Chapter 2

The transcriptome of the mouse olfactory system.

Some of the material presented in this chapter has been previously published in reference [295]. I confirm the work is my own unless otherwise stated and I have ownership of the copyright for its reproduction.

Much is known about the structure and composition of the MOE and VNO, and of the essential components expressed by their sensory neurones to achieve olfactory detection and signalling. Much less understood is the transcriptional profile of these tissues and their individual components, how specialised they are and whether transcriptional differences can account for observed disparities in olfactory-mediated behaviours. On the study of the olfactory system, the sensory neurones are, undoubtedly, the most interesting component, given their role in signal detection and information processing. The regulation of the characteristic expression pattern of the receptors –monogenic and monoallelic– is still incompletely understood. Not surprisingly, a lot of work has been carried out to tackle the fundamental questions regarding receptor expression and regulation.

The sequencing of the mouse genome allowed the use of computational methods to search for the the complete repertoire of OR and VR genes, based on homology searches with the few dozen experimentally obtained sequences from rat and mouse sensory neurones[58, 65]. Recent analyses of the mouse genome have identified 1,130 OR genes with an intact open reading frame (ORF) and an additional 236 pseudogenes[296].

In the case of VR genes, there are 392 V1R genes (239 intact)[136] and 279 V2R genes (158 intact)[65]. If all these receptor genes are indeed true ORs and VRs, they should be expressed in the MOE and VNO respectively. Expression of several OR genes has been detected in other tissues, and some have been shown to play a role in functions other than olfaction, such as sperm chemotaxis[297, 298]. While some can have both olfactory and non-olfactory functions, others might be specifically expressed outside the olfactory system and should not be considered part of the receptor repertoire.

The study of the expression of receptor genes has been a daunting task for several reasons. First, there is a very large number of genes to analyse; this makes approaches such as *in situ* hybridisation slow and labour intensive. Secondly, both ORs and VRs tend to have close paralogous sequences; in some cases, two receptors can be nearly identical which makes their discrimination challenging. And third, each receptor is expressed in only a small subset of the cells from the whole population found in the tissue, which leads to very low expression levels when analysing total tissue RNA samples.

The expression of a few hundred receptor genes has been confirmed by *in situ* hybridisation assays. In an effort to produce a high-throughput approach, Firestein and colleagues developed custom microarrays containing probe sets for the majority of the annotated OR and VR sequences available at the time. For the ORs, the Celera mouse genome was used, one of the earliest versions available; since then, the OR repertoire has changed considerably and many receptors have been identified as artefacts of assembly errors. In the case of the VR genes, a more recent assembly of much better quality was used. For both OR and VR genes, probes were designed based on computationally predicted 3' untranslated regions (UTRs) for two reasons: the array is biased to detect this portion of the transcripts, and UTR sequences tend to be more variable compared to the CDS. However, no evidence of the accuracy of these predictions was provided[299, 300]. Despite all these drawbacks, the authors reported that 817 (62.55%) of the ORs represented in their array were enriched in the MOE compared to the VNO[299]; and 168 (54.5%) and 98 (32.1%) V1R and V2R genes respectively were expressed in the VNO[300]. These studies represent the highest number of receptors with validated expression in their cognate tissue.

Microarrays have the advantage of providing space for thousands of probes; this makes it possible to interrogate all the genes annotated in the genome at the same time in a given sample. The high-throughput character of microarrays made them a very popular technology and have been widely used in different species, tissues and experimental designs. Yet, they have several limitations. First, microarrays suffer from high

background levels of hybridisation, that is, signal is recorded irrespective of the presence of a transcript; this makes them less sensitive to accurate estimation of gene expression levels, especially for lowly expressed genes where it becomes difficult to differentiate true signal from background. Secondly, some probes are more efficient than others in their hybridisation efficiency. Third, probes for genes with close paralogs can cross-hybridise, also making their accurate detection problematic. And fourth, the information obtained with the array is limited to known genes and relies on the correct annotation of the genome; thus, any novel genes cannot be detected[301].

A different technology, the NanoString nCounter platform, allows the simultaneous quantification of expression levels for hundreds of genes. Even though it does not provide genome-wide coverage, the sensitivity of the assay is much greater, especially for lowly expressed genes that can not be detected accurately by microarrays. This technology is based on the design of two probes for each gene: a *capture* probe, that contains 35 to 50 nucleotides complementary to the mRNA of interest, coupled to an affinity tag such as biotin; and a *reporter* probe, also consisting of 35 to 50 nucleotides of the target mRNA sequence, plus a tag that contains a specific sequence of different fluorophore-labelled RNAs that make up a unique code for identification. The assay consists on hybridising the RNA sample of interest with the capture and reporter probes; the hybridised probes are then fixed to a solid surface coated with streptavidin and imaged to identify the code of each molecule. Finally, counts for each type of targeted mRNA are obtained[302]. Using this technology, Khan et al. were able to design specific probes for 592 OR genes, based on their CDS; for the rest of the repertoire, probes were not specific enough to differentiate closely related genes. From the tested ORs, 577 (97.5%) were detected reliably in wild-type mice. They further demonstrated the sensitivity and specificity of the system with two experiments: 1) RNA was obtained from OSNs that expressed a particular OR gene and, as expected, a high signal was obtained for that OR only; and 2) RNA was prepared from a mouse strain that has a deletion in a cluster of about a hundred OR genes in chromosome 9; for the deleted genes, expression levels were no different from background while the rest of the repertoire was correctly detected[186]. Despite the great sensitivity and specificity of this system, it is restricted to the study of those receptor genes that are divergent enough to be unambiguously detected by short probes.

Nowadays microarrays are being replaced by high-throughput RNA sequencing (RNA-seq), given the drop in the cost of sequencing and the advances in the development of both robust protocols for library preparation and bioinformatic analysis tools for pro-

cessing the data. RNAseq has several advantages for transcriptome profiling. It has virtually no noise since all sequencing reads are obtained from the RNA present in the sample. It is quantitative: the number of sequenced reads per gene is proportional to its expression level, and there is no saturation of the signal; increasing expression of a gene results in increasing numbers of sequencing reads. It does not depend on *a priori* knowledge of the genome sequence or annotated genes, allowing for the discovery of new genes. Also, this results in a comprehensive capture of the different splice isoforms for a given gene, some of which may not be known. The major limitation of RNAseq is that it is based on short reads; these can sometimes map to several regions of the genome, which makes them ambiguous since their true place of origin can not be determined[303]; the increase in length of sequencing reads (from 32 bp to over a hundred bp) has ameliorated this problem, in conjunction with the development of paired-end sequencing that also increases the amount of sequence obtained from each molecule. Nonetheless, some regions of the genome are still affected by a high proportion of ambiguous reads –also called *multireads*– especially those containing genes with close paralogs.

At the beginning of my PhD, I had access to RNAseq data from both the MOE and VNO of C57BL/6J adult mice. Given the clear advantages of RNAseq over other transcriptomic technologies, I conducted a series of validation experiments to assess the suitability of this method to accurately profile the expression of the olfactory system and, mainly, of the receptor genes.

