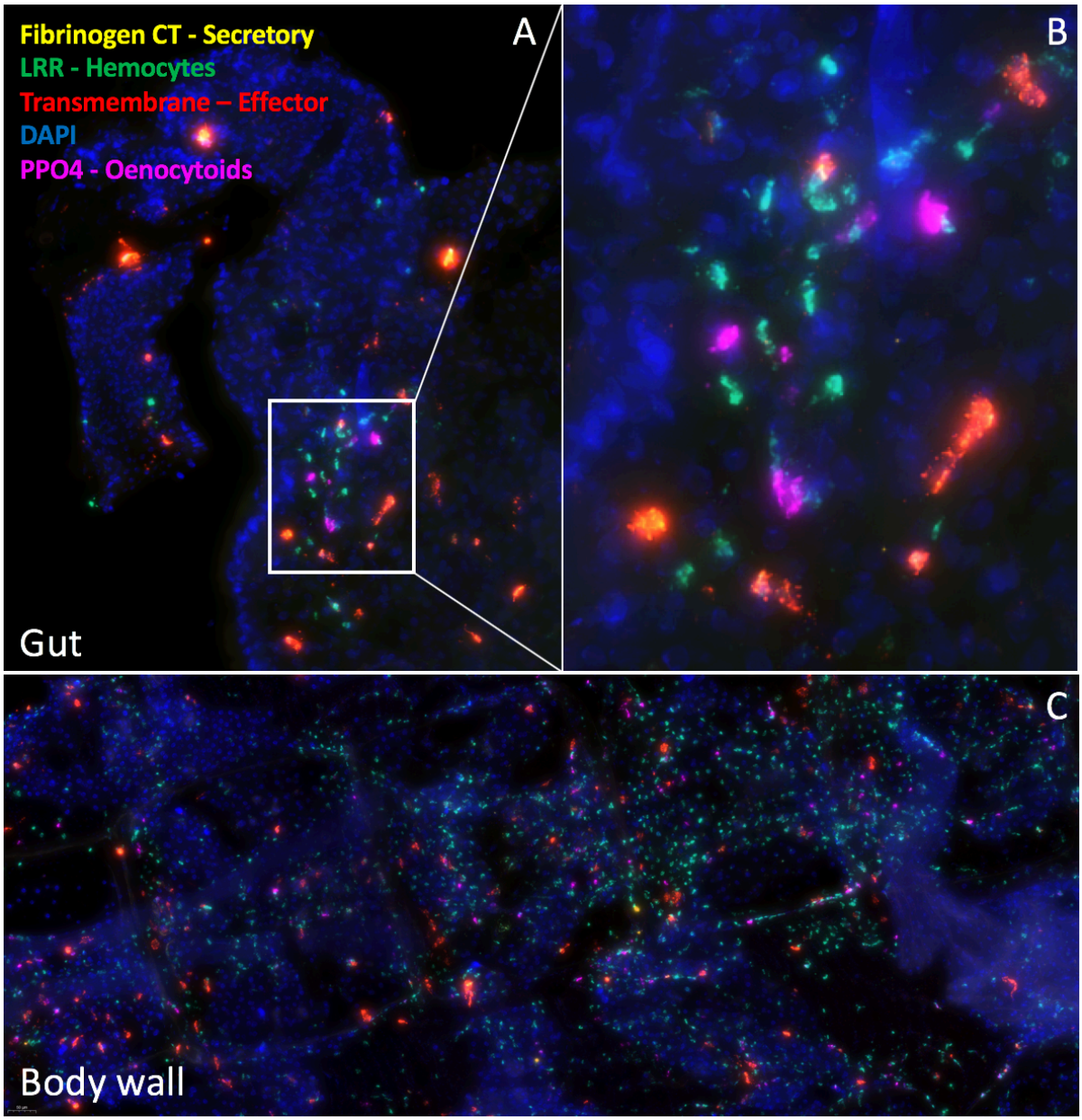

Fig. III.18 Mosquito midguts and bodies contain all subtypes of sessile hemocytes (A) A 20x view of the proximal part of a blood-fed control mosquito gut, with RNA-FISH of hemocytes (green, LRR probe), secretory cells (yellow, Fibrinogen C Terminal), effector cells (red, transmembrane), and nuclear counterstain (blue, DAPI) on whole mounts of *Anopheles* mosquito. **(A)** A 40x magnification of the gut. **(C)** A 20x whole mount view of a mosquito body wall with the same probe of above. All imaged with a slide scanner.

3.1.8 Distinct states within each cell type

While initially conservative in our clustering as to only capture true cell types rather than cell states, thresholding was then relaxed to identify subtler grouping of cells, which could theoretically split existing cell types into cell states, differentially responding to stimuli. There was hidden diversity within the original mapping, especially in the large granulocyte cluster. We observed a central disc of cells, surrounded by two separate hemi-discs. Importantly, the central group contained more cells from baseline conditions, whereas the two hemi-discs featured more active cells (blood fed and *P. berghei*-infected) [Fig. III.19A-B]. After iterating clustering until all clusters had at least more than 20 meaningful marker genes (adjusted p value <0.05) and were well-mixed among samples and conditions, we identified four additional cell states. Fat body cells divided into an additional cell state that sat between baseline cells and activated cell types based on the UMAP and the marker genes (see table III.6 below for top 10 genes, as well as figures III.19 and III.20). From the same figures and tables prohemocytes also split in two: a more active state defined by increased expression of hemocyte / granulocyte genes and a more inactive state with decreased gene expression. Granulocytes showed the largest transcriptional diversity, splitting into three different cell states: one putative baseline state, as well as two different types of more activated granulocytes [Fig. III.19C]. The baseline granulocyte cluster contained the highest number of inactivated cells (sugar-conditions), whereas activated cells came either from blood-fed or even more so from *P. berghei*-infected samples [Fig.19A, Fig. IV.1, Fig. IV.2]. A heatmap of the top 10 marker genes for each cell state more clearly showed how putative prohemocytes and granulocytes sat in a transcriptional programming continuum. Oenocytoids on the other hand still formed a distinct separate group on the UMAP, as well as on the marker genes heatmap. Furthermore, the heatmap also showed how within the prohemocyte-granulocyte group baseline granulocytes and prohemocytes were more similar to each other, whereas Type 1 and Type 2 active granulocytes show larger transcriptional differences.

Functional classes of mosquito hemocytes

Putative inactive prohemocytes

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP011828	4.98E-47	0.4644	0.843	0.725	cathepsin L
AGAP010163	2.85E-39	0.3039	0.943	0.961	60S ribosomal protein L38
AGAP007740	1.14E-36	0.2630	0.96	0.966	60S ribosomal protein LP1
AGAP012100	2.03E-36	0.2586	0.966	0.977	40S ribosomal protein S26
AGAP000305	5.64E-26	0.2972	0.877	0.739	SPARC
AGAP002464	2.68E-23	0.2907	0.95	0.909	secreted ferritin G subunit
AGAP029054	7.29E-17	0.3604	0.739	0.645	nimrod B2
AGAP002422	1.94E-15	0.5140	0.591	0.56	CLIP-domain serine prot
AGAP002465	5.35E-15	0.3314	0.804	0.78	ferritin heavy chain
AGAP013186	3.70E-07	0.2842	0.15	0.282	None

Putative active prohemocytes

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP004936	2.69E-63	0.58799	0.873	0.65	None
AGAP011119	7.93E-54	0.554097	0.843	0.638	None
AGAP011228	7.70E-47	0.445928	0.981	0.792	None
AGAP002464	1.29E-45	0.488801	0.974	0.908	secreted ferritin G subunit
AGAP005611	7.84E-37	0.50457	0.775	0.65	None
AGAP000305	1.15E-20	0.30061	0.899	0.745	SPARC
AGAP002465	2.32E-19	0.365025	0.854	0.773	ferritin heavy chain
AGAP011828	8.19E-19	0.297401	0.86	0.73	cathepsin L
AGAP002422	1.36E-18	0.475993	0.654	0.551	CLIP-domain serine prot
AGAP002878	6.99E-13	0.519473	0.509	0.408	Cystatin-like protein

