Chapter 3

Decomposing the WOM: from tissue to single-cell.

The material presented in this chapter has been previously published in reference[308]. I confirm I have ownership of the copyright for its reproduction. All the work presented here has been done in collaboration; the experiments that led to the collection of cells and tissues were performed by others, but I have solely analysed all the data generated. The results presented are my own work unless otherwise stated.

The data presented in the previous chapter provides a generalised profile of the transcriptome of the two most prominent and better studied components of the mouse olfactory system. Such gene catalogues, however, represent the averaged expression of each gene across the different cell types found within each tissue. The sensory neurones are the primary focus in this dissertation, given their role in olfactory signalling, and particularly the receptor repertoire. Therefore, I established a collaboration project with Peter Mombaerts and Mona Khan to specifically isolate the OSNs of the MOE. For this, a transgenic mouse was used, that is engineered to express GFP from the OMP locus (OMP-GFP)[309]. OMP is a protein specifically expressed in the sensory neurones from the MOE and VNO[310], and it is not expressed in sustentacular or basal cells[311]. Its expression can be identified after both OR genes and *Adcy3* have been activated. Additionally, the expression domain of *Omp* doesn't overlap with *Gap43*, a marker of immature OSNs. Therefore, OSNs that express OMP are considered mature[312]. Analysis of an OMP KO animal revealed that the general structure of the MOE is unaffected and

the ratio of mature to immature neurones also remains normal; but the MOB is 15% smaller. EOG recordings upon stimulation with several odours resulted in decreased recordings, with slower kinetics for both the initiation and recovery of the response. Nonetheless, OSNs were still able to be activated and mice were not anosmic[313]. The OMP-GFP mouse is a knock-in, where the CDS of *Omp* was substituted for GFP. To avoid altering the response kinetics of the OSNs, all experiments were conducted in heterozygous animals.

3.1 The transcriptome of the olfactory sensory neurones.

To characterise the transcriptome of the OSNs alone, dissociated cells from the WOM of heterozygous OMP-GFP mice were separated based on their expression of GFP, by fluorescence-activated cell sorting (FACS). The GFP^+ –and therefore OMP^+– cells were retained, and the populations from several animals were pooled to obtain 10 million cells. This is the number of mature OSNs typically found in a single adult mouse[224]. Three biological replicates were collected and the RNA extracted from the OSNs was used to construct libraries for RNAseq (Table B.1 in Appendix B). Since the OMP-GFP mouse is in a mixed genetic background (129P2 X B6), three WOM samples from single animals were also processed for comparison[1]. All six samples were sequenced at high depth on the Illumina platform and data was analysed including the extended models for receptor genes (Table B.2 in Appendix B).

As observed previously, biological replicates were highly correlated (rho $>$ 0.96, p-value $<$ 2.2e-16) for both the WOM samples and the pooled OSNs. In order to reveal genes that are preferentially expressed in the OSNs or in other cell types, I performed a differential expression (DE) analysis. 67.6% of all the genes expressed were significantly different between the OSNs and the whole tissue (FDR $<$ 5%), with 45.8% of these being expressed more abundantly in the OSNs (Figure 3.1). From these, 790 have a fold-change greater than 3 and 50.1% are OR and TAAR genes. To explore their functions, I performed a gene ontology (GO) analysis that revealed enrichment for terms related to the olfactory transduction pathway and G-protein coupled amine receptor activity, along with genes related to synaptic vesicles, branching morphogenesis of a nerve and peptide hormone processing. In contrast, 5,227 genes were expressed higher in the WOM

[1]The collection and RNA extraction from WOM samples and pools of FAC-sorted OSNs were performed by Mona Khan.

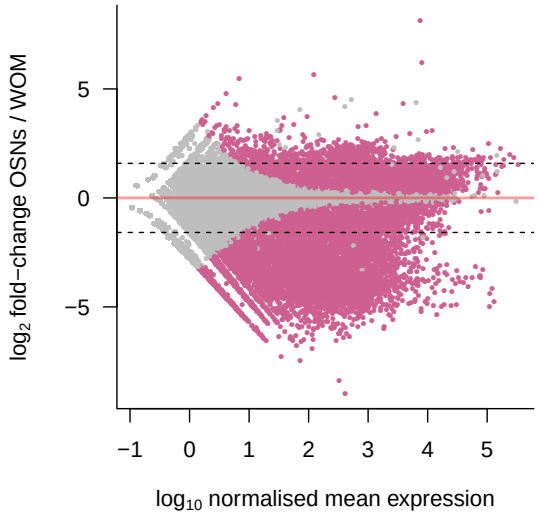

Figure 3.1 – Differentially expressed genes between the OSNs and WOM. MA plot showing the mean expression for each gene against its corresponding fold-change value between the OSN and WOM samples. The red line represents equal expression in both groups. Significantly DE genes are in pink (FDR < 5%). Dotted lines represent a threshold of fold-change of ± 3.

(fold-change < 0.33, FDR 5%), and over half of these were expressed at least ten times higher than in the pooled OSNs; this suggests that they are likely restricted to the non-neuronal cell types found within the WOM samples.

A similar study was performed by Sammeta et al.[314] using microarrays, to compare the transcriptomes of the GFP^+ versus the GFP^- populations of an OMP-GFP mouse. The genes that were reported as enriched in the OMP^+ population in this dataset tend to be more abundant in the pooled OSN samples; and the genes expressed specifically in the WOM are classified as enriched in the OMP^- population (Figure 3.2), thus supporting the accuracy of the DE analysis results.

Next, I compared the expression of the receptor repertoire. As expected, both OR and TAAR genes were generally expressed at higher levels in the OSNs than in the WOM. There was a median fold-change increase of 2.56 in overall expression, and a similar increment is observed for the canonical markers of mature OSNs (Figure 3.3). The overall expression values were highly correlated (rho=0.95, p-value < 2.2e-16), which indicates that the repertoire's expression increases as a whole, maintaining the proportions between receptors observed in the whole tissue. Only 19 OR genes that are present in the WOM samples are lacking in the OSN pools. However, all these are expressed at very low levels and 13 (68.4%) are annotated pseudogenes. As observed previously, nearly the complete repertoire of OR genes is expressed, as assessed by unique counts. Over 95% of the OR genes have at least one fragment mapped in both the WOM and

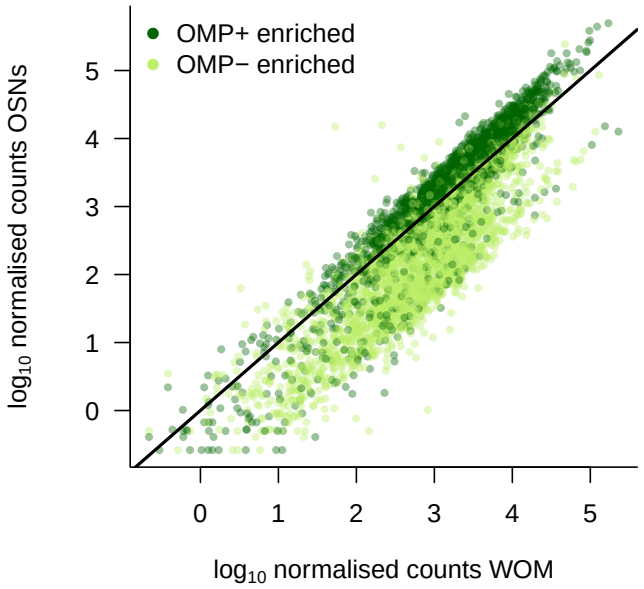

Figure 3.2 – **Comparison to Sammeta et al.** Scatter plot of the expression values in the WOM versus OSN samples. Genes are coloured depending on whether Sametta et al. classified them as enriched in the OMP^+ cells (dark green) or the OMP^- cells (light green). The line indicates the 1:1 diagonal. Genes above the diagonal are expressed higher in OSNs and tend to be enriched in the OMP^+ sample, whereas genes below the diagonal are expressed higher in the WOM and are enriched in the OMP^- sample.