2.1 Transcriptome profiling by RNAseq.

In order to study the complete profile of genes expressed in the MOE and VNO of wild-type adult B6 mice, these tissues were dissected from both male and female mice and RNA was extracted to construct libraries for RNAseq. All samples were sequenced at high depth to ensure capture of lowly expressed genes, such as the receptors. The VNO samples, composed of the sensory neuroepithelium, progenitor and non-neuronal supporting cells, underlying glandular and cavernous tissue and a blood vessel with blood[1], yielded a mean of 37.1 $\pm$ 3.6 million (SD) paired-end fragments per sample. For the MOE, dissections included the underlying lamina propria, glandular tissue, olfactory nerves and blood vessels with blood, in addition to the olfactory epithelium[304]; for clarity, I will refer to these as the whole olfactory mucosa (WOM). WOM samples yielded 46.4 $\pm$ 4.3 million (SD) paired-end fragments on average (Table B.1 in Appendix B). All sequencing fragments were mapped to the mouse reference genome with a splice-

aware algorithm; on average, 80.96±1.2% (SD) and 87.78±3.0% (SD) of the VNO and WOM data respectively mapped unambiguously (Table B.2 in Appendix B). To estimate expression levels, I counted the number of fragments that mapped to each gene model, as annotated in the Ensembl database. In order to be able to compare expression levels across different samples, I normalised these raw counts to account for the depth of sequencing. Detailed methods can be found in Appendix A.

I first assessed the variation in gene expression among the three biological replicate samples for each sex and tissue. All pairwise comparisons showed very high correlation levels (Spearman's rank correlation coefficient rho > 0.95, p-value $< 2.2e\text{-}16$), with only small sets of genes showing unusually variable values among replicates (Figure 2.1). I therefore averaged the normalised counts for each gene in each tissue.

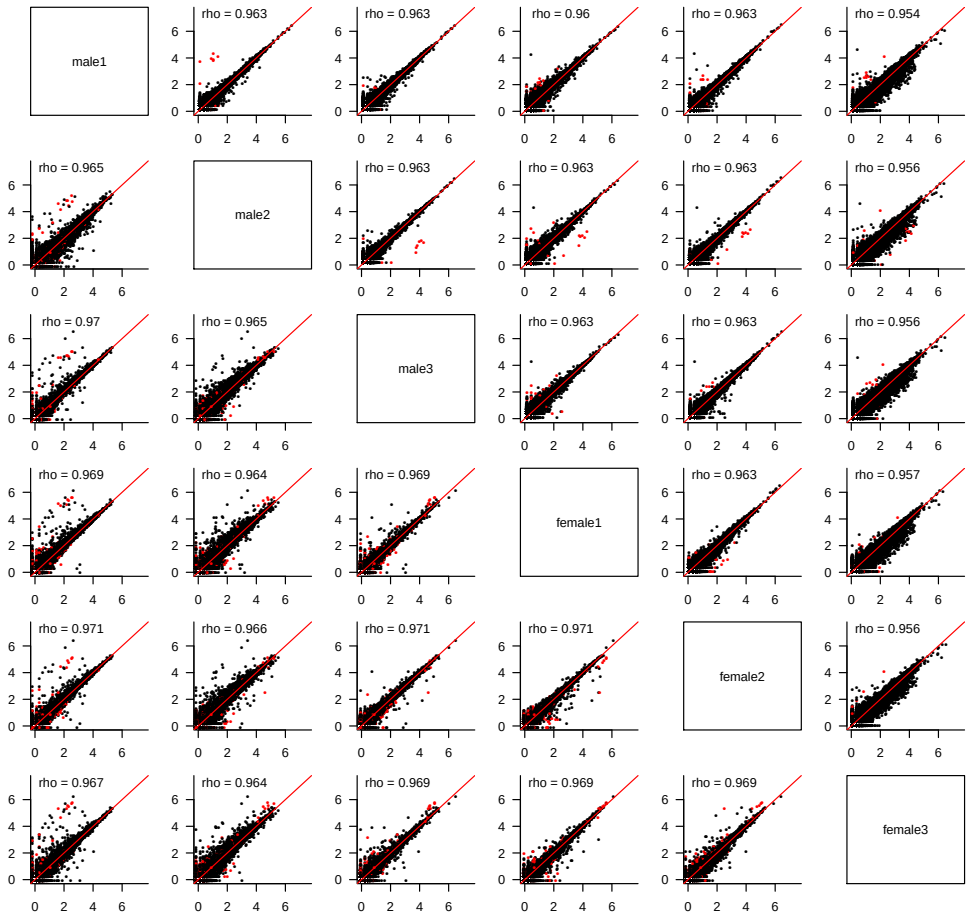

Figure 2.1 – Correlation between biological replicates. Scatter plots of the $\log_{10}$-transformed normalised counts for each pairwise comparison of biological replicates for the VNO (on top of the diagonal) and WOM (undeneath the diagonal) data. The 1:1 diagonal is in red. Samples are indicated in the diagonal. The Spearman's correlation value is on the top left corner. Genes with a coefficient of variation greater than the 90th percentile are highly variable between replicates, and are highlighted in red.

A total of 11,215 (29.01%) and 10,543 (27.28%) genes in the VNO and WOM respectively have no fragments mapped in any replicate, suggesting they are not expressed in that tissue. The expression of the remaining genes shows a bimodal distribution of low- and high-expressed genes characteristic of RNAseq datasets[305]. These can be decomposed into two normal-like overlapping distributions (Figure 2.2), and each gene can be assigned to either with a degree of confidence. Low-expressed genes typically do not have active chromatin marks, are enriched in non-functional mRNAs and, unlike the high-expressed genes, lack correlative protein expression data[305]. For the VNO and WOM respectively, 56.5% and 52.65% of the expressed genes are in the distribution of high-expressed genes with 50% probability or more.

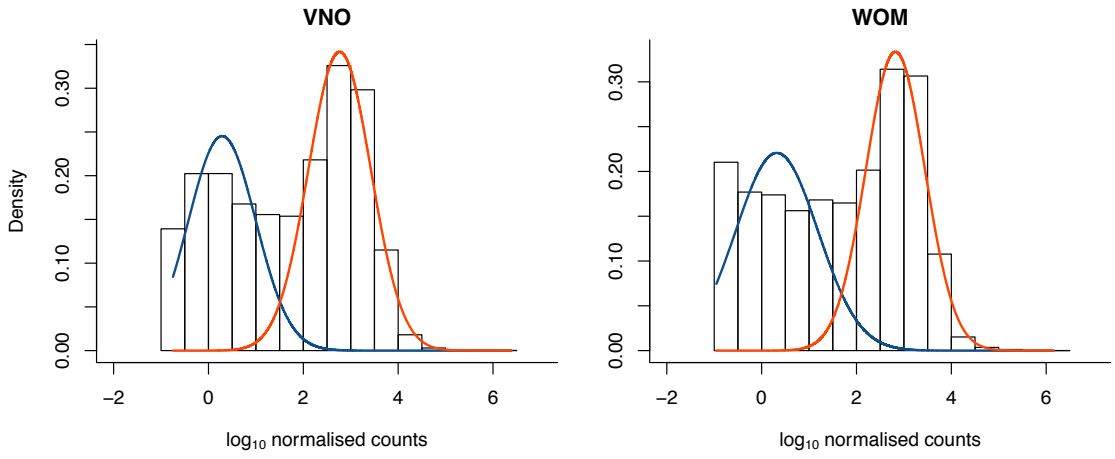

Figure 2.2 – The transcriptome shows a bimodal distribution of low- and high-expressed genes. Histograms showing the distribution of the mean $\log_{10}$-transformed normalised counts for all those genes expressed in at least one replicate, for the VNO and WOM. Superimposed are two normal-like distributions calculated from the data, that decompose the total bimodal distribution. The blue distribution contains all those genes expressed at low levels, while the red distribution represents highly expressed genes.