Putative baseline granulocytes

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP011228	4.6E-101	0.74197	0.988	0.796	None
AGAP011119	7.82E-93	0.682409	0.946	0.628	None
AGAP004936	2.29E-76	0.630795	0.939	0.646	None
AGAP007312	1.03E-65	0.66471	0.781	0.428	None
AGAP006278	7.85E-62	0.583197	0.893	0.583	None
AGAP005611	2.25E-58	0.519989	0.915	0.632	None
AGAP002594	1.50E-57	0.602139	0.743	0.426	apolipoprotein D
AGAP000790	2.86E-56	0.799228	0.47	0.196	None
AGAP000305	4.56E-56	0.516235	0.961	0.74	SPARC
AGAP000964	4.50E-51	0.672595	0.66	0.353	None

Putative granulocytes T2

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP006367	6.0E-165	1.427458	0.547	0.104	None
AGAP004916	1.54E-89	1.210717	0.38	0.087	None
AGAP004164	8.04E-80	0.974785	0.543	0.181	glutathione S-transf delta cl. 1
AGAP003016	1.11E-79	0.930581	0.446	0.125	mesenceph. neurotroph hmlg
AGAP029139	7.64E-76	0.98333	0.604	0.238	None
AGAP007120	1.16E-72	0.720407	0.901	0.584	nucleoside-diphosphate kinase
AGAP004743	2.90E-70	0.838938	0.657	0.275	Transmembr. emp24 containing
AGAP009194	5.10E-67	1.183577	0.407	0.124	glutathione S-transf. epsilon 2
AGAP005861	1.00E-66	0.877063	0.428	0.131	Translocon-associated subun β
AGAP004918	1.90E-60	1.094499	0.596	0.282	fibrinogen

Putative granulocytes T1

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP011828	7.3E-109	0.943378	0.983	0.73	cathepsin L
AGAP009156	6.93E-97	1.016372	0.505	0.118	None
AGAP004993	4.51E-93	1.109549	0.84	0.427	laminin subunit alpha
AGAP009201	1.17E-92	1.130115	0.842	0.481	collagen type IV alpha
AGAP011974	7.29E-88	1.013233	0.732	0.291	Class C Scavenger Receptor
AGAP002599	9.63E-83	0.916165	0.818	0.387	polyubiquitin
AGAP002016	3.58E-82	0.988187	0.545	0.158	iron/zinc purple acid phosphata
AGAP002879	3.12E-73	0.8705	0.78	0.357	cathepsin F
AGAP028157	1.02E-70	0.824397	0.452	0.12	None
AGAP013509	1.26E-70	0.947402	0.72	0.322	carboxylesterase clade H, 1

Putative fat body baseline T.1

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP010968	0	2.657141	0.702	0.053	CLIPA9
AGAP008013	2.8E-303	1.993149	0.418	0.013	None
AGAP005563	2.5E-290	2.843697	0.731	0.084	Tret1
AGAP011792	5.1E-269	2.177422	0.541	0.039	CLIPA7
AGAP006275	7.7E-261	2.351344	0.86	0.156	None
AGAP008227	7.3E-258	2.216784	0.737	0.097	trehalose 6-phosphate synth
AGAP002588	4.2E-254	1.689412	0.38	0.014	None
AGAP013060	6.1E-250	1.889122	0.804	0.123	None
AGAP008688	1.0E-245	2.040984	0.392	0.017	None
AGAP006177	3.9E-245	1.748359	0.406	0.02	None

Functional classes of mosquito hemocytes

Putative fat body T1

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP004203	8.4E-162	3.006847	0.782	0.096	vitellogenin
AGAP007940	3.8E-126	2.764072	0.721	0.109	Reticulon-like protein
AGAP006548	2.6E-124	2.550861	0.912	0.214	glycine cleavage system H
AGAP002593	8.3E-116	2.110959	0.435	0.035	outer membrane lipoprot Blc
AGAP001065	1.9E-104	2.542809	0.769	0.15	glycine hydroxymethyltransf
AGAP004700	7.0E-102	2.252557	0.381	0.03	None
AGAP010046	1.58E-89	2.525451	0.293	0.019	None
AGAP009173	2.84E-84	2.202421	0.381	0.037	fructose-1,6-bisphosphatase I
AGAP001116	5.13E-80	1.918763	0.442	0.054	D-amino-acid oxidase
AGAP002198	7.81E-78	2.063797	0.463	0.062	glycine N-methyltransferase

Putative fat body T2

Gene	P_val_adj	Avg_logFC	Pct.1	Pct.2	Annotation
AGAP003473	3.3E-163	2.480769	0.865	0.305	None
AGAP003474	6.5E-160	2.15225	0.992	0.955	None
AGAP005888	1.2E-135	1.620302	0.945	0.563	None
AGAP002632	1.5E-105	2.280139	0.701	0.265	None
AGAP004203	9.03E-93	2.308033	0.437	0.091	vitellogenin
AGAP012571	3.67E-91	1.310944	0.673	0.222	None
AGAP008011	3.37E-85	1.437902	0.382	0.072	None
AGAP008004	7.54E-82	1.195067	0.813	0.409	None
AGAP028386	2.64E-81	1.502057	0.799	0.469	NADH dehydr subunit 6
AGAP028373	3.34E-77	1.474149	0.626	0.23	NADH dehydr subunit 3

Table III.6 Marker genes for each cell state cluster. P_val_adj = P value adjusted for multiple testing. Avg_logFC = average log fold change for the gene between cluster of interest and other clusters. Pct.1 = percentage of cells in cluster of interest where gene is detectable. Pct.2 = percentage of cells in other clusters where gene is detectable. Annotation = electronic annotation of gene.