OSN samples; this number increases to 98.9% if only the ORs annotated as functional are considered.

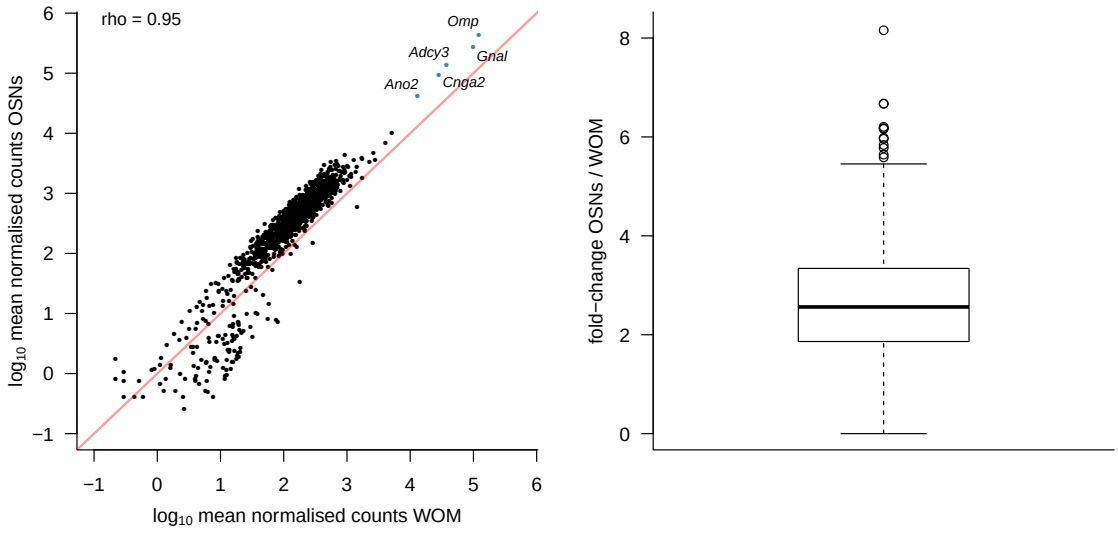

Figure 3.3 – **Receptor expression in the WOM vs the sorted OSNs.** On the left, scatter plot of the expression of all OR and TAAR genes (black) in the WOM versus the sorted OSNs, and of five canonical markers of mature OSNs (blue). The red line represents the 1:1 diagonal. The majority of genes are expressed more abundantly in the sorted OSNs. On the right, a boxplot of the fold-change between the OSNs and the WOM for the receptor expression levels. The median increase in expression is 2.56.

All together, I have characterised the transcriptome of the neuronal component of the MOE, revealing many genes that are preferentially expressed in OSNs. Additionally, thousands of genes can now be classified as specific to the other cell types of the MOE and surrounding tissue. The deep RNAseq strategy used is effective in characterising the expression of the whole receptor repertoire in this transgenic mouse strain also. By separating the OSNs from all the other cell types, the depth of sequencing devoted to receptor genes is increased, but the proportionality of the different neuronal subpopulations present within the MOE is conserved. The very high correlation between the expression estimates for the receptors in both sample types argues in favour of their accuracy and reproducibility.

3.2 Mature OSNs segregate into two distinct populations.

When cells were dissociated from the WOM of OMP-GFP animals and sorted based on their GFP intensity, it was evident that two different populations of GFP^+ cells are present in the samples (Figure 3.4). Cells from one population have much more intense fluorescence levels than the others, so I will refer to these as the GFP^{high} and GFP^{low} populations. To investigate further this distinction, pools of 10,000 cells from each population were obtained from three different animals for RNAseq (Table B.1 in Appendix B)[2]. The expression estimates between replicates were highly correlated ($0.87 < \text{rho} < 0.91$, p-value $<$ 2.2e-16) despite the smaller number of cells used for RNA extraction. However, one of the GFP^{high} samples yielded only a small number of sequencing fragments compared to the others, and it was excluded from downstream analyses (Table B.2 in Appendix B).

As a first control, I examined the levels of *Omp* expression in each population. In all samples there was clear, robust expression of this marker gene, but the GFP^{high} samples had a consistent 1.52 fold increase in expression (p-value = 0.02482, t-test; Figure 3.5A). All the GFP^+ cells should have the transcriptional profile of mature OSNs, since they robustly express *Omp*. To confirm this, I compared the RNAseq data profiles to a microarray dataset that characterised 670 and 565 genes as preferentially expressed in mature and immature OSNs, respectively[315]. Both the genes enriched in mature and immature OSNs were expressed at similar levels in both the GFP^{high} and GFP^{low}

[2]Sorting experiments, RNA extraction and library preparation were performed by Luis Saraiva.

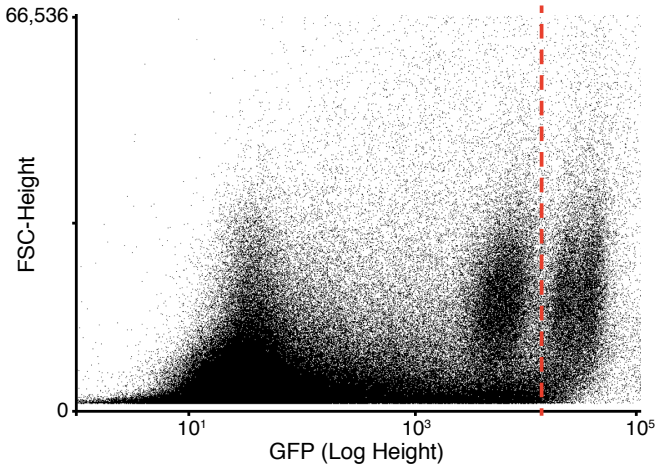

Figure 3.4 – FACS plot of OMP-GFP animals. Each dot is a cell plotted according to its GFP fluorescence intensity (x-axis). For the GFP$^+$ cells, two populations are visible. The red dotted line roughly separates them. Image kindly provided by Luis Saraiva.

populations (p-value = 0.59, t-test), but the genes from mature OSNs were expressed significantly higher than the ones from immature neurones (p-value = 1.73e-07 for GFP^{low}, and p-value = 1.35e-07 for GFP^{high}, t-test; Figure 3.5B).