A comparison of the transcriptomes of both tissues revealed that 61.43% of the genes are differentially expressed (DE) with a false discovery rate (FDR) of less than 5%. To explore these further, I selected the genes with a fold change of four or higher and performed a functional terms enrichment analysis. As expected, those expressed higher in the VNO were enriched for VR genes, which contribute to the response to pheromones and detection of chemical stimulus involved in sensory perception. Additionally, the calcium signalling pathway, ionic and voltage-gated channel activity and regulation of vasoconstriction were significantly enriched as well. For the WOM, enriched genes were dominated by ORs and those involved in the olfactory transduction pathway, cAMP-mediated signalling and sensory perception. In addition, there was enrichment of ionic and ligand-gated channels and regulation of neurone, oligodendrocyte and glia differen-

tiation. In contrast, 'housekeeping' genes were expressed at similar levels in both tissues (Figure 2.3). All together, these data indicate that RNAseq is a highly reproducible technique to study the olfactory system, despite the heterogeneity in the composition of each tissue; and that the quantification of expression levels accurately reflects the principal functions of each tissue.

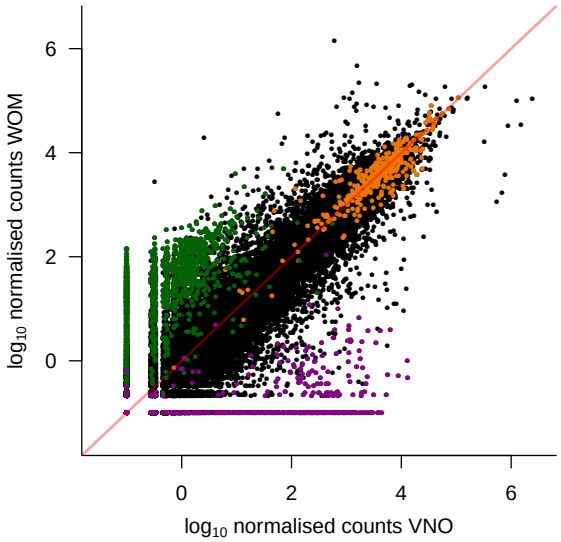

Figure 2.3 – The transcriptome of the VNO and the WOM. Scatter plot of the expression values for each gene in the VNO versus the WOM; 0.1 was added to the counts to be able to visualise those genes that are expressed in one tissue only. The VR genes (purple) are largely undetected in the WOM but clearly expressed in the VNO. The opposite is true for the OR genes (green). In contrast, housekeeping genes (orange) tend to be expressed at similar levels in both tissues. The red line represents the 1:1 diagonal.

2.1.1 Comparison to alternative methodologies.

As mentioned previously, the most popular technology for transcriptome profiling before the advent of high-throughput sequencing was expression microarrays. To assess whether RNAseq is a better technology to study the olfactory system, I used commercial Illumina microarrays to profile six more biological replicate VNO and WOM samples[1]. For both tissues, the overall expression values are correlated (rho = 0.66 for the VNO, 0.67 for the WOM, p-value < 2.2e-16). However, the RNAseq expression estimates are considerably different for genes that are either very lowly or very highly expressed. As can be observed in Figure 2.4A, the microarray intensity values flatten for the genes in both ends of the distribution, because of high background noise and saturation of the fluorescent signal; yet, in the RNAseq data, expression values span several orders of magnitude. Thus, the

[1]Microarray data are available in the ArrayExpress database under accession number E-MTAB-2163.

dynamic range of expression levels that can be detected is much broader for RNAseq compared to microarrays.

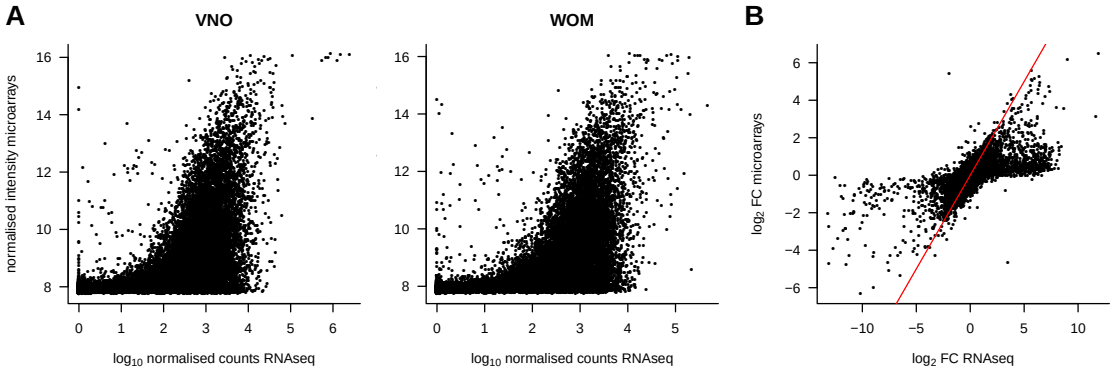

Figure 2.4 – Comparison of the RNAseq expression values versus the microarray intensity data. A) Scatter plot of the expression estimates for each gene from the RNAseq data compared to the microarrays. **B)** Comparison of the fold-change (FC) for each gene as assessed by both technologies. The red line indicates the 1:1 diagonal.

From the genes with expression signals above background in the microarray data, 11,499 can be directly mapped to an Ensembl gene id present in the RNAseq data. From these, 66.96% are reported as DE between the VNO and WOM samples (FDR $< 5\%$), a proportion similar to the results from the RNAseq data. 76% of the DE genes as assessed by the microarray are also DE from the sequencing data and the overall estimated fold-changes from both platforms are correlated (rho = 0.79, p-value $<$ 2.2e-16). However, a subset of the genes with large fold-changes identified as DE by RNAseq have small differences in the microarrays (Figure 2.4B). Previous studies have shown that the genes that are called as DE by only one technology validate well for the RNAseq calls but poorly for the microarrays[301], which suggests that these robust changes detected in the RNAseq data are likely to be true positives.

Quantitative real time PCR (qRT-PCR) is accepted as the gold standard for expression profiling, so next I compared both the RNAseq and the microarray expression estimates to a panel of qRT-PCR TaqMan gene expression assays[2]. I included data for genes with and without a known function in olfactory and vomeronasal signalling that cover the whole range of expression values observed (Figure 2.5). The correlation is considerably higher with the RNAseq data (rho = 0.87 for the VNO and 0.92 for the WOM, p-value $<$ 2.2e-16) than with the microarray estimates (rho = 0.70 for the VNO and 0.78 for the WOM, p-value $<$ 2.2e-16), indicating that RNAseq is better suited for transcriptome profiling in the olfactory system. As observed previously, these comparisons also

[2]TaqMan expression estimates were obtained and kindly provided by Maria Levitin.

suggest that microarrays are not sensitive enough to accurately detect genes expressed at low levels (Figure 2.5). Thus, the strong correlation between the qRT-PCR and the RNAseq data, strongly suggests that the expression estimates generated by sequencing are reproducible and specific, and provide a comprehensive characterisation of olfactory transcriptomes at a wide range of expression values.