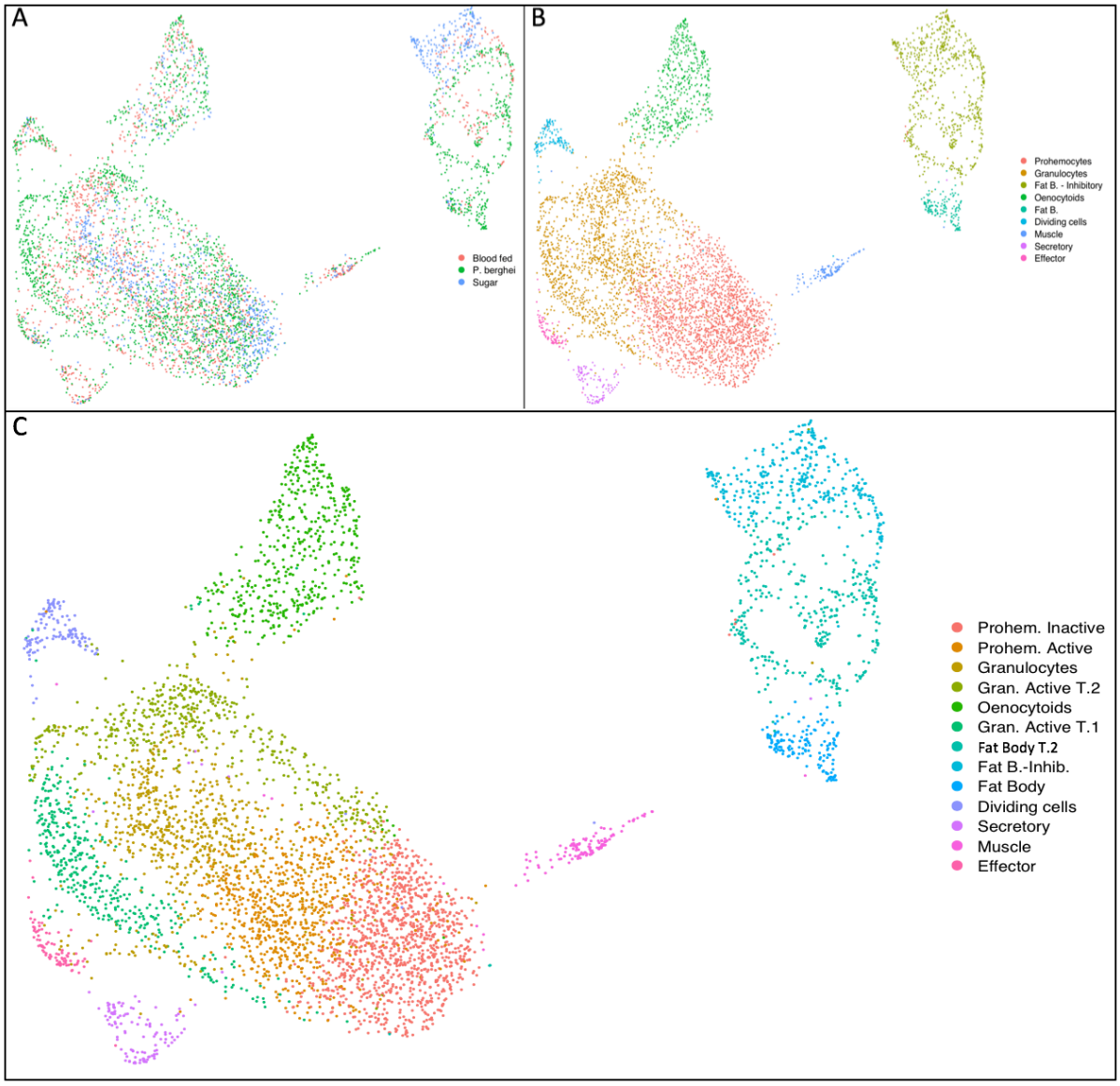

Fig. III.19 Diversity within cell types. (A) UMAP coloured by experimental condition. Within the putative granulocyte cluster, cells from sugar-fed (in blue) mosquitoes segregated from blood-fed mosquitoes (red), and more so *P. berghei* mosquitoes (green) **(B)** UMAP of cells clustered with 0.3 resolution (conservative subdivision identifying cell types) **(C)** UMAP of cells clustered with 0.7 resolution to identify cell states within the larger cell types.

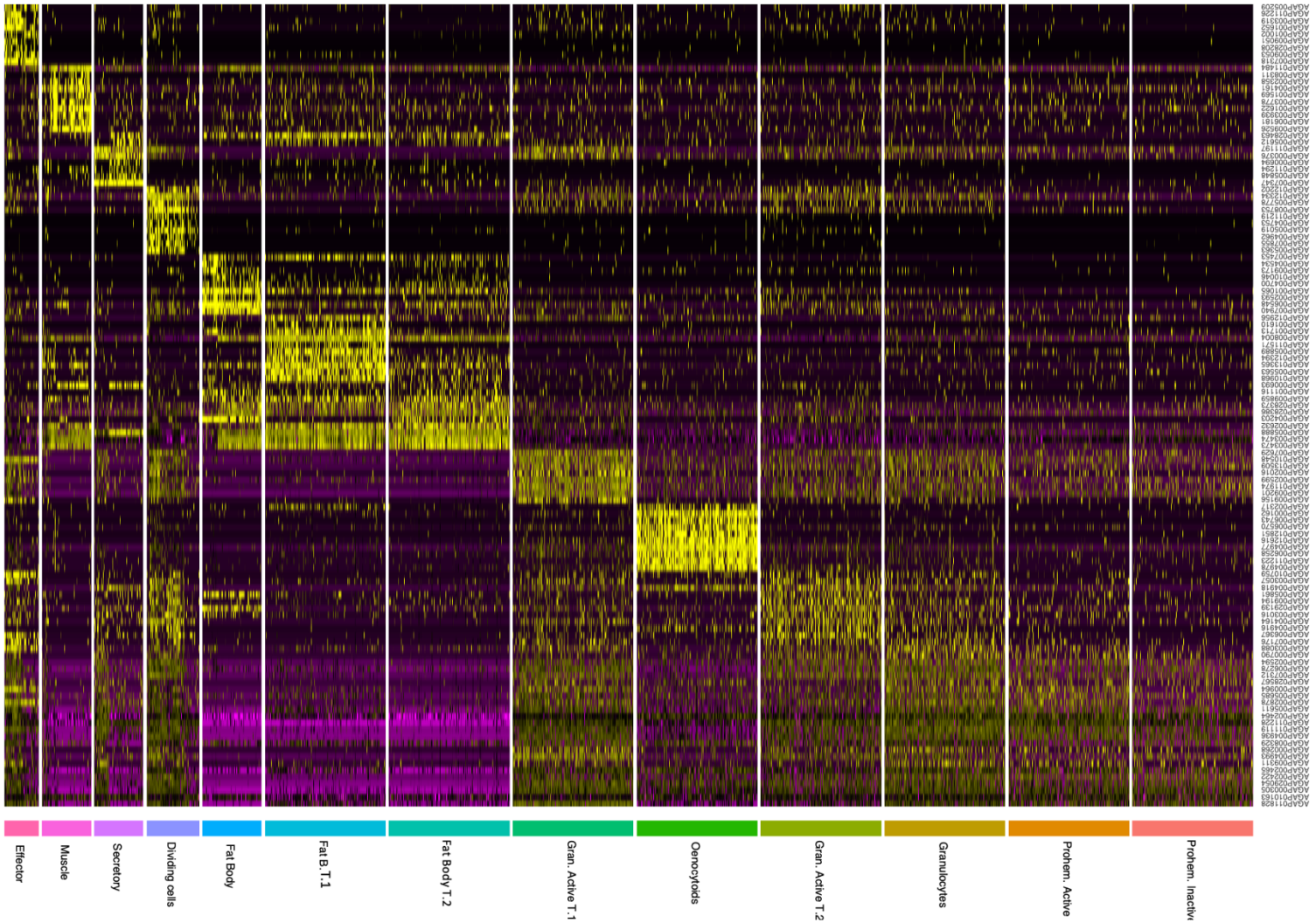

Fig. III.20 Heatmap of top ten gene biomarkers for each cell type or state. DE genes were identified with the Wilcoxon rank-sum test. The P values were adjusted for multiple testing using the Bonferroni correction. P-adjusted values < 0.001, ordered by average log fold change between cluster of interest vs all other cells. Down-sampled to 300 cells per cluster for clarity.

3.1.9 Distinct hemocyte lineages in *A. gambiae* mosquitoes

Hemocytes differentiation dynamics are unclear. To understand whether prohemocytes are true stem cells or a separate lineage we used cellular states subdivision to perform lineage tree reconstruction with the partition-based graph abstraction (PAGA) method. By combining clustering and pseudotemporal algorithms we were able to infer hemocyte trajectories and differentiation paths. We chose PAGA as it was recently shown to be the most accurate and robust lineage analysis software for complex datasets [311]. As a positive control, PAGA correctly identified fat body cells and muscle cells as separate clusters with no close connection to other cell types. Oenocytoids were also shown to be disconnected from other hemocyte subtypes, indicating a wholly separate lineage, while all other cell states were connected along a linear differentiation trajectory with inactive baseline prohemocytes at one end, moving towards active prohemocytes and granulocytes, before splitting into three different lineages. Secretory cells formed their own lineage from baseline granulocytes, while the two intermediate activated granulocyte cell subtypes split into either effector granulocyte subtypes or dividing granulocytes. Dividing cells reverted back into activated granulocytes type 2, replenishing the granulocyte cell pool after immune activation. We were thus able to identify a branching event centred on granulocytes thanks to an unsupervised network analysis. Nodes were identified with Seurat and connected by PAGA into a biologically meaningful network.

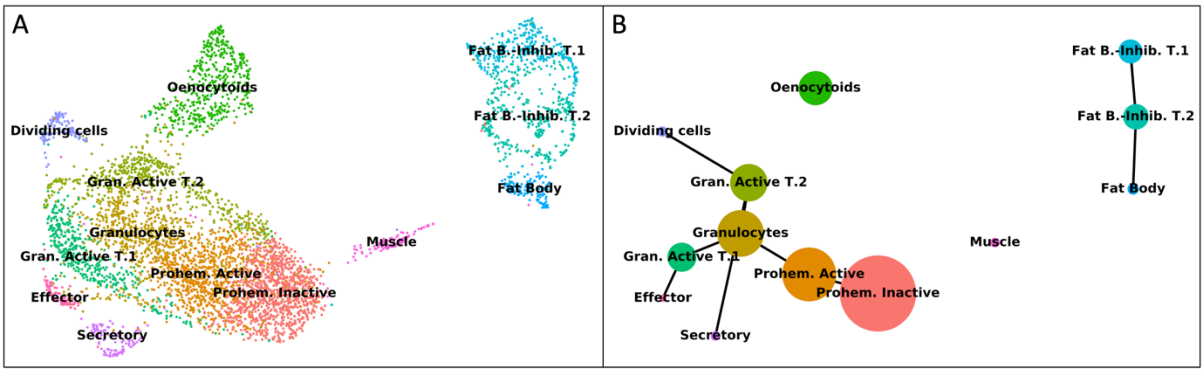

Fig. III.21 Cell lineages in adult *Anopheles*. (A) Graphical mapping of cell states with UMAP **(B)** Unsupervised PAGA network analysis of *Anopheles* hemolymph cells uncovers separate lineages and a branching event. Nodes correspond to clusters identified with Seurat while edges are putative cluster transitions.