To identify other genes that differentiate these two populations of OSNs I performed a DE analysis. This revealed 537 significantly DE genes (FDR 5%), 420 (78.2%) of

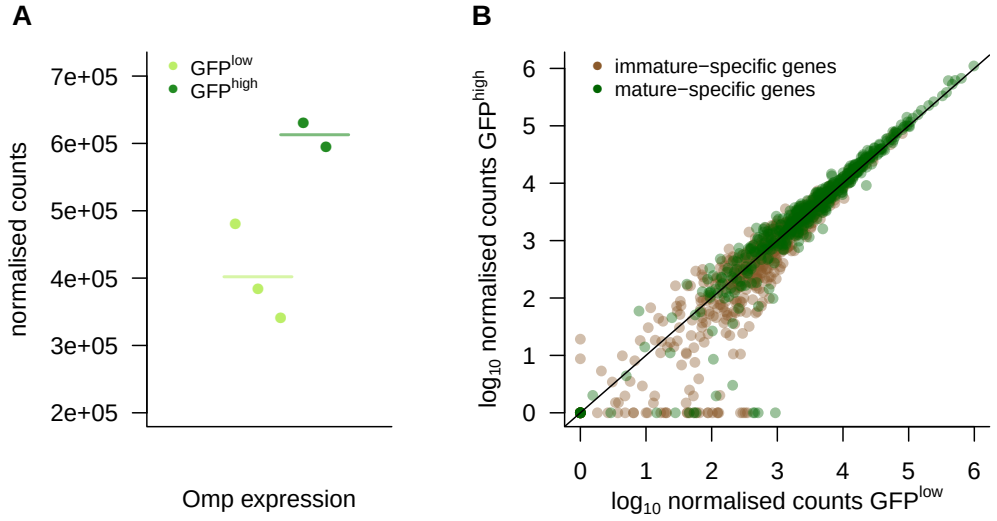

Figure 3.5 – GFP$^+$ neurones are mature. A) Both the GFP^{low} and GFP^{high} populations express high levels of *Omp*, but the GFP^{high} samples have 1.52 times higher expression. **B)** Scatter plot of the normalised expression in the GFP^{low} versus the GFP^{high} samples. The line indicates the 1:1 diagonal. Genes are coloured according to whether they were characterised as enriched in mature (green) or immature (brown) OSNs. All genes are expressed at similar levels in both populations, but the genes from mature OSNs are expressed at higher levels than those from immature neurones.

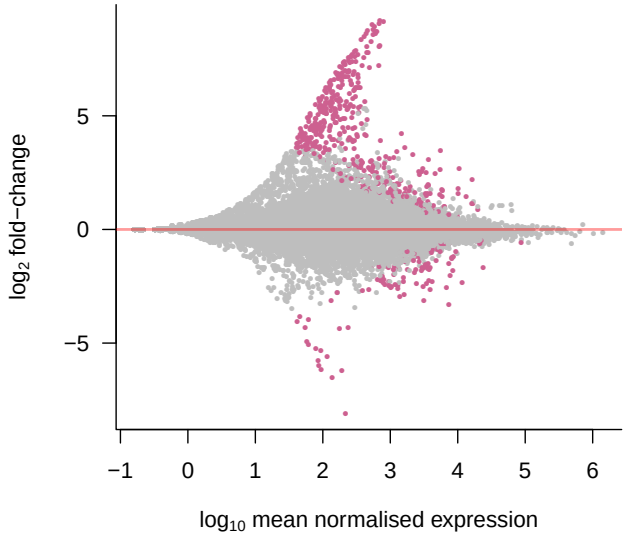

Figure 3.6 – Differentially expressed genes between the GFP^{low} and the GFP^{high} cells. MA plot showing the mean expression for each gene against its corresponding fold-change value between the GFP^{low} and the GFP^{high} samples. The red line represents equal expression in both samples. Significantly DE genes are in pink (FDR $< 5\%$).

which were more highly expressed in the GFP^{low} population (Figure 3.6). This set of genes is enriched for terms related to development, morphogenesis, negative regulation of neuronal differentiation and positive regulation of cell proliferation. Therefore, it appears that the different levels of *Omp* correlate with the level of maturation of the cells. The GFP^{low} cells are still in the process of downregulating genes involved in proliferation and have yet to achieve final differentiation.

These data indicate that there exists a previously unknown subdivision of mature OSNs characterised by the levels of *Omp* expression. While both populations have a characteristic expression profile of mature OSNs, the neurones from the GFP^{low} population are less mature than the rest.

3.3 RNAseq of single OSNs.

I have demonstrated that the mouse MOE is composed of over a thousand different subpopulations of OSNs, each characterised by the OR gene they express. We were interested to know how homogeneous are the transcriptomes of neurones from these different subpopulations. To this end, a Fluidigm C1 microfluidic system was utilised to

capture single OSNs from the FAC-sorted GFP^{high} population from one mouse[3]. This system uses a fluidic circuit to isolate single cells into individual reaction chambers, that can be examined under the microscope to ensure that a single cell is present. After capture, the system performs all the necessary reactions to produce cDNA from each cell (cell lysis, mRNA reverse transcription and PCR amplification) that can then be used to prepare libraries for RNAseq (Table B.4 in Appendix B). From the 96 wells in the capture chip, 58 contained single cells, 8 were empty and the remaining contained more than one cell and/or had visible debris contamination (Figure 3.7A). Only single cells were considered in further analyses.

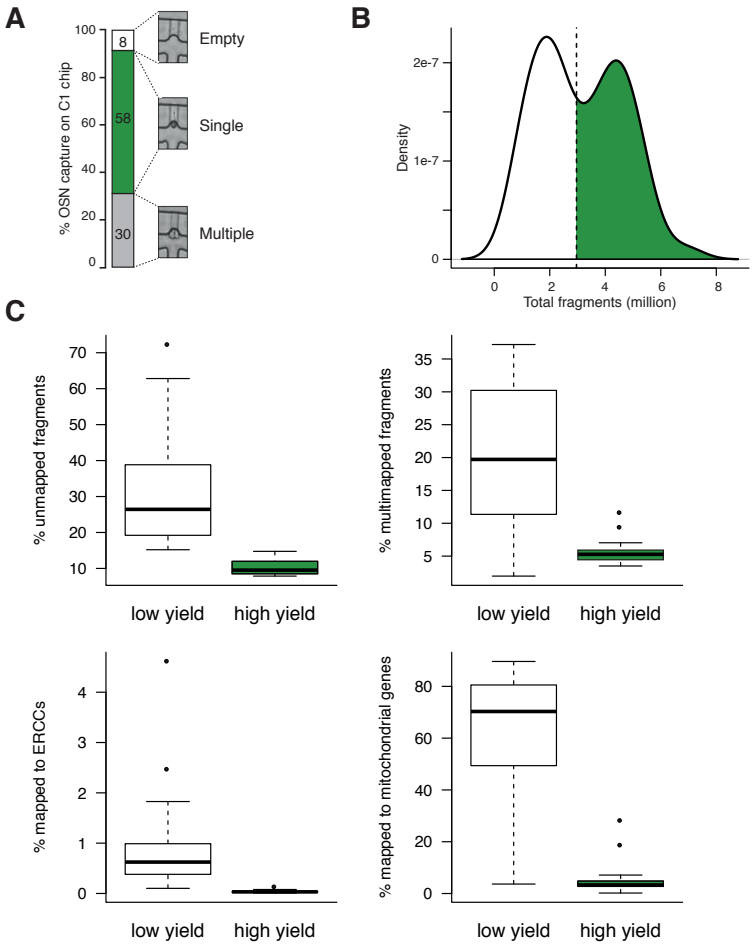

Figure 3.7 – **Quality control of the single-cell RNAseq data. A)** Proportion of wells in the C1 capture chip that were empty (8), contained one cell (58) or more than one cell and/or debris contamination (30). Representative bright field images of each case are shown. **B)** Density distribution of the total number of sequencing fragments obtained for the 58 single-cell samples. Decomposition of the distribution into two normal-like distributions splits the data at 2.96 million (dotted line), separating the samples into those with low (white) and high (green) yield. **C)** Several mapping statistics are clearly different for the low and high yield samples. Schematic in A was kindly provided by Luis Saraiva.

[3]All the capture and library preparation experiments were performed by Luis Saraiva.