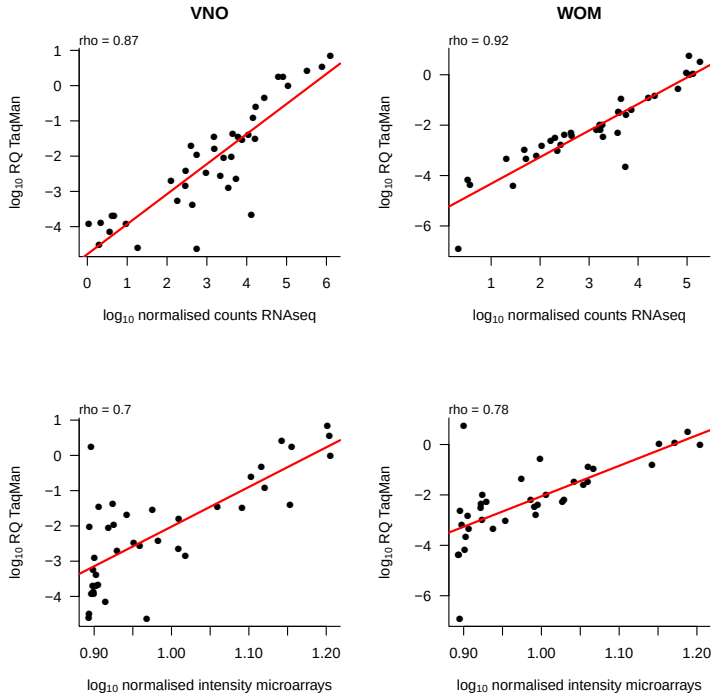

Figure 2.5 – Comparison with qRT-PCR TaqMan expression assays. The normalised RNAseq (top) or microarray (bottom) data was compared to expression estimates obtained by qRT-PCR TaqMan probes. The Spearman's correlation value is indicated in the top left corner. In red is the regression line. *B2m*, *Bpifa1*, *Calm1*, *Casp3*, *Clca1*, *Cnga2*, *Dmbt1*, *Dnase1*, *Eef1a1*, *Expi*, *Gnai2*, *Gnal*, *Gnao1*, *Gucy2e*, *H2-Aa*, *Ltf*, *Nelf*, *Olfr1213*, *Olfr123*, *Olfr124*, *Olfr1262*, *Olfr1347*, *Olfr1507*, *Olfr1509*, *Olfr222*, *Olfr691*, *Olfr692*, *Omp*, *Slc8a1*, *Stk32a*, *Trpc2*, *Vmn1r90*, *Vmn2r1*, *Vmn2r29*, *Vmn2r3* and *Xist* were assayed in both tissues and *Vmn2r121*, *Vmn2r28* and *Vmn2r6* in VNO only since amplification failed with WOM samples.

2.2 Expression of the receptor repertoire.

So far I have shown that the RNAseq data is reproducible and accurate at different expression levels. However, properly assaying the receptor repertoire has additional challenges. The monogenic expression of receptors in sensory neurones means each individual receptor is expressed in only a small subset of cells. Therefore the expression of any given receptor within the whole epithelium is low. The GRCm38 mouse assembly contains 1,250 annotated OR genes and 530 VR genes. To ensure that these represent the

complete repertoires, I took the cDNA sequences for the mouse VR genes as previously reported[65, 136], and aligned them to the genome with BLAST. I recovered 32 Ensembl genes that were not annotated as a VR gene, but that perfectly matched a VR cDNA sequence. These were included in subsequent analyses (Table B.3 in Appendix B). An additional 19 VR and 4 OR genes matched genes not annotated as receptors with high identity. These are annotated as novel genes and most likely represent additional members of the receptor repertoire. Proper analysis and annotation with new gene names is necessary for all these, which is beyond the scope of this dissertation; therefore, these were not considered in further analyses. It is worth noting that the above strategy was only intended to recover genes that are present in Ensembl but not labelled as VR or OR genes. Other methodologies such as hidden markov models would be more appropriate to discover new genes that possess the structure, domains and motifs that characterise these multigene families.

I first analysed the overall expression distribution for each class of receptors in their cognate tissue. In both cases, after normalising for gene length, the receptors in the repertoire do not have equal abundances, as may be expected if each receptor had equal probability of being chosen for expression by the OSNs. Instead we observe a large dynamic range of expression: a few receptors are expressed at high levels and the vast majority of the repertoire is expressed at relatively low levels. For the VR genes, the most highly expressed receptor, *Vmn2r89*, has a value of 4,363.4 normalised counts and only 22 other receptors have more than 1,000 normalised counts. In contrast, the median expression is 20.69 and 37.95 for V1R and V2R genes respectively (Figure 2.6). 436 VR genes (77.3%) have at least one fragment mapped uniquely. From the remaining 127, 70.1% are annotated pseudogenes, and 62 (48.8%) have multireads, which indicates that at least a fraction of these are expressed. 65 VR genes have no mapped fragments, either unique or multi-mapped, but these are all annotated as pseudogenes.

In the case of the OR genes the most abundant, *Olfr1507*, is expressed at 2,456.1 normalised counts and only 12 other receptor genes are above 500 normalised counts. The median expression is 26.88 normalised counts (Figure 2.7). Despite their relatively low abundance, 1,182 (94.56%) of all the annotated OR genes have at least one fragment mapped to their exonic region. Of the remaining 68 genes, 50 (73.5%) are annotated as pseudogenes and 17 have multireads that could indicate expression. Only 9 putatively functional OR genes have no evidence of expression in the WOM whatsoever (*Olfr115*, *Olfr141*, *Olfr504*, *Olfr564*, *Olfr574*, *Olfr834*, *Olfr1053*, *Olfr1061*, *Olfr1367*).