Functional classes of mosquito hemocytes

We then confirmed the connections between clusters in the granulocyte lineage with a different method, diffusion maps. Like PCA, diffusion maps are another popular dimensionality reduction technique. However, diffusion mapping is a non-linear dimensionality reduction technique which aligns cells based on transcriptional similarities rather than clustering them. Hence, diffusion components (DCs) emphasize transcriptional transitions, which is particularly useful when analysing processes that are continuous, as for instance differentiation. Our data set showed DC1 to recapitulate the interconnectivity of prohemocytes, active prohemocytes, granulocytes, and active granulocytes type 1 and 2. These existed in a continuum of differentiation which includes dividing cells, whereas effector, secretory, and diving cells split along their independent trajectories [Fig. III.22A-B]. A DC1 vs DC3 plot showed that rapidly dividing cells and active granulocytes type 2 sat on a common differentiation trajectory, as expected from PAGA lineage tracing [Fig. III.22C]. DC1 vs DC3 also showed the opposite lineage (effector cells) emerging from active granulocytes type 1 [Fig. III22.D]. DC2 recapitulated hemocyte cell maturity: young, proliferating cells sat diametrically opposite to mature effector cells such as effector and secretory hemocytes [Fig. III.22E].

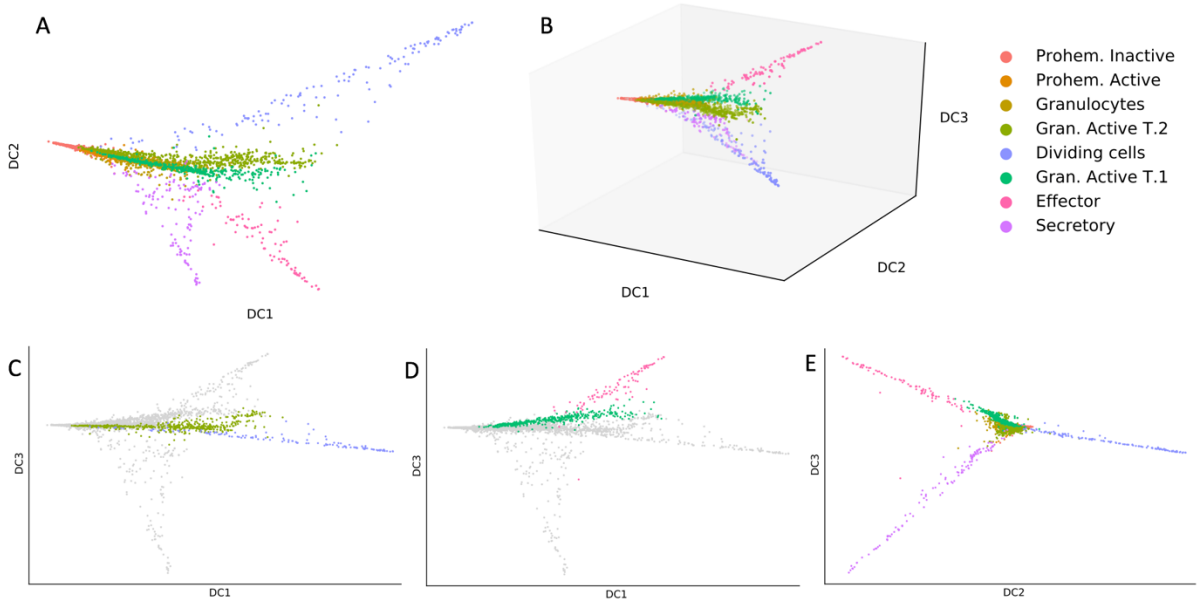

Fig. III.22 Diffusion maps confirm hemocyte lineages. (A) 2D diffusion map of granulocytes **(B)** 3D diffusion map of granulocytes **(C)** Diffusion Component 1 (DC1) vs DC3 plot highlights transition between dividing cells and granulocytes T.2 **(D)** DC1 vs DC3 plot highlights transition between effector cells and granulocytes T.1 **(E)** DC2 showcases hemocyte maturity, with proliferating cells on the right and differentiated states on the left.

Lastly, hemocyte lineages were also confirmed with the lineage analysis software package Slingshot, another highly rated lineage tracing software. It does not perform as well as PAGA when dealing with complex dataset containing multiple separate lineages, but it does work well in branching analyses [311]. As such, we subset our dataset to only include the three interconnected granulocyte-prohemocytes branches, and then run Slingshot. The results confirmed PAGA and diffusion maps findings. Slingshot identified three separate lineages originating in the inactive, baseline prohemocytes, moving into active prohemocytes and standard granulocytes, before branching alternatively into Type 2 active granulocytes and dividing granulocytes, or into Type 1 active granulocytes and then effector or secretor cells. Cells were ordered along a pseudotemporal dimension showing the differentiation of each hemocyte lineage. Pseudotime reconstruction was comparable between Slingshot and diffusion maps, with in blue baseline inactive prohemocytes, and in yellow the terminal effector states or proliferating cells. The central basal granulocyte cluster appeared once again to be the main branching point of the prohemocyte-granulocyte system [Fig. III.23].

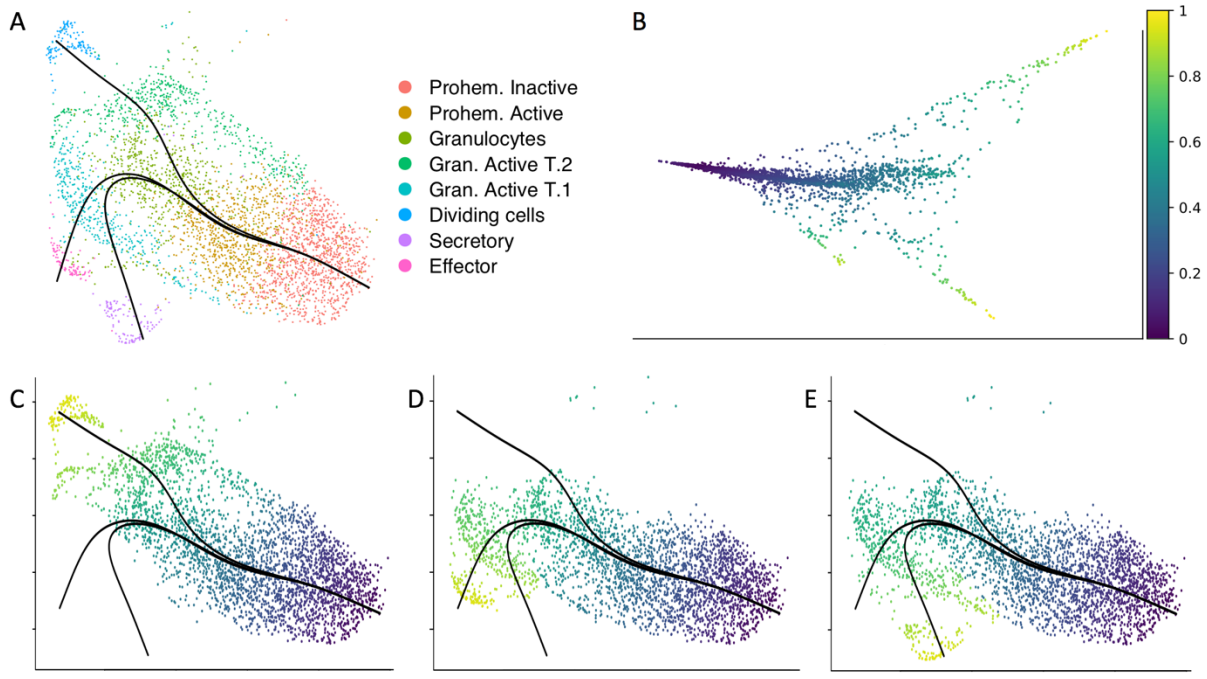

Fig. III.23 Slingshot lineage tracing and pseudotime reconstruction of granulocytes and prohemocytes (A) Slingshot analysis after subsetting non-hemocytes and oenocytoids. **(B)** Pseudotime reconstruction on DC1 vs DC2 **(C-E)** Pseudotime reconstruction with Slingshot for each separate lineages from prohemocytes to **(C)** Dividing **(D)** Effector **(E)** Secretory cells.