The quality of the RNAseq data produced from a single cell is greatly influenced by the amount and integrity of the starting material. Including cells with poor quality data in normalisation steps can be detrimental to the downstream analyses of cells with good quality data; therefore, it is imperative that these are identified and excluded at an early stage[316]. The distribution of total fragments obtained from each of the 58 single-cell samples was clearly bimodal (Figure 3.7B). Deconvolution into two normal-like distributions revealed 28 cells with low yield (mean of 1.7 million) while the remaining 30 were sequenced at significant higher levels (mean of 4.4 million; p-value $<$ 2.2e-16, t-test). To determine if the lower yield was a result of sequencing poor quality libraries, I analysed the mapping statistics (Figure 3.7C). The low-yield group of samples had a much higher percentage of unmapped fragments (31.97% on average versus 10.21% in the high-yield group; p-value = 5.739e-08, t-test) as well as multi-mapped fragments (20.27% versus 5.49%; p-value = 4.018e-08, t-test). The proportion of fragments that mapped to ERCC spike-ins was over 20 times higher in the low- versus high-yield groups (p-value = 5.576e-05, t-test). Furthermore, most of the uniquely mapped fragments from the low yield samples aligned to mitochondrial genes (on average 60.53%, compared to 4.77% in the high-yield samples; p-value = 6.327e-12, t-test). Together these suggest that the starting material for the samples with low yield was of poor quality and were therefore excluded. I focused subsequent analyses on the 30 high-yield samples.

Next, I compared the OR gene expression profiles of these samples as a function of their capture location on the C1 capture chip and the library preparation plate. Ten cells (OSN 157, 178, 185, 191, 207, 214, 218, 223, 255 and 263) had evidence of two highly expressed OR genes. OSN 263 had high counts for *Olfr55* and *Olfr239*, two adjacent OR genes that are 99% identical. Closer inspection of the sequencing data revealed that the fragments assigned to *Olfr239* were in fact mismapped, since a BLAST alignment mapped them back to *Olfr55*. Therefore, I set the counts of *Olfr239* to zero. For the remaining nine cells, in four cases (OSN 178, 191, 223 and 255) I found that one of the OR genes was expressed in another sample located in the immediately adjacent well, suggesting evidence of carry-over. The other five cells (OSN 185, 207, 214, 218 and 257) did not share an OR gene with a sample in an adjacent well. I independently reassessed these nine cells through all previous quality control criteria and could not distinguish the four cells with evidence of carry-over from the five cells without it. A recent report found that up to 20% of cells captured on a C1 microfluidic system contain two cells that are not visible in the microscopy images[317]. Thus, to take a conservative approach and to reduce the possibility of including samples containing a second, visually obscured cell

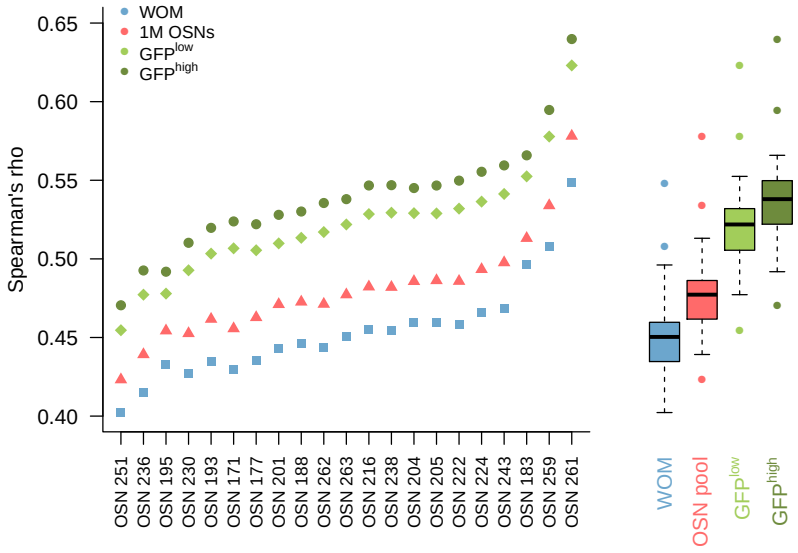

Figure 3.8 – Correlation of single OSNs to other datasets. The Spearman correlation coefficient is plotted for each single OSN (x-axis) against the previous population datasets. The single cells correlate better with the isolated OSNs that with the WOM, and better with the GFP^{high} than the GFP^{low} neurones. On the right, the cumulative data for each comparison is presented as boxplots; dots are outliers.

or contaminating debris in subsequent analyses, we elected to exclude all 9 from further study. This procedure resulted in a final dataset of 21 high-quality single cell samples.

On average, 4.4 million paired-end fragments were obtained for each sample (Table B.4 in Appendix B), 2.7 of which could be mapped to annotated genes. These cover a mean of 4,717 ± 175.8 (SEM) genes per single cell; collectively, 13,582 different genes were expressed in at least one cell, which represents 74.2% of the genes expressed in the GFP^{high} OSNs. To confirm that the sequenced cells indeed correspond to GFP^{high} OSNs, I compared the transcriptome of each single cell to the expression profile of the WOM, the pooled 10 million OSNs and the bulk GFP^{low} and GFP^{high} populations. All single cells correlated better with the OSN samples than with the WOM (p-value < 2.2e-16, paired t-test), and better with the GFP^{high} than with the GFP^{low}(p-value < 2.2e-16, paired t-test; Figure 3.8).

Next, I examined the expression of a set of genes that are canonical markers for the different cell types found in the WOM and the VNO, the other tissue where OMP^{+} cells are abundant. All 21 cells showed robust expression of the genes characteristic of mature OSNs (such as *Omp*, *Gnal*, *Cnga2*, *Ano2* and *Adcy3*), and low or no expression of markers of other cell types, including immature cells and VSNs. Also, none of the 21 cells expressed GC-D or TAAR genes (Figure 3.9). Thus, this confirms that the cells

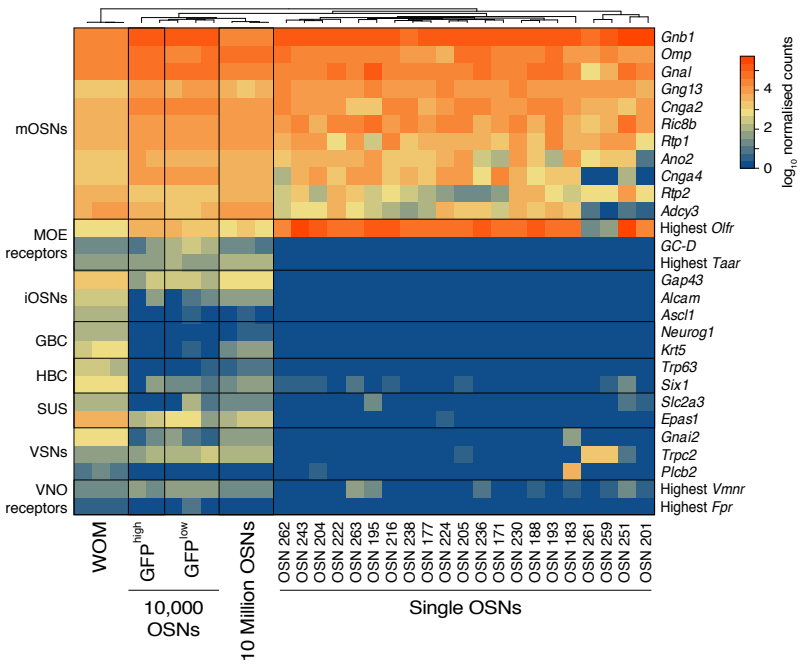

Figure 3.9 – **Expression of canonical markers.** Heatmap representing the expression levels for different genes that are canonical markers for various cell types found in the WOM and VNO. Single cells express all the marker genes of mature OSNs and virtually none of the genes from other cell types. mOSNs, mature OSN; iOSNs, immature OSN; GBC, globose basal cell; HBC, horizontal basal cell; SUS, sustentacular cell.

sequenced correspond to mature OSNs, with no contamination from other cell types.