Importantly, the expression estimates for both OR and VR genes are consistent

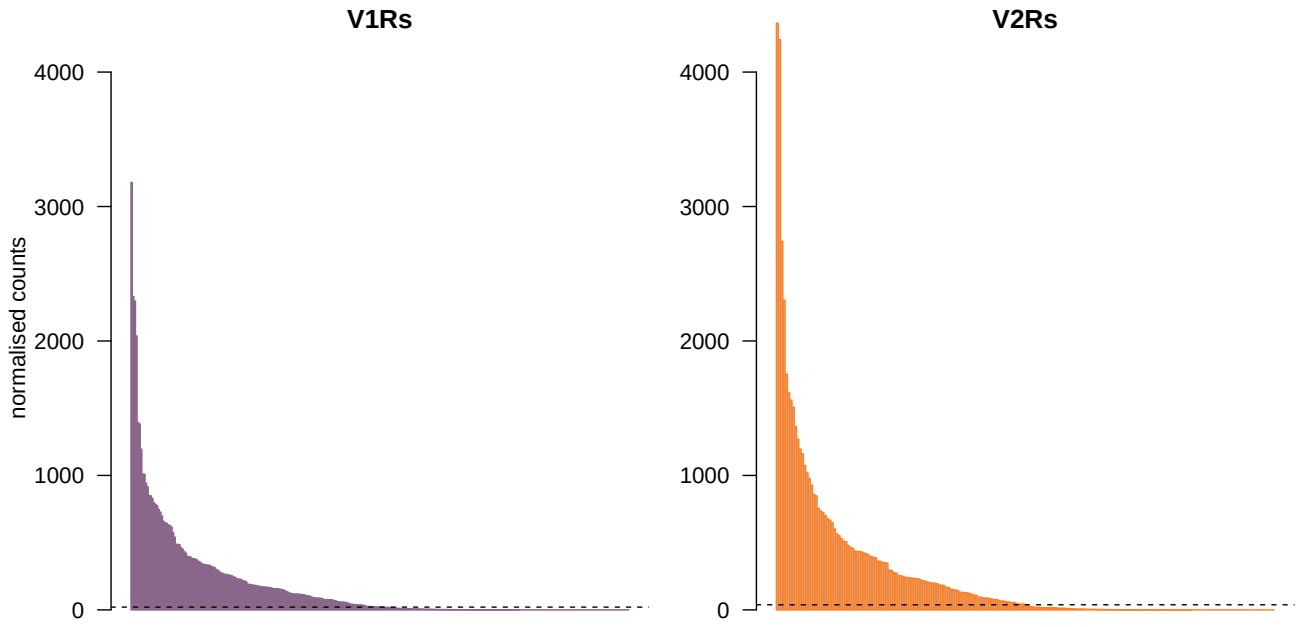

Figure 2.6 – **Expression of the VR repertoire.** Each bar represents the mean normalised expression of a V1R (purple) or V2R (orange) gene; counts have been normalised for gene length as well as depth of sequencing. The horizontal dotted line represents the median expression.

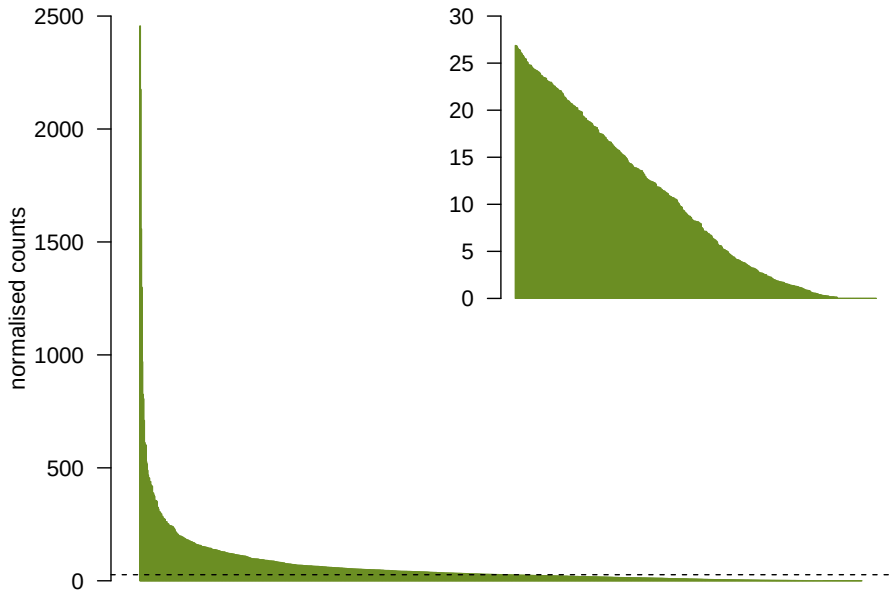

Figure 2.7 – **Expression of the OR repertoire.** Each bar represents the mean normalised expression of an OR gene; counts have been normalised for gene length as well as depth of sequencing. The horizontal dotted line represents the median expression. The inset contains all those receptors with expression values below the median.

between biological replicates. The median coefficient of variation (CV; standard deviation / mean) is 0.27 and 0.31 for the VR and OR genes respectively, suggesting that the uneven distribution observed is stereotypical. I could not identify any apparent pattern in the expression of either VR or OR genes based on cluster or genomic location. A difference in expression abundance is evident for genes and pseudogenes, with the former being expressed at much higher levels than the latter, for both VR and OR genes (p-value < 2.2e-16, Mann-Whitney test; Figure 2.8A-B). This is consistent with previous reports that observed the expression of pseudogenes was extinguished progressively after an OSN chose a functional receptor[207]. Interestingly, class I OR genes and V1R genes are also expressed at significantly lower levels than class II OR genes and V2R genes, respectively (p-value = 0.0077 for VRs and 0.0004 for ORs, Mann-Whitney test; Figure 2.8C-D).

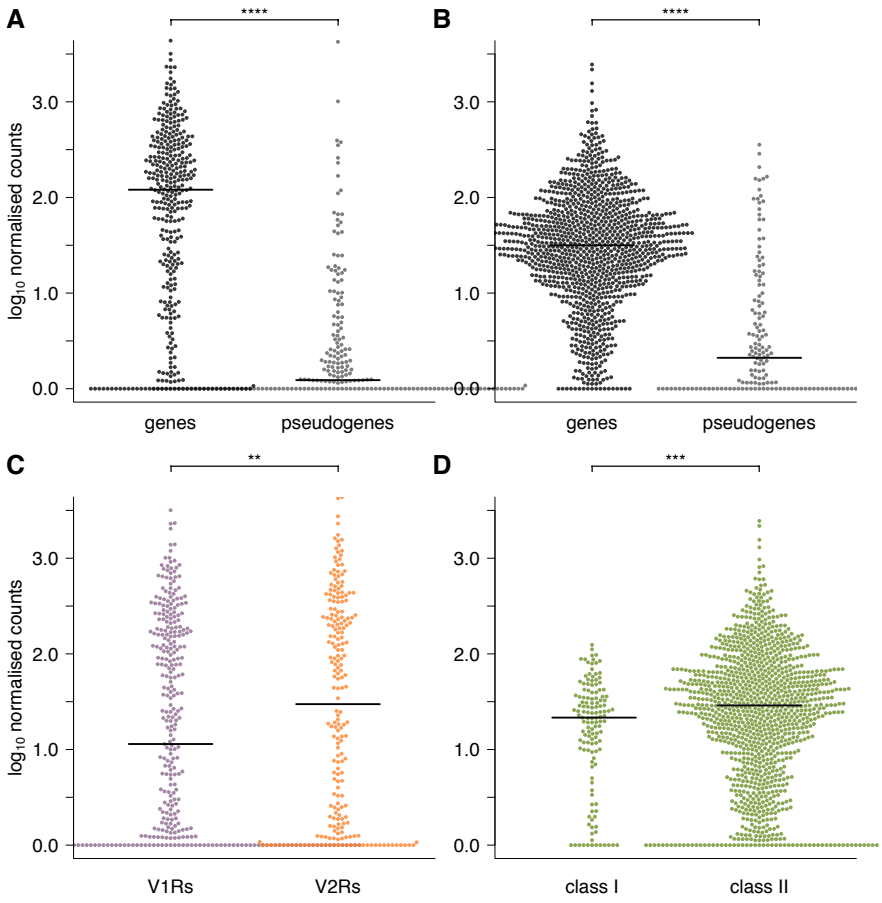

Figure 2.8 – Genes are expressed higher than pseudogenes. A-B) The normalised expression of each VR (A) and OR (B) gene is plotted, separated by whether they are annotated as intact genes or pseudogenes. Pseudogenes are expressed at much lower levels for both receptor types. **C-D)** Same but now separated by receptor class. V1R genes, as a whole, are expressed lower than V2R genes, and class II OR genes are expressed higher than class I genes. Mann-Whitney test; ** < 0.01, *** < 0.001 and **** < 2.2e-16.