Functional classes of mosquito hemocytes

After trajectory identification, generalized additive models (GAMs) were fitted with the package tradeSeq, estimating one smoother per lineage with a negative binomial distribution. A total of 1018 highly expressed genes were filtered for the analyses. The TMM effective library size was internally used as offset by the model, which also allowed to fit zero inflated negative binomial to deal with zero inflation. After filtering for Wald test score >150 and a p-value <0.001 we identified 57 DE genes whose expression changed along lineage 1 (prohemocytes to granulocytes to rapidly dividing), 28 DE genes for lineage 2 (prohemocytes to granulocytes to secretory), and 40 for lineage 3 (prohemocytes to granulocytes to effector cells). Lineage 1 DE genes included PPO6, fibrinogen, cofilin, actin 5C, ARP2/3 complex, and many ribosomal transcripts. Lineage 2 DE genes featured cecropin, LYSC1, collagen Type IV alpha, laminin subunit alpha, cathepsin, LRIM16A, actin 5C, SPARC, class C scavenger receptor. DE genes for lineage 3 were largely similar to lineage 2, further demonstrating their similarity. LITAF3 (LL3), laminin gamma 1, LRIM16B were however specific for this lineage [Fig. III.24]. Overall many marker genes identified with Seurat were also independently found in this independent pseudotime-based analysis.

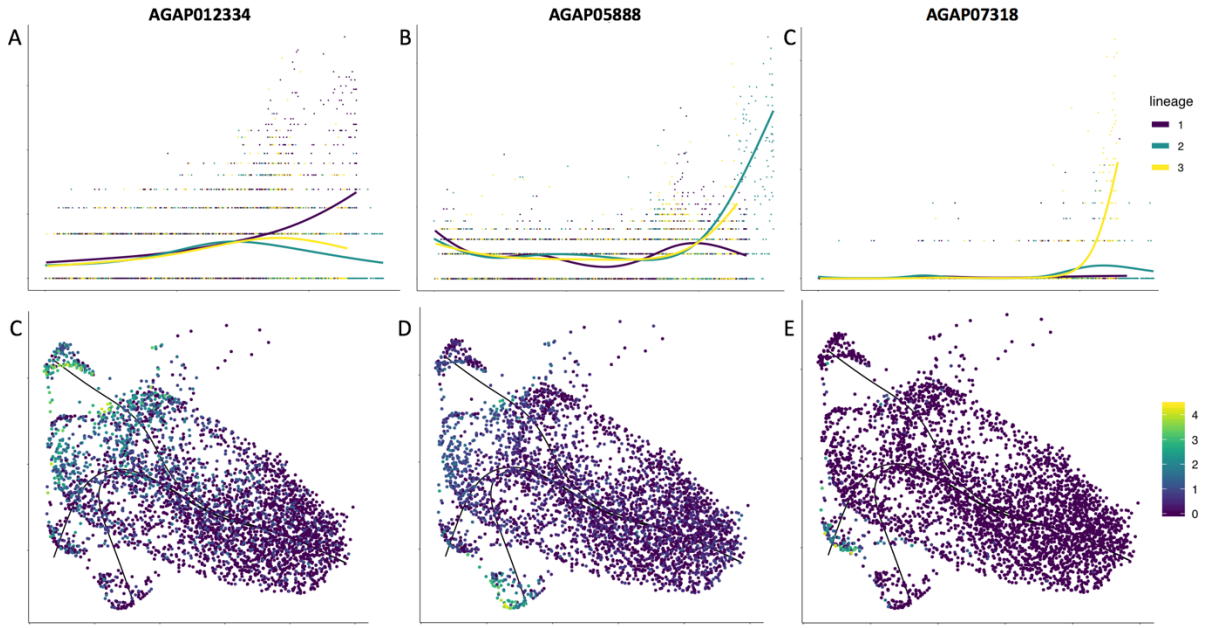

Fig. III.24 DE analysis of lineage-specific genes based on Slingshot pseudotime. (A-C) Smooth curves showing expression by pseudotime for the top three DE genes for each lineage **(D-E)** Corresponding expression of the top 3 DE genes on UMAP of prohemocyte-granulocyte lineage. Blue low transcript counts, yellow highest transcript counts.

Lastly, we analysed correlative microscopy images to help validate our lineage tracing hypotheses. Putative intermediate and early stages of both hemocytes and oenocytoids could be found, defined by a smaller cell size, smaller nuclei, lower expression of marker genes, and rounder morphology. Finally, less mature forms were likely to have less, or be void of, pseudopodia. The images are consistent with a cell development hypothesis that holds prohemocytes as the starting point, before branching differentiated cell types, both for LRR8+ hemocytes and oenocytoids [Fig. III.25 and Fig. III.26].

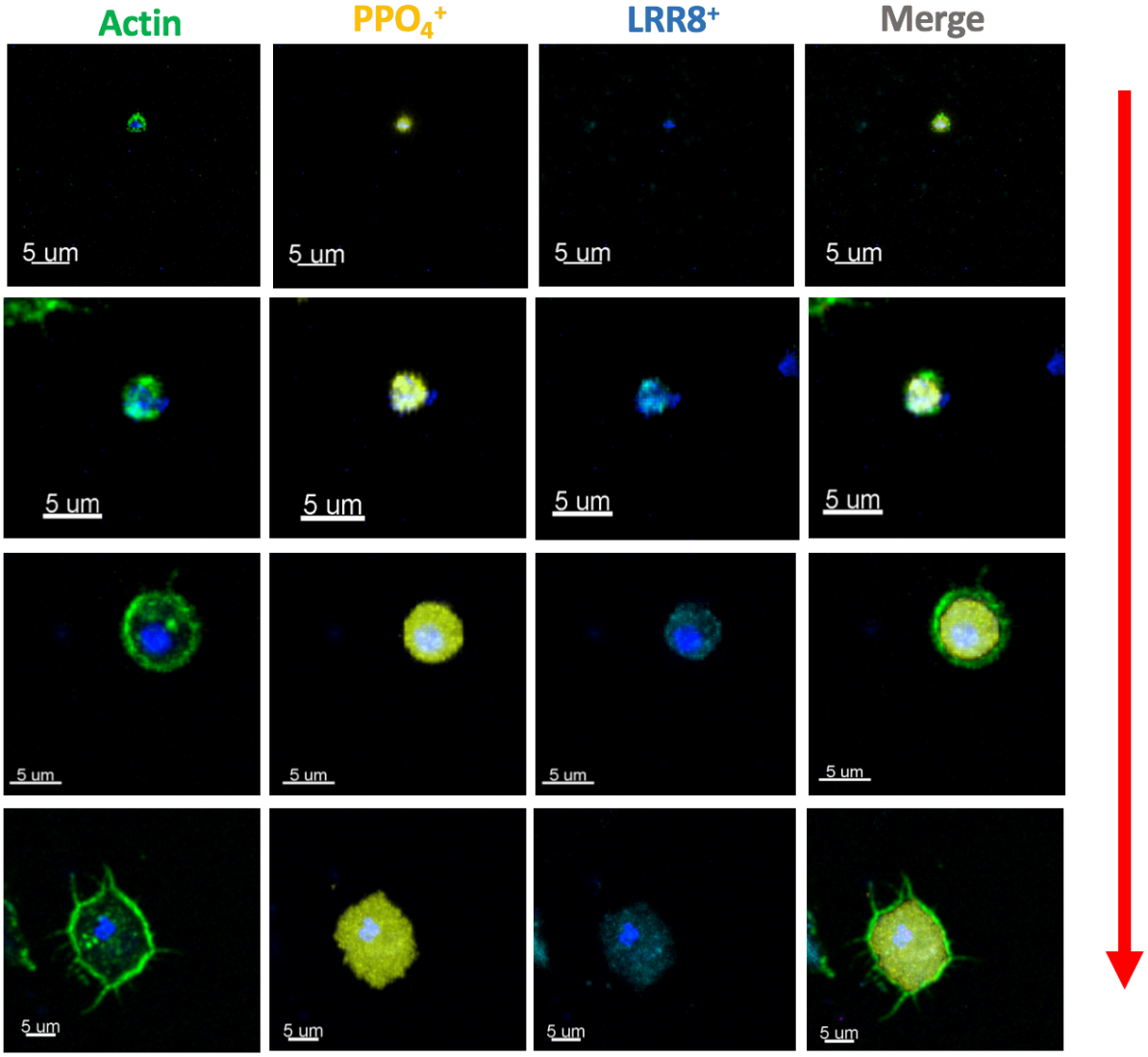

Fig. III.25 Oenocytoid lineage. Red arrow indicates trajectory of maturation. Correlative microscopy. 63x merged, RNA-FISH, and morphological (green, actin) view of circulating hemocytes (blue, LRR8 probe), and oenocytoids cells (yellow, PPO4).