3.3.1 Heterogeneity between single OSNs.

Single-cell RNAseq data suffers from much higher technical variation than bulk RNAseq, because the amount of starting material is very low. Many genes can have very different normalised counts between technical replicates taken from the same pool of total RNA. For example, a gene can have a hundred to a thousand counts in one replicate and zero in the other; only genes that are expressed at high levels are consistent[318]. Therefore, the variation observed between two different single cells will be a combination of the large technical noise plus true biological variation. Consistent with this, the correlation between any two single OSNs was only moderate ($0.45 < \text{rho} < 0.61$, p-value $< 2.2\text{e-}16$), and there was great variation in the expression levels of low and moderately expressed genes (Figure 3.10).

To address the question of how heterogenous are the transcriptomes of OSNs expressing different OR genes, I calculated the coefficient of variation (CV) for all those genes with more than a thousand normalised counts in at least one single OSN. At this expression level, variation should largely reflect biological variability instead of technical noise.

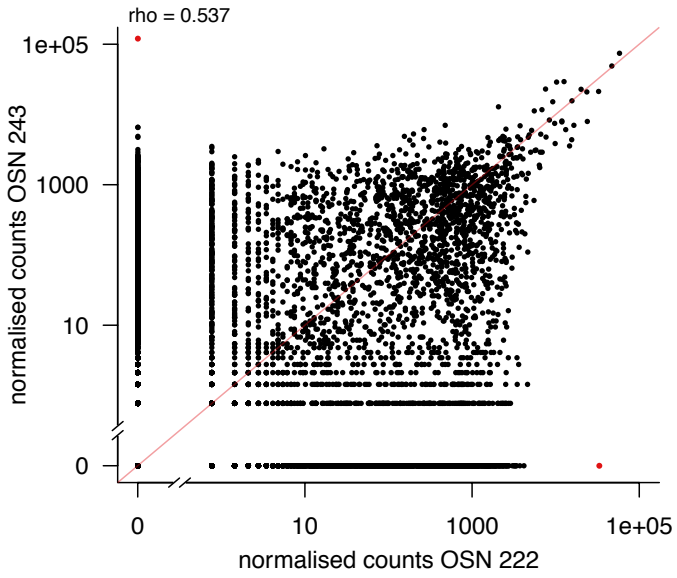

Figure 3.10 – Correlation between two single OSNs. Scatter plot of the normalised counts of two different single OSNs; 0.1 has been added to the counts before computing the logarithm, to be able to show the genes not expressed in one sample. The red line corresponds to the 1:1 diagonal. The Spearman rank correlation coefficient is indicated in the top left corner. In red are higlighted the two OR genes abundantly expressed by each OSN.

From these, 598 showed highly variable expression across cells (CV > 4; Figure 3.11A). These were relatively evenly distributed across the single OSNs (Figure 3.11B) except for one unusually variable cell (OSN 183) which contained 15% of all the genes; these were enriched in GO terms related to chemokine receptor binding, cytokine binding and activity, antigen processing and presentation and regulation of lymphocyte activation, which suggest a stressed cellular state. In contrast, the remaining 509 genes are only enriched in G-protein coupled receptor signalling and transduction. The significance of this term stems from the variable expression of OR genes, but there are also a few orphan GPCRs (*Gpr32*, *Gpr123*, *Gpr125* and *Gpr160*). A similar enrichment analysis on protein domains identified the expected seven-transmembrane receptor domain present in all ORs, but also a significant enrichment for a zinc-finger motif (C2H2 type) and a KRAB box domain, both found within a group of zinc finger proteins (ZFP). Thus, different OSNs are distinguished by the OR gene they express; ZFP genes might also pattern different subtypes of OSNs, but a larger sample is necessary to assess the significance of this finding and whether their expression is correlated with distinct subtypes of OSNs.

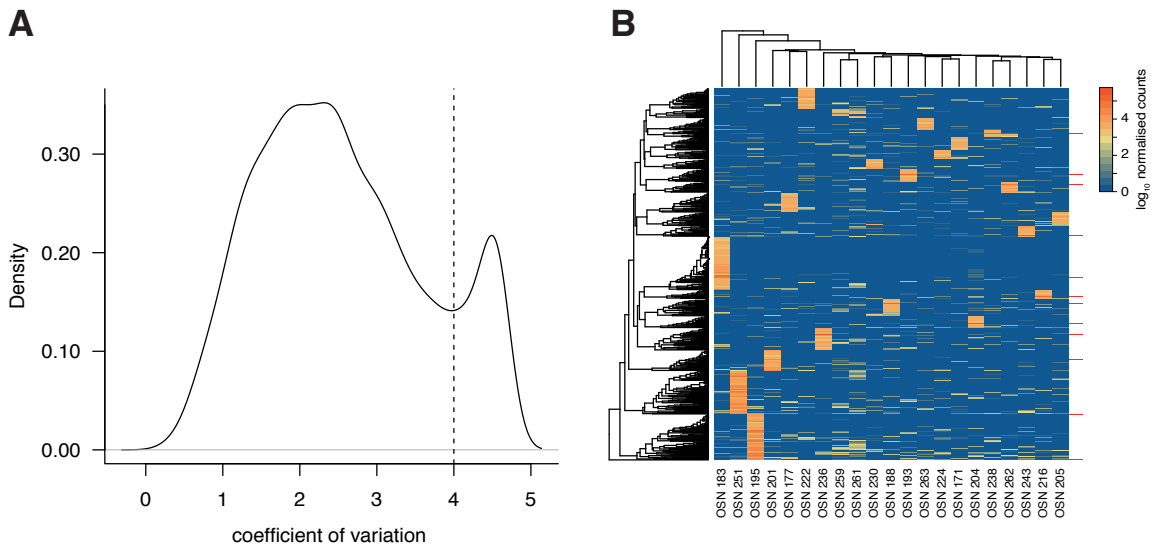

Figure 3.11 – Highly variable genes pattern single OSNs. A) Density distribution of the coefficient of variation (CV) for all those genes that have at least 1,000 normalised counts in at least one cell. **B)** The genes with a CV > 4 are highly variable between cells and only expressed in one or a few samples. To the right of the heatmap, a red line indicates the gene is an OR.

3.3.2 Monogenic expression of OR genes.

The expression of OR genes in OSNs of the MOE is considered to be monogenic. Many experiments have supported this paradigm, but it has never been conclusively proven[74]. Most evidence has come from both single and double *in situ* hybridisation, sometimes in OSNs expressing a receptor tagged with a reporter protein, to show that a given neurone does not express two receptors at the same time. But in all cases only a (very) restricted set of receptors has been tested; the lack of coexpression between these has then been extrapolated to the complete repertoire. Also, in a scenario where the coexpression of a given OR gene with any other is random, only a very small number of OSNs would be expected to have a particular combination of two receptors and, therefore, it would be very challenging to detect by traditional methods. Furthermore, since *in situ* hybridisation requires specific probes, many times it is not possible to distinguish between closely related receptors. RNAseq of single cells allows, for the first time, to investigate the complete receptor gene repertoire in each cell, and assess how many different ORs are expressed.