Finally, I asked whether any VR genes are expressed in the WOM, or if OR genes are found in the VNO. One VR gene, *Vmn2r29*, is expressed in the WOM at a level that is higher than the median OR gene expression (32.1 normalised counts). This is consistent across all six replicates, suggesting there may be additional populations of sensory neurones in the MOE expressing receptors different to ORs. In the case of the VNO, 11 OR genes are expressed higher than the VR gene median, with *Olfr124* as the highest (496.7 normalised counts) followed by *Olfr692* and *Olfr1509* (262.2 and 126.1 normalised counts respectively). Both *Olfr124* and *Olfr692* consistently display higher expression values in the VNO than in the WOM.

2.2.1 Comparison to other methodologies.

In the commercial microarrays, there are probe sets for 178 VR genes and 1,051 OR genes, that can be mapped directly to a gene annotated in the Ensembl database; these represent 31.6% and 84.1% of the repertoires, respectively. The correlation between both datasets is only moderate for the OR receptors (rho = 0.54, p-value < 2.2e-16) and weak for the VR genes (rho = 0.30, p-value < 2.2e-16). As observed in Figure 2.9A, a proportion of the receptors accumulate in the lower detection threshold of the microarray intensity signal, close to the background level of the array; those genes show varied levels of expression in the RNAseq data, spanning several orders of magnitude. Therefore, the poor correlation can be attributed to the low sensitivity of the microarray to estimate expression of lowly expressed receptor genes.

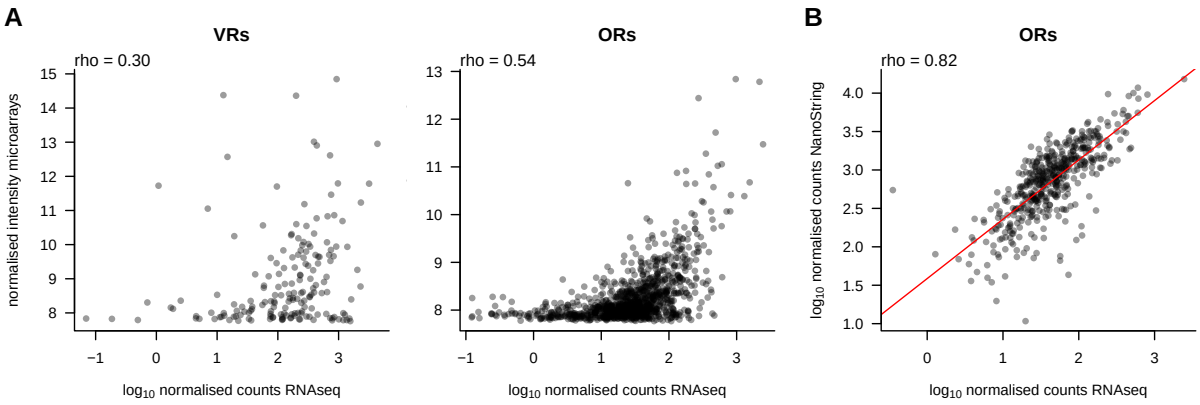

Figure 2.9 – Comparison of the receptor expression across different platforms. A) Scatter plots of the expression values for VR and OR genes in the RNAseq data versus the microarrays. Many receptors have intensity values around 8 in the arrays, the threshold between signal and noise; these genes are scattered along several orders of magnitude in the RNAseq data. **B)** Comparison of the expression values for OR genes in the RNAseq data versus the NanoString nCounter estimations. Values are highly correlated. In red is the regression line. For all, the Spearman's correlation coefficient is indicated in the top left corner.

Khan et al. profiled 531 OR genes in C57BL/6 animals, using NanoString nCounter [304]. When compared to the RNAseq expression levels, the NanoString normalised counts are highly correlated (rho = 0.82, p-value < 2.2e-16; Figure 2.9B). The agreement of these two technologies, which are based on very different detection principles, provides further support for the accuracy in the quantification of expression of receptor genes by RNAseq.

2.2.2 Sensitivity of RNAseq to detect lowly expressed receptor genes.

The expression of the majority of the receptor genes is very low, but they still display a dynamic range of expression levels (inset in Figure 2.7) that are consistent between biological replicates (0.81 < rho < 0.92, p-value < 2.2e-16). At such low expression values, the variance of the data increases considerably, but the high correlation coefficients imply that the relationships between the different receptors are preserved. To further test whether this range of expression values is meaningful, I analysed RNAseq data obtained from the WOM of mice with a homozygous deletion of the second cluster of OR genes in chromosome 9 (S. Xie, P. Feinstein, and P. Mombaerts, unpublished data; [306])[3]. Of the 94 deleted OR genes within the cluster, 83 (88.3%) had no counts in any of three biological replicates; importantly, these same genes show expression estimates up to 301 normalised counts in control animals (Figure 2.10). The 11 remaining genes have just one or two fragments mapped in only one of the replicates, resulting in normalised counts of less than 0.4.

This suggests that RNAseq suffers from virtually no noise and that very low expression levels reflect real expression in the samples profiled. Therefore, I propose that the RNAseq methodology presented here is able to detect the expression of most of the putative functional receptors in complex samples of whole tissue from the olfactory system, despite their low expression levels. This provides, for the first time, almost complete evidence of expression for the computationally predicted receptor genes in their cognate tissue, which supports their role in olfactory signalling.

2.2.3 The multiread problem.

The major limitation of RNAseq is that it is based on short fragments; these are not always unique in the genome, which means that by reading their sequence, one cannot

[3]RNA from these animals was kindly obtained and provided by Mona Khan.

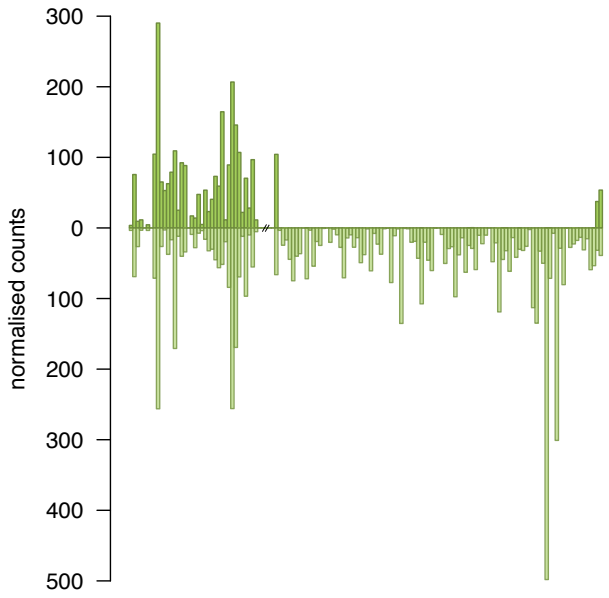

Figure 2.10 – Expression of a deleted OR cluster. The mean normalised expression for each receptor in chromosome 9 is plotted for a mouse with a homozygous deletion of an OR cluster (dark green); the break in the x-axis separates the two clusters present in this chromosome. While expression is detected for the first cluster, the OR genes from the deleted cluster are not expressed, except for the ones at the borders, which were not included in the deletion. The corresponding expression in wild-type animals is plotted as a mirror image (light green).

know their true place of origin. These multireads occur much more often in genes with close paralogs and may, therefore, represent a big problem for accurately estimating the expression of receptor genes. To assess this, I first calculated the *uniqueness* of each receptor gene with different lengths of reads; that is, I counted how many of all the possible n-mers obtained from a given receptor gene can be mapped unambiguously to the genome, with $n = (32, 76, 100)$ nucleotides (nts), which are common read lengths.