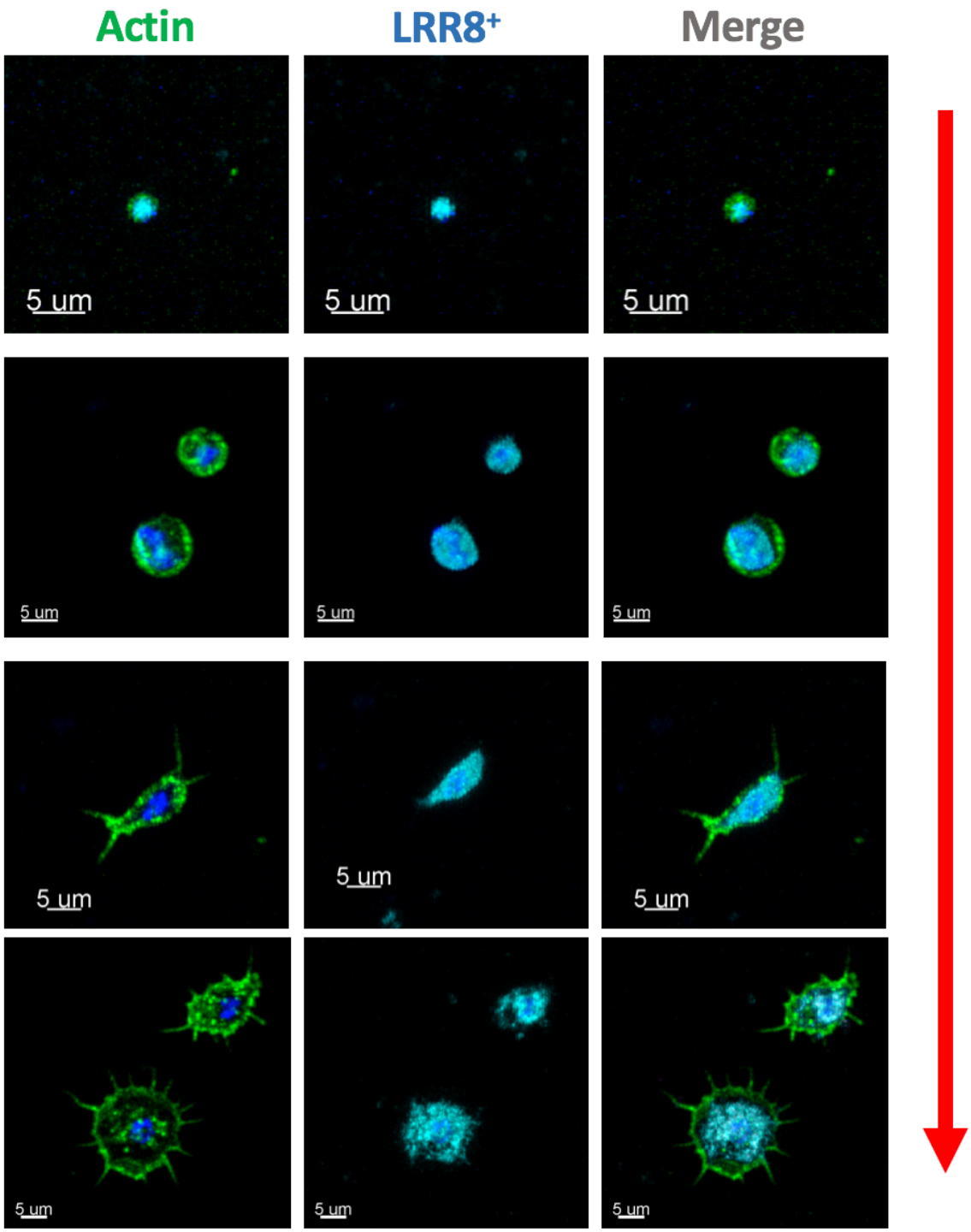

Fig. III.26 Granulocyte lineage. Red arrow indicates trajectory of maturation. Correlative microscop. 63x morphological (green, actin), RNA-FISH (blue, LRR8 probe), and merged view of circulating hemocytes.

3.1.10 Correlation of *Aedes* and *Anopheles* hemocytes

To assess which of the newly discovered putative cell types are shared between anopheline and culicine mosquitoes, we also analyzed the single-cell transcriptome of 3123 cells from *A. aegypti*, a vector for several viral diseases including yellow fever, dengue, chikungunya and Zika. As with *Anopheles*, a dimensional reduction plot shows both canonical hemocytes and other cell types with mostly fat body signatures [Fig. III.27-28]. We once again identified canonical oenocytoids (two subtypes, HC1 and HC2), granulocytes (HC4 and HC5), prohemocytes (HC3), dividing granulocytes (two subtypes, HC6 and HC7), secretory granulocytes (HC8). Fat body cells were characterised by a heightened complexity, with five different cell states recognised (FBC1-5).

A cross-species correlation after a logistic regression and multinomial learning approach further supported our cell type identification, and revealed similarities and differences with *Anopheles* hemocytes. Two clusters (AaHC1 and AaHC2) both have conserved transcriptome signatures for oenocytoids compared to *Anopheles* oenocytoids (AgHC1): 99% and 77% correlation respectively. We again detected different granulocyte subtypes, including antimicrobial peptide secreting cells (94% correlation with *Anopheles* secreting granulocytes), and dividing granulocytes (87% with *Anopheles* progenitor cells). Granulocytes and prohemocytes are again positioned on a continuum of transcriptomic similarity, with four different cell states, including a proliferating S-phase granulocyte cluster (AaHC6) without a clear *Anopheles* equivalent. Granulocytes once again express laminins, leucine-rich repeat proteins, scavenger receptors, Toll receptor 5, and the transcription factor Rel2 [Fig. III.28]. However, effector cells (AgHC5) lack an obvious counterpart in *Aedes*.