First, I analysed the expression of all OR and TAAR genes and pseudogenes, and identified 476 receptor genes with at least one fragment mapped in at least one single OSN; all of these were OR and no expression of TAAR genes was detected. However, the great majority of these OR genes (86%) had less than one third of their gene

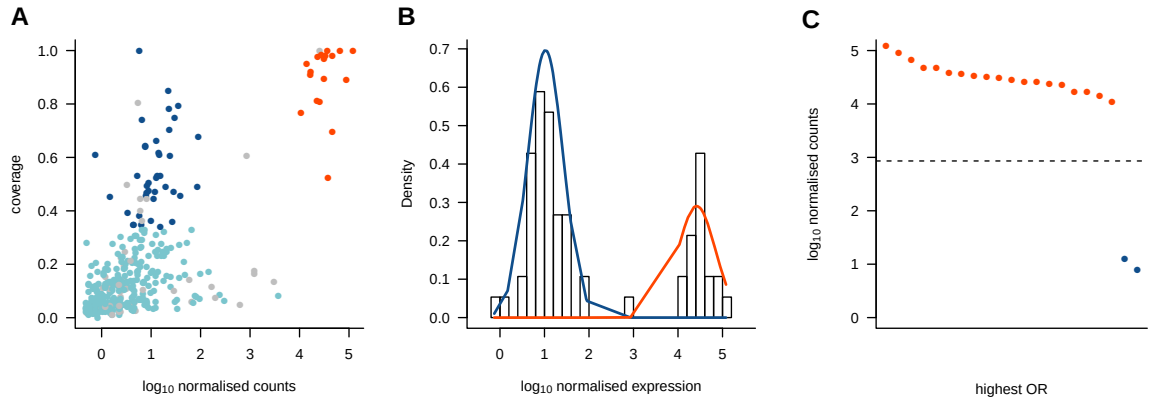

Figure 3.12 – OR expression in single OSNs. A) Scatter plot of the normalised expression of each OR gene in each single OSN against the proportion of the gene covered by at least one read. Pseudogenes are in grey. The majority of the OR genes are expressed at very low levels and have low coverage (light blue). The other 15% have coverage $> 1/3$ and **B)** are bimodally distributed into high (red) and low (dark blue) expressed OR genes. Two normal-like distributions were fit to the data to separate the genes based on their expression levels. **C)** Normalised counts for the highest OR expressed in each single OSN. In 19 of the 21 cells an OR is expressed at very high levels (red), while two OSNs show very low OR expression (dark blue). The horizontal dotted line indicates the expression value where the two normal distributions from (B) intersect, and represents the threshold to separate the high- from the low-expressers.

length covered by sequencing fragments, which suggests non-specific transcription and/or mismapped reads (Figure 3.12A). The remaining 65 OR genes segregated into two distinct distributions, depending on their expression level (Figure 3.12B). A total of 45 OR genes were expressed at low levels, with mean expression of only 15.9±2.7 (SEM) normalised counts, while the other 20 were expressed on average at 36,162.5±6,238.7 (SEM) normalised counts. The latter had sequencing fragments mapped along the majority of the full transcript, with median coverage of 0.93.

Next, I looked at the OR genes expressed in each single OSN. The intersection between the two distributions in Figure 3.12B (855.98 normalised counts) can be used to define whether an OR is expressed at low or high levels. In 17 of the 21 single OSNs there was a single OR gene, annotated as functional, expressed at high levels. Two more cells expressed an annotated OR pseudogene at similarly high levels*: Olfr1191-ps1* and *Olfr1224-ps1*. Closer inspection of these two genes, revealed that they encode full-length ORFs of 318 and 311 amino acids each, and these align to other mouse OR genes with high identity; *Olfr1224-ps1* is annotated as protein-coding in Ensembl. Thus, both of these genes are likely to encode a functional receptor. Together, 19 of the 21 single OSNs express a single putatively functional OR gene at great abundance (Figure 3.12C). Indeed, the OR genes rank on average as the 6^{th} most abundantly expressed gene in the transcriptome of these OSNs; only *Stoml3*, *Gnb1*, *Malat1* and *Calm1* are consistently expressed at higher levels. All the expressed OR genes are Class II except for *Olfr556*.

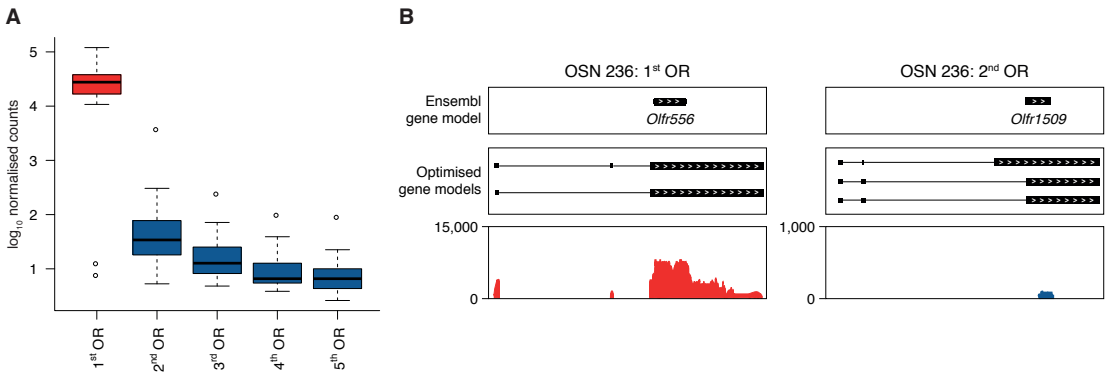

Figure 3.13 – Monogenic expression of OR genes. **A)** Boxplots for the five most abundant putatively functional OR receptor genes expressed in each single OSN, in decreasing order. Most OSNs express one OR gene at very high levels (red). The next most abundant OR gene is, on average, a thousand times lower. **B)** Representative example of the first and second most abundant OR genes expressed in a particular OSN. On the top panel is the gene model present in Ensembl, with the reconstructed models from the RNAseq data below. Each black box represents an exon and the lines joining them are introns; the arrow heads indicate the strand of the gene. At the bottom is a coverage plot of the sequencing reads. For the most highly expressed receptor (red) a large number of sequencing reads cover the full gene model, but for the second most abundant OR gene (blue) there is only a small number of reads that do not span the whole gene. Note the difference in scales for each plot.

In addition to the single abundant functional OR, the 19 single OSNs had between 11 and 28 other OR genes with evidence of expression, but all had extremely low normalised counts. After excluding annotated pseudogenes, the most highly expressed OR gene is on average over 1,000 times more abundant than the next highest OR gene expressed (Figure 3.13A-B). In two cases, a pair of single cells expressed the same OR gene; OSN 171 and OSN 177 each express *Olfr728*, and OSN 22 and OSN 263 each express *Olfr55*. The OR genes expressed at low levels in these two pairs of cells are not shared to a greater degree as do any other two cells, thus suggesting that these lowly expressed genes are not coordinated with the expression of the abundant OR gene, nor are they the product of mismapping.

To assess whether these low levels of OR expression could be biologically meaningful or whether they are more likely to represent leaky expression, I analysed publicly available single cell RNAseq data from 96 mouse T-helper lymphocytes[319] and 288 single mouse embryonic stem cells[320], which were captured and processed in a similar manner as the single OSNs, in the same facility. In these datasets, some cells expressed up to 101 OR genes, but all at very low levels (Figure 3.14). Therefore, this suggests that the low expression of a fraction of the OR repertoire is mainly the result of non-specific transcription and has no biological significance. The contrast with the highly expressed receptor, along with the recapitulation of this pattern in non-olfactory cells, strongly argue in favour of a monogenic expression pattern of OR genes in OSNs.