In general, OR genes are much more unique than VR genes (p-value $<$ 2.2e-16, Mann Whitney test). Sequence fragments of 32 nts are very often multimapped; at this fragment length, 43.7% and 15.8% of the VR and OR genes respectively have less than 50% uniqueness. However, a steep increase in the uniqueness values occurs when the sequence length is increased to 76 nts and then a modest gain when fragment length reaches a 100 nts (Figure 2.11). These values, however, are a worst-case scenario, since paired-end reads will improve the uniqueness of many genes. Therefore, RNAseq based on reads of 76 or 100 nts should be able to differentiate between most receptors. Nonetheless, even at 100 nts, 53 (9.4%) VR genes have uniqueness of 0, and 105 (18.6%) have uniqueness $<$ 25%; these are virtually inaccessible for any current profiling technique.

Despite the ability to detect expression of most receptor genes, those with low unique-

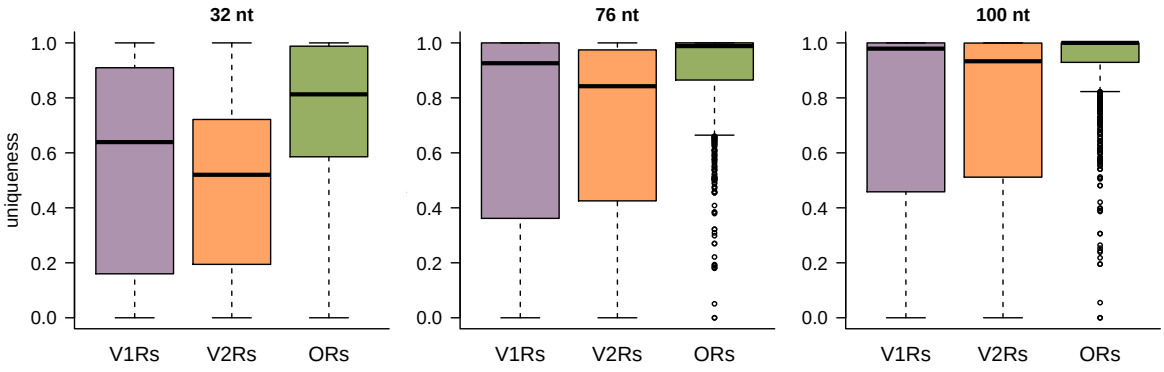

Figure 2.11 – Uniqueness of the receptor sequences. Boxplots indicating the uniqueness values for V1R, V2R and OR genes, for read length of 32, 76 and 100 nucleotides long.

ness will be consistently underestimated. I therefore analysed how much information is lost in multireads and whether this affects some classes of receptors more than others. For each receptor, I counted the number of different sequencing fragments that map to it; therefore, each multiread was counted at least twice, once for each receptor it maps to. On average, 37% and 94.1% of the multi-mapped fragments had only two alignments, for VR and OR genes respectively, with the rest mapping to three or more different locations. These counts were then normalised to account for depth of sequencing and gene length, and compared to the corresponding unique counts (Figure 2.12).

Consistent with the uniqueness calculations, there are many more multireads assigned to VR than to OR genes, and there are some receptor families that have many more multireads than others. For example, there is a cluster of 84 V1Rs in chromosome 7 that have very low unique counts (median = 0.1), and a low but consistent number of multireads (median = 50.01); accordingly, the uniqueness of these genes is very low (median = 0; mean = 0.12). For the VR repertoire, 44.85% of the receptors have more multireads than unique counts; in contrast, for the OR genes, this is the case for only 9.9%. Therefore, the expression estimates of some classes of VR genes are not accurately accounted by RNAseq, but the majority of the OR repertoire is not significantly affected by this problem.

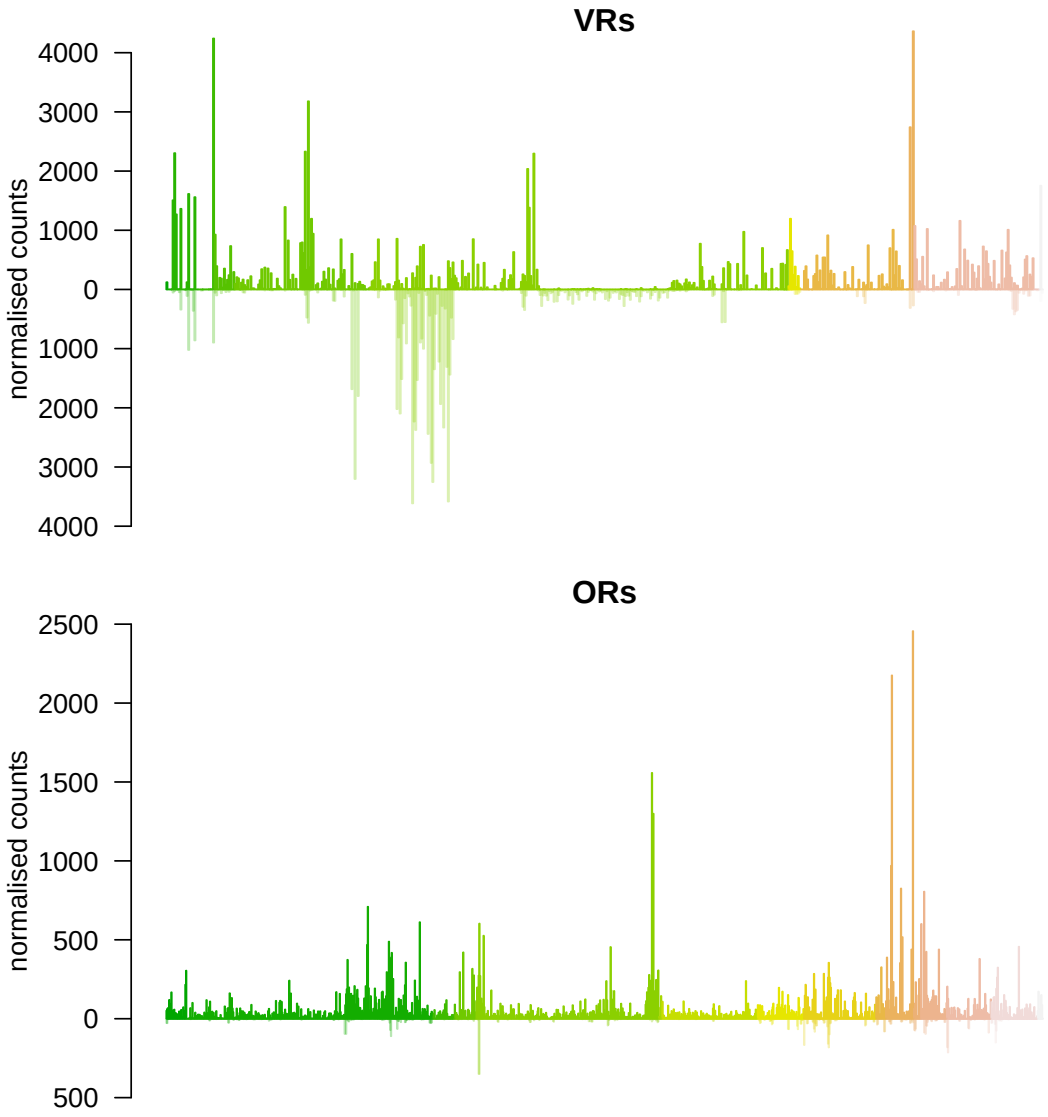

Figure 2.12 – **Multireads mapped to VR and OR genes.** Barplot of the normalised counts for each VR (top) and OR (bottom) gene; receptors are ordered according to their chromosomal location and each chromosome is a different colour. As mirror images are the corresponding normalised counts of multireads for each receptor.