Functional classes of mosquito hemocytes

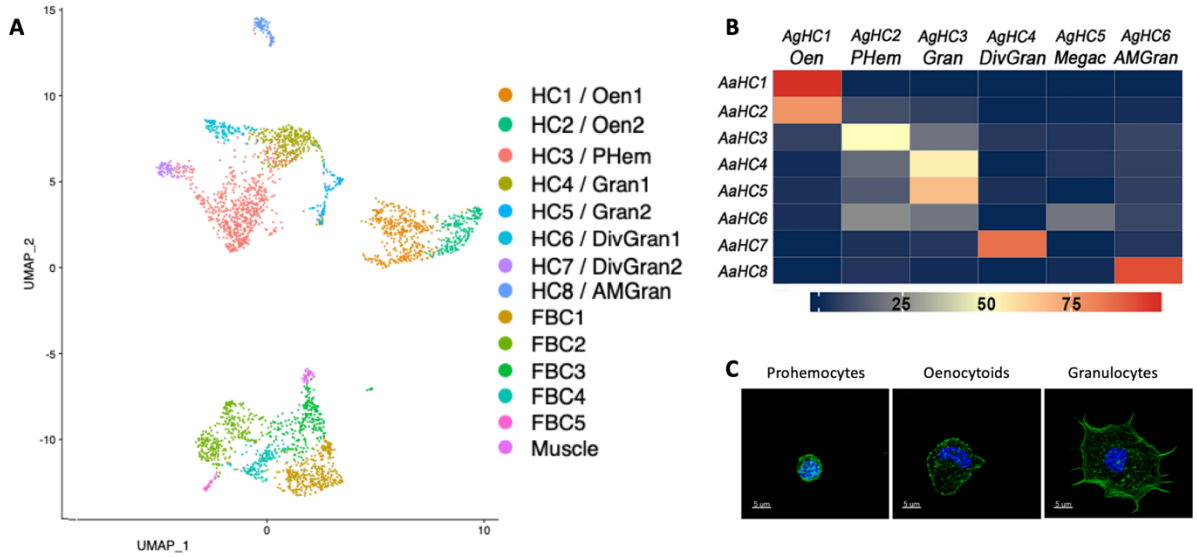

Fig. III.27 Characterisation *Aedes aegypti* hemocytes and correlation with *Anopheles*
(A) UMAP of 3123 *A. aegypti* hemocyte clusters colored by cluster identity with Seurat clustering. **(B)** Heatmap showing probability of each *A. aegypti* hemocyte cell in the cluster belonging to each one of the *Anopheles* cell types after logistic regression and multinomial learning approach. *Ag*, *Anopheles*; Aa, *Aedes*. Oen, oenocytoids; Div Gran, dividing granulocytes; Gran, granulocytes; Mega, megacytes (effector); AM Gran, secretory granulocytes; PHem, prohemocytes. **(E)** *Aedes* hemocyte morphology. Stained with phalloidin (actin) in green and Hoechst (nuclei) in blue. Scale bar: 5 µm.

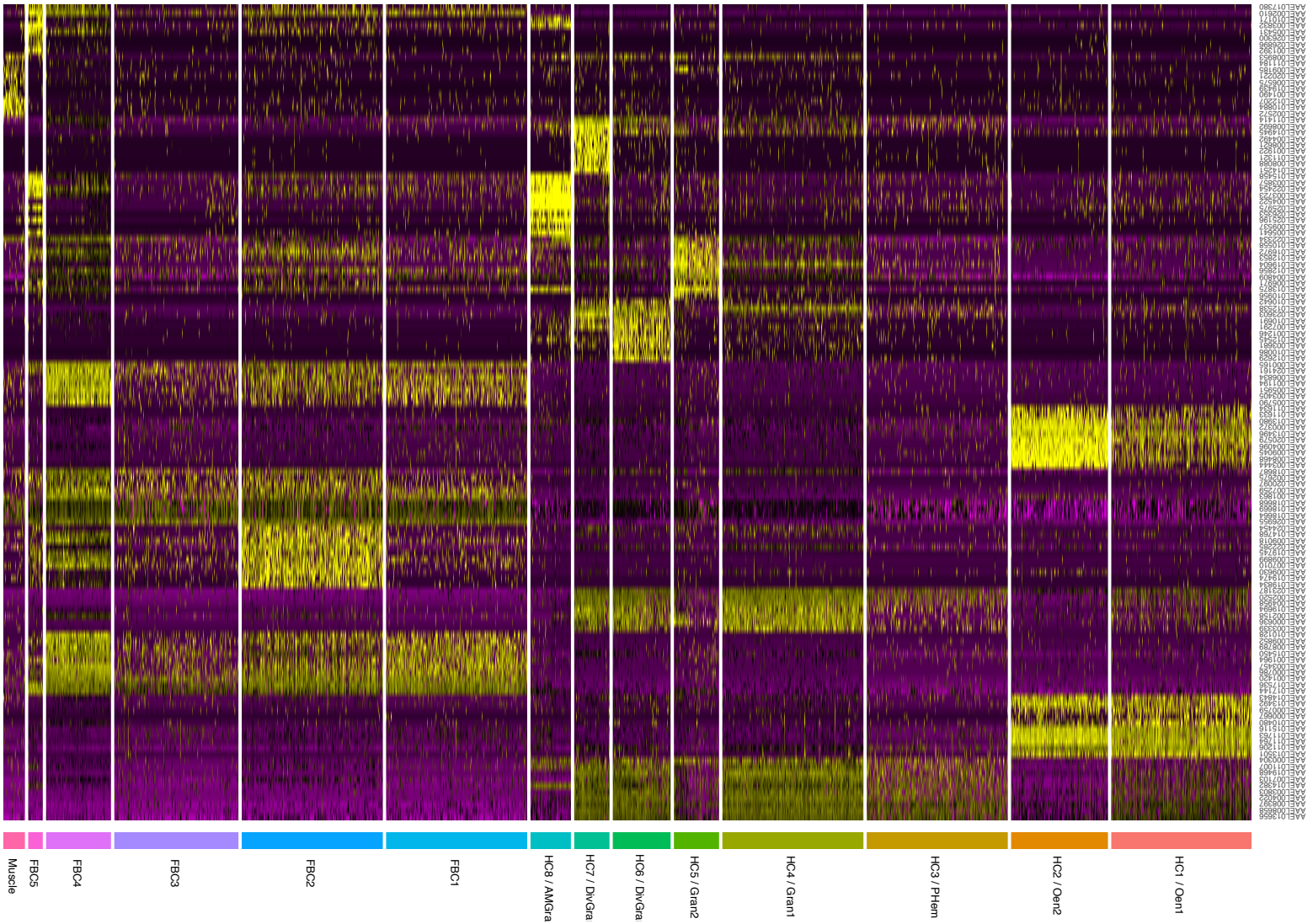

Fig. III.28 Heatmap of top ten gene biomarkers for each *Aedes* cell type or state. DE genes were identified with the Wilcoxon rank-sum test. The P values were adjusted for multiple testing using the Bonferroni correction. P-adjusted values < 0.001, ordered by average log fold change between cluster of interest vs all other cells. Down-sampled to 300 cells per cluster for clarity.

4 Discussion

Clustering analysis with Seurat, diffusion maps, lineage tracing with PAGA and Slingshot, and RNA-FISH validation make us posit that 6 hemocyte cell types exist in the hemolymph of mosquitoes. These include three main types already known: prohemocytes, granulocytes, and oenocytoids. In addition, we found novel cell types, namely dividing hemocytes, effector hemocytes, and secretory hemocytes. We classified cell types by taking into consideration both the RNA content of cells - using the number of UMIs per cell as a proxy - as well as the analysis of the differentially expressed genes between each cell cluster. Prohemocytes were characterised by a low number of UMIs (yet distinct from background), consistent with the high nuclear-cytoplasmic ratio and small overall size. Conversely, granulocytes were transcriptionally active, had large diameters, and exhibited high UMIs. Oenocytoids were intermediate in size, RNA content and number of UMIs.

Furthermore, when looking more in detail into cell expression, prohemocytes split into two main cell states within their larger group: inactive and active prohemocytes. Granulocytes showed the largest diversity, compatible with their effector functions. They subdivided into baseline, Type 1 and Type 2 active granulocytes. While baseline granulocytes were well represented in sugar-fed, blood-fed, and infected conditions, that was not the case for Type 1 and Type 2 active granulocytes, which were enriched in blood-fed and *P. berghei* infected conditions. Blood feeding has been shown to activate and induce granulocyte proliferation, in keeping with our results[147]. Thus, T.1 and T.2 granulocytes appear to be activated granulocyte states, and lineage tracing analysis indeed suggests they are alternative granulocyte activation trajectories. Whereas Type 1 active granulocytes appeared to give rise to dividing cells, the other differentiation branch split from baseline granulocytes into Type 2 granulocytes and then effector or secretory hemocytes. Indeed, effector hemocytes were characterised by high expression of LITAF (LPS-Induced TNF-alpha transcription factor), AGAP007318 (an uncharacterised membrane protein upregulated with *P. berghei* infection [349]), Toll proteins, and ficolins. LL3 had been previously shown to control oocysts numbers, but the cell population responsible for the phenotype was unknown [186]. We hypothesize these cells to

be the elusive immune effectors responsible for *Plasmodium* oocyst control. Secretory hemocytes on the other hand constitutively expressed proteins with N-terminal signal peptides for secretion either into circulation or lysosomes, such as LYSC1, TEP3, Ficolins, cecropins, and defensins. Granulocytes, oenocytoids, prohemocytes could be found both in circulation as well as in sessile form, and the same for effector and secretory hemocytes. We did not however find dividing cells in tissues. It is possible replicating granulocytes exist only briefly in this cell state, or alternatively that only circulating hemocytes replicate.