As discussed in Chapter 2, many OR genes have several transcripts that differ in their

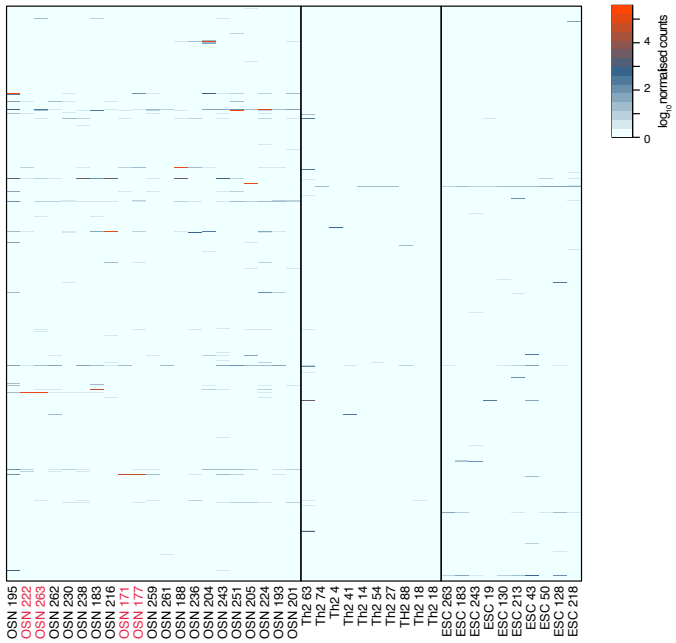

Figure 3.14 – OR expression in several single-cell datasets. Heatmap of the normalised expression levels for the complete OR repertoire; each row represents a different receptor gene. All 21 single cells are presented alongside 10 representative T-helper lymphocytes (Th2) and 10 embryonic stem cells (ESC). In the non-olfactory cells, there is a variable number of OR genes with low normalised counts, similar to what is observed for OSNs. The two pairs of single OSNs that express the same abundant OR gene are highlighted in red.

UTR structure and a few also have potentially different protein coding isoforms. None of the OR genes expressed in the single OSNs include genes with alternative protein isoforms, but most of them have several different transcripts. For these, the sequencing reads support expression of several distinct transcripts with alternative UTRs. This demonstrates that within a single OSN, the different isoforms of the OR gene expressed abundantly are present.

3.3.3 Monoallelic expression of OR genes.

The OMP-GFP mouse line is in a mixed genetic background of 129P2×C56BL/6 strains [309]. This allows to discriminate which allele is expressed, for all those OR genes with SNPs between the two strains. Therefore, to assess whether the abundant OR gene is expressed in a monoallelic fashion, I mined the mouse genomes project (MGP) data[321] to obtain the variable nucleotide positions within the highly expressed ORs. Of the 19 abundant OR genes, nine had at least one SNP within exons. As an example, *Olfr55* is expressed in both OSN 222 and OSN 263, and has 15 SNPs across its transcript. For OSN 222, there are 54,677 reads that map across these, and 54,563(99.79%) support the

nucleotides found in the C57BL/6 genome. Similarly, there are 52,503 reads spanning variable positions in OSN 263, 52,456 (99.93%) of which have the variant of the 129P2 allele (Figure 3.15A). Since both these cells originated from the same mouse, these data directly demonstrate that *Olfr55* is expressed monoallelically. For the other seven genes with SNPs, five expressed exclusively the C57BL/6 allele and two expressed the 129P2 allele; in all these cases, at least 99.57% of the reads supported expression of one of the alleles (Figure 3.15B). Thus, I have directly demonstrated that OR expression in single OSNs, not only is monogenic, but it conforms to an extremely tight monoallelic expression pattern.

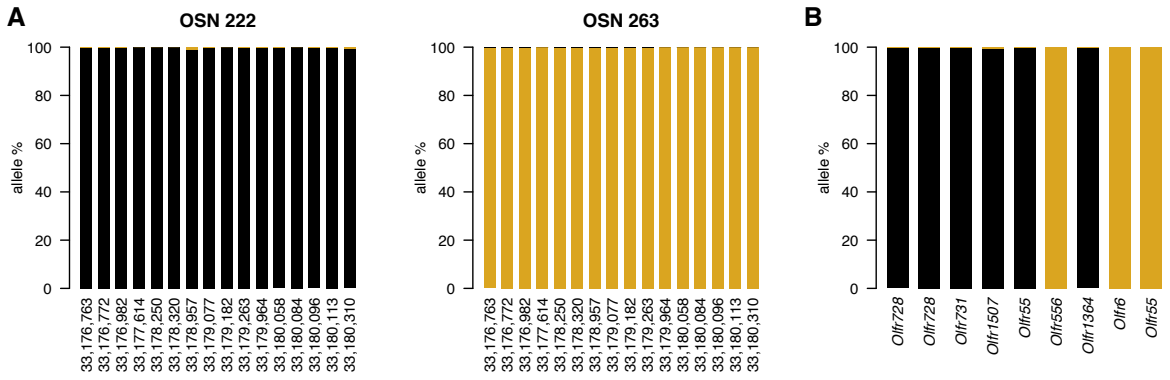

Figure 3.15 – OR expression is monoallelic. A) Stacked barplots of the proportion of sequencing reads supporting the C57BL/6 (black) or 129P2 (golden) alleles for the *Olfr55* gene in OSN 222 and OSN 263. Each bar represents a different SNP. **B)** Same but for the cummulative data for all SNPs in each of the other OR genes expressed at high levels that have SNPs. In all cases, a single allele is supported by over 99.5% of the data, indicating very tight monoallelic expression.

3.3.4 Identification of a novel type of OSN.

Two of the sequenced single OSNs (OSN 259 and OSN 261) did not express any annotated OR gene at high levels (Figure 3.12C); they have some expression of ORs but all are very low (the highest are at 12.66 and 7.7 normalised counts respectively) and rank at similar levels to what is found in the T-helper lymphocyte and ES cell transcriptomes. I could not detect expression of other chemoreceptor genes either, such as vomeronasal or taste receptors. From Figure 3.9 it is evident that neither of these cells express *Adcy3* or *Cnga4*, both essential components of the canonical signal transduction pathway, but they express many other markers of mature OSNs such as *Omp*, *Gnal* and *Cnga2*. In contrast, they express *Trpc2*, a gene characteristic of VSNs. During the course of this PhD, two subpopulations of OSNs in the MOE were shown to express *Trpc2*[131].

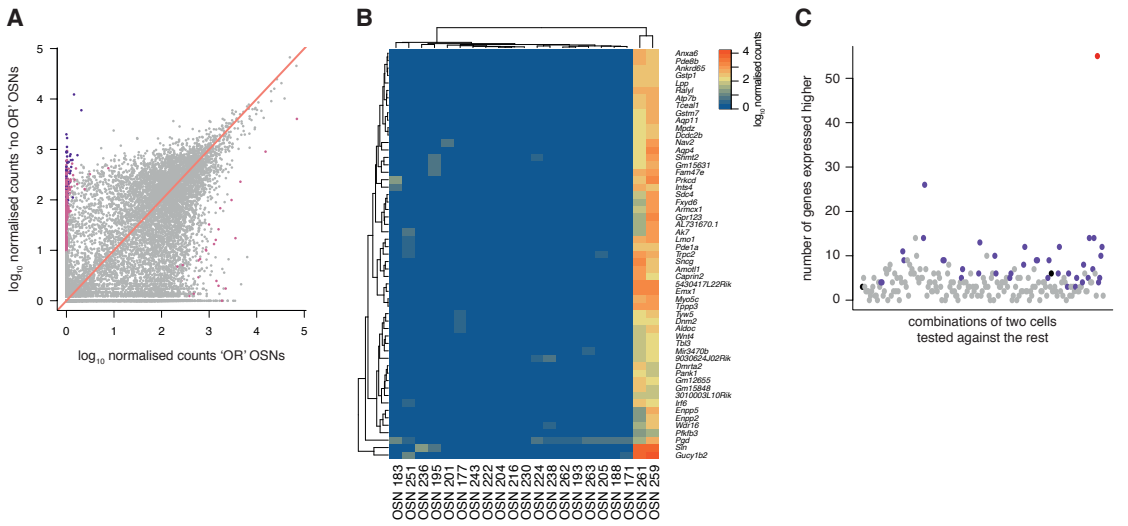

Figure 3.16 – Characteristic expression profile of a novel type of OSN. A) Scatter plot of mean normalised expression values for the whole transcriptome in the 19 OR-expressing cells (x-axis) versus the two no-OR cells (y-axis). The red line is the 1:1 diagonal. Genes significantly DE (FDR < 5%) are in pink, and those consistently expressed in both no-OR cells are in purple. **B)** Heatmap of the 55 DE genes that are consistently expressed higher in the no-OR cells compared to the other 19 cells. **C)** Number of genes that are consistently expressed higher when two cells are compared to the other 19. Each dot represents a specific combination of two samples. In purple are all those combinations that include one of the no-OR cells, and in red the combination of both no-OR cells.

In order to characterise these cells better, I compared their transcriptomes to those of the other 19 –OR-expressing– cells. Differential expression analysis revealed 494 significantly DE genes (Figure 3.16A) but only 55 that were consistently expressed in both cells (Figure 3.16B). In order to determine if this number of shared DE genes is statistically meaningful, I performed the same analysis for all 210 possible combinations of two among the 21 single OSNs. This pair of 'no OR' cells have over twice as many DE genes as any other pair of cells (Figure 3.16C) which suggests that they are indeed different from the rest and posses a set of genes specifically expressed in this novel type of OSNs.

The most abundant DE gene is *Gucy1b2*, a soluble guanylyl cyclase, followed by sarcolipin (*Sln*) and *Emx1*, a transcription factor involved in neuronal fate specification. *Trpc2* is ranked as the 39^{th} most abundant gene in the transcriptomes of these cells. In collaboration, Luis Saraiva and Masayo Omura validated the expression of some of the DE genes by *in situ* hybridisation, using *Trpc2* as a marker of this type of OSN. They found that *Gucy1b2* and *Sln* indeed are expressed in the MOE, sparsely distributed within the OSN and sustentacular cell layers (Fig 3.17A-B). By two-colour *in situ* they could confirm that the *Gucy1b2*$^+$ cells define a subset of the *Trpc2*$^+$ cells in the MOE (Fig 3.17C), consistent with the single-cell RNAseq data and recent reports[131, 322]. Additionally, *Gucy1b2* was found to be coexpressed with *Sln* and *Sncg* (a less abundant

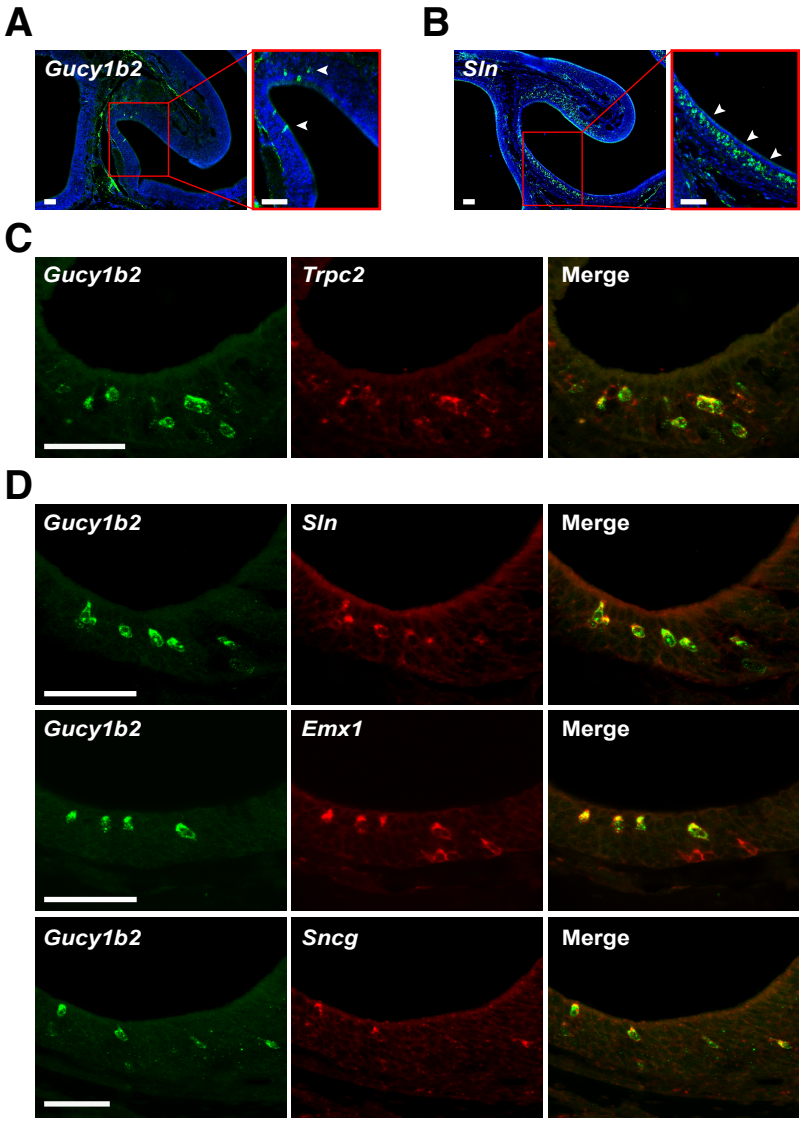

Figure 3.17 – Validation of the DE genes in no-OR cells. A-B) Cryosections of adult mouse MOE hybridised with cRNA probes for the two most highly expressed DE genes *Gucy1b2* (A) and *Sln* (B). The hybridisation signals are sparsely distributed within the MOE. **C)** Two-color *in situ* hybridisation of the top ranked marker (*Gucy1b2*) with *Trpc2*. Some but not all of the *Trpc2*$^+$ cells in the MOE also coexpress *Gucy1b2*. **D)** Two-color *in situ* hybridisation of the top ranked marker (*Gucy1b2*) with other differentially expressed genes, *Sln*, *Emx1* and *Sncg*. Arrowheads point to labelled cells. Scale bars, 50 μm. Image kindly provided by Luis Saraiva and Masayo Omura.

DE gene, ranked 7^{th}) and to partially overlap with cells expressing *Emx1* (Fig 3.17D). Thus, these two cells appear to be examples of the recently discovered type B *Trpc2*$^+$ cells in the MOE, which are now characterised by at least a handful of other genes, and clearly represent a distinct subtype of OSN.

In all, RNAseq has demonstrated to be a very powerful technology to definitely answer questions that are fundamental for the understanding of the olfactory system. The

data presented here strongly supports the monogenic character of OR expression, but a larger sample size will be necessary to extend this conclusion to more OSN types; the inherent challenges of working with single OSNs, however, will make this a difficult task. Interestingly, the monoallelic character of OR expression is extremely tight, with nearly the entirety of the sequencing data supporting expression of a single allele. Therefore, it seems like the expression of one abundant OR allele is very tightly controlled, with only a small degree of leaky transcription from a small subset of OR genes.