Genes of interest that should be followed up include AGAP009201, encoding for the collagen type IV, highly expressed in circulating hemocytes and the basal lamina and shown to be important to reduce oocyst load, to increase phagocytic capabilities of hemocytes, and to modulate LRIM1 [324]. In our study AGAP009201 was highly expressed in prohemocytes and all granulocytes, including dividing cells. LRR (AGAP004017 and AGAP004016) are leucine-rich repeat proteins highly expressed in circulating hemocytes (in our data in all hemocytes, including some oenocytoids). Of interest AGAP004016 was shown to be a *Plasmodium* agonist [324]. Both LL3 and LL1 are highly expressed in effector hemocytes and are part of the LITAF family (LPS-induced tumor necrosis factor alpha factor) and have important roles in *Plasmodium* control and immune modulation [185]. AGAP011223 was one of the top genes in oenocytoids and encodes fibrinogen-related FBN8 (FREP57), which was shown to promote phagocytosis and have a role in anti-*Plasmodium* defences [324]. Finally, among cell cycle genes and transcription factors we have NF-X1-type zinc finger protein NFXL1, orthologue to *Drosophila* 'nessun dorma', a top gene marker for dividing cells, but with an unknown role in hemocyte replication [350].

There likely exist four distinct hemocyte lineages in the mosquito. Two main lineages, the prohemocyte – granulocyte lineage, and the oenocytoids lineage, are distinct as shown by clustering, lineage tracing analyses, and correlative microscopy. Prohemocytes have long been thought to be the stem cells of the mosquito immune system. In this dataset there was no direct evidence for prohemocytes to be stem cell-like, but prohemocytes do appear to be a pool of

inactive, immature immune cells that the mosquito can draw upon when challenged, or when overloaded with nutrients such as after blood-feeding. Under these conditions, cell activate and replicate. We saw cellular activation shifts in all cell types, with prohemocytes becoming active prohemocytes and granulocytes. Baseline granulocytes morphed into two active subtypes, which also functioned as intermediate stages before terminal effector and secretory cells, and dividing cells. It appears thus more likely that with blood-feeding and infection granulocytes undergo a rapid activation and replication, and that prohemocytes are recruited at the same time to also become active granulocytes, some of which can then go on to replicate. Whether these replicating and active cells can return to an inactive prohemocyte state is yet unknown, and we did not find direct evidence for replicating stem cells in our *Anopheles* dataset. In the correlative experiment dataset however, we did find a large number of small cells (prohemocytes) expressing markers of cell maturity such as LRR (granulocytes) and PPO4 (oenocytoids), supporting microscopically the hypothesis that all hemocyte subtypes, including oenocytoids, derive from prohemocytes.

Recent studies have shown prohemocytes to have phagocytic capabilities and thus to partially resemble granulocytes [192]. Consistently we showed that prohemocytes and granulocytes exist on a continuum of activation and development. The prohemocyte-granulocyte combined lineage split into three subtypes: a) one lineage differentiated from prohemocytes into granulocytes, then active granulocytes type 2 and finally dividing granulocytes, replenishing the granulocyte cell pool after blood feeding, b) two other lineages instead branched off together into active granulocytes type 1 before splitting into effector cells and c) secretory cells. Oenocytoids on the other hand appear to be a completely separate lineage. We did not find evidence of transcriptomic transition between prohemocytes and oenocytoids, but we did find likely transitions between prohemocytes and oenocytoids with correlative microscopy. Prohemocytes are also the smallest of hemocytes, and few genes per prohemocyte could be captured. The transitions could have thus been missed. Importantly, all three lineage tracing algorithms (PAGA, diffusion maps, Slingshots), as well as Seurat agreed with one another, reinforcing our confidence in the hypothesised lineages. PAGA in particular

is well suited to identify connections between cell types in complex datasets. No clusters were removed in the PAGA analysis, and yet the algorithm still correctly identified a transcriptomic relationship between all fat body cells, whereas muscle cells formed a separate cluster of its own. Surprisingly, oenocytoids were also disconnected from all other cell types. Indeed, even when the PAGA threshold was lowered to capture less confident inter-cluster connections, oenocytoids still did not connect to any other clusters, even when fat body cells and hemocytes did. The lack of connection between fat body cells and hemocytes amounts to a positive control, and we thus conclude that oenocytoids and hemocytes either sit on different lineages that likely arose during the embryonic and larval stage, or that the depth of coverage of our dataset did not allow for the connection to be determined transcriptomically, as few transcription factors or lowly expressed genes could be found in prohemocytes. After subsetting the prohemocytes-granulocytes family we then run separate Slingshot and diffusion maps analyses to confirm the data found through PAGA. And indeed, when visualising diffusion component 1 vs component 3 we could observe direct transitions from active granulocytes type 2 to rapidly diving cells, as well as from type 1 granulocytes to effector hemocytes, indicating a differentiation process. Furthermore, DC1 vs DC2 and the 3D visualisation of the first three diffusion components also showed the secretory subtype emerging from granulocytes.

Slingshot – another top-rated lineage tracing software – further supported our hypothesis, recapitulating the differentiation process we had observed with PAGA. A pseudotime analysis of the three branches also showed some of the genes involved in the transitions. Keeping in mind that most cell cycle genes were not included in the lineage analysis due to the strict filtering requirements, many of the genes Seurat identified as markers for each cell type were also independently found in the pseudotime-based analysis. For example, lineage 1, which traces prohemocytes to dividing cells, featured PPO6 and fibrinogen. Of interest, in humans and mice extravascular fibrinogen has been shown to induce macrophage chemokine expression via Toll-like receptor 4, leading to increased immune surveillance at sites of increased inflammation [351]. In our dataset, granulocytes type 2 and many oenocytoids expressed fibrinogen and fibronectin-like transcripts. It may be that these cells are

immunogenic sensors leading to fibrinogen deposition and activation, followed by mitotic division of granulocytes (dividing cells). Lineage 2 genes featured cecropin, LYSC1, collagen Type IV alpha, laminin subunit alpha, once again transcripts that were gene markers of granulocytes type 1 and secretory cells with Seurat. Lineage 3 genes were very similar but LITAF3 (LL3), laminin gamma 1, and LRIM16B were specific for effector cells.

These conclusions were reinforced by the parallel results in our *Aedes* dataset. The cell types originally discovered or confirmed in *Anopheles* were largely conserved between the two species, and thus possibly of functional importance. Because of the increased number of genes per cell we were able to detect more granular details, including two different oenocytoid cell and dividing granulocytes cell states. Interestingly however, effector cells were not detected at all in the *Aedes* dataset. Furthermore, the gene marker (TM7318) defining them is only present in anophelines of the *Cellia* subgenus (malaria vectors in Africa and Asia). We speculate these cells may thus have specific functions in African and Asian Anopheles, potentially connected to immune priming and *Plasmodium* responses (see Chapter IV).

Fat body cells and muscle cells were captured in both species, either because they naturally slough off into the hemolymph, or because the shear stress of the anti-coagulant buffer injection, or the tearing of the abdomen, dislodges them. Fat body cells had two main transcriptomic states: baseline and active. The active fat body cell was highly metabolic, characterised by the expression of canonical markers such as vitellogenin. Conversely, baseline fat body cells expressed a plethora of immune genes, both pro and anti-inflammatory, although many of the top markers are known for dampening the immune system. Inactive fat body cells were characterised by high expression of CLIPs, lectins, LRIMs, APL1C, and SRPNs, in addition to regulatory genes of the PPO cascade, such as apolipophorins and phenoloxidase inhibitor protein. This cell type appears to specifically express *Plasmodium* infection agonists. For example, CLIPA9 expression increases oocystsis load [352], and both CLIPA7 and CTL4 stop parasite melanisation [353]. LRIM17 is downregulated after infection to activate an effective immune response [354], and LYSC1 and CLIPA14 knock-down mosquitoes exhibit

increased resistance to *P. berghei* and bacterial infections [355, 356]. SRPN2 also appears to aid malaria parasites [357]. Interestingly, with blood-feeding or infection there was a shift towards a metabolically active, and immunologically permissive fat body. The loss of immune inhibition by the fat body and the concurrent activation of immune cells in the hemolymph suggests the mosquito immune response is tightly integrated with its metabolic functions, with different organs interacting to provide an optimal immune response at each phase of the mosquito life.