2.2.4 Complete annotation of the gene models.

An advantage of RNAseq over other expression profiling techniques is that it is not restricted by a catalog of known transcripts. Using the sequencing data directly, it is possible to reconstruct the most parsimonious set of transcripts that would generate the observed reads. This can be used to annotate new transcript isoforms for known genes, or to identify completely novel genes. For most of the receptor genes, only the CDS is annotated, based on computational predictions. However, there is evidence that at least some receptors contain additional non-coding exons and complex UTRs that undergo alternative splicing[62]. I used Cufflinks[307] to generate *de novo* assemblies in order to identify full length transcripts for OR and VR genes. Samples were sequenced at sufficient depth to produce new, extended receptor gene models for 913 (73%) OR and 246 (43.7%) VR genes. I identified additional exons for many of the receptor genes: 866 and 68 OR genes have exons 5' and 3' to the CDS, respectively; and 163 and 79 VR genes have exons 5' and 3' to the ORF (Figure 2.13).

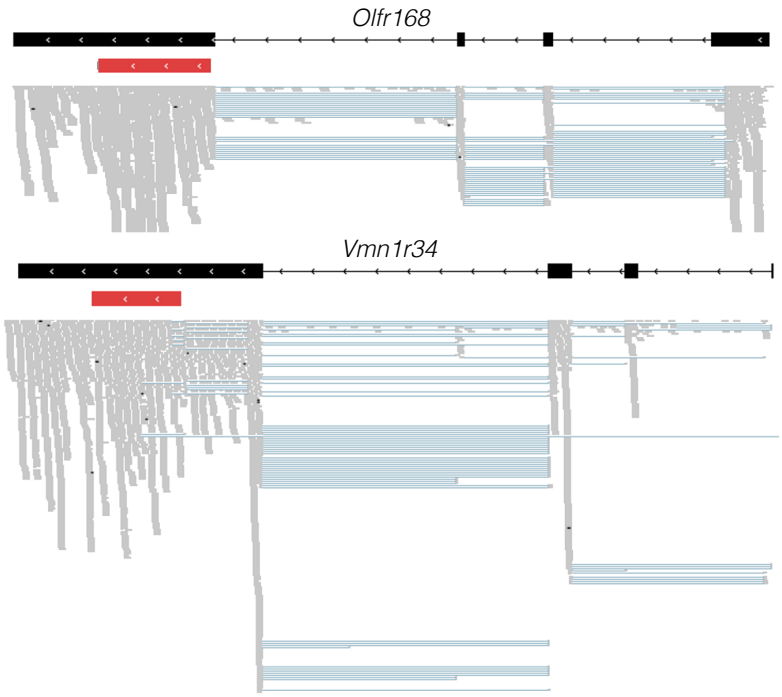

Figure 2.13 – Full length gene models for OR and VR genes. Examples of the generated gene models for an OR (top) and a V1R (bottom) are shown in black. Boxes represent exons and the lines joining them are introns; the arrow heads indicate the strand of the gene. The existing Ensembl annotations for the genes are shown in red with their UTRs in grey. Underneath, the sequencing reads are shown in grey; those that span exon-exon junctions are joined by blue lines.

OR and V1R genes typically have coding regions that span a single exon, but I identified 54 OR and 15 V1R genes where at least one of the reconstructed transcripts

has an intron within the protein coding sequence (as annotated in Ensembl). The predicted ORFs for most of these transcripts are truncated due to a premature stop codon. But for 17 OR and 3 V1R genes the ORF is of typical length, and could encode a putatively functional receptor. All but one (*Olfr332*) are annotated in Ensembl and classified as protein coding.

To investigate cases of alternative splicing, I retained all the multi-exonic receptor gene models and counted the number of alternative isoforms produced. 70% of VR genes have between 1 and 4 isoforms while 85% of OR genes have 1 to 3 isoforms (Figure 2.14). A few receptor genes have more than 8 different isoforms (38 VR and 10 OR genes) but in most of these cases isoforms have alternative transcription start sites (TSS) or exons that differ in length by just a few nucleotides; therefore, several of the final transcripts differ only very slightly.

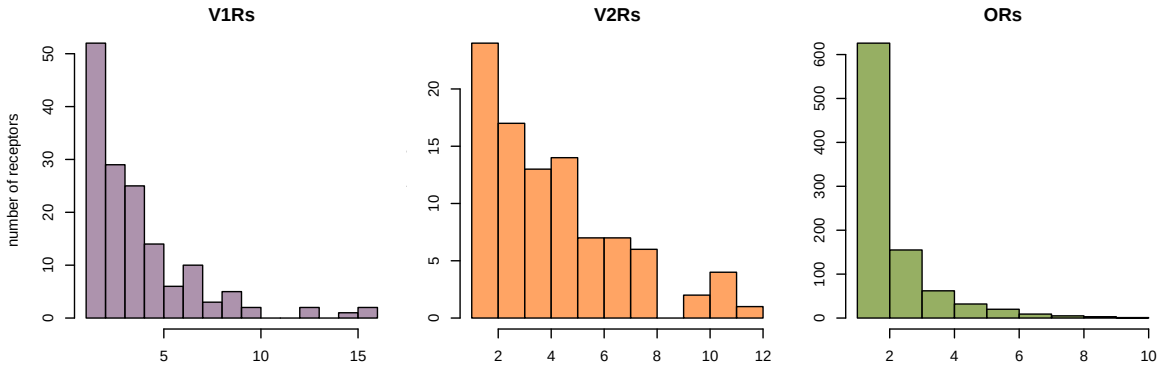

Figure 2.14 – Number of transcripts per receptor gene. Histograms of the number of different transcripts isoforms for each V1R, V2R and OR gene.

Next, I calculated the length for each receptor gene based on the existing Ensembl and the new reconstructed models. The median length for both the Ensembl OR and V1R gene models is about 950 nts, while the corresponding reconstructed gene models are now around 2,500 nts long. The median length of Ensembl V2R genes is 2,559 nts, while for the corresponding reconstructed gene models is 2,912 nts (Figure 2.15A). The increase in length is dominated by a long 3' UTR. UTR sequences are more variable than the CDS[299, 300] and, therefore, could help to differentiate between highly similar OR and particularly VR transcripts. I therefore assessed the uniqueness of these new gene models and found a large increase in the proportion of unique sequence for the V1R (p-value < 0.0001, Mann Whitney test) and V2R gene models (p-value < 0.0001, Mann Whitney test); a more modest increase was apparent in OR genes (p-value $= 0.044$, Mann Whitney test; Figure 2.15B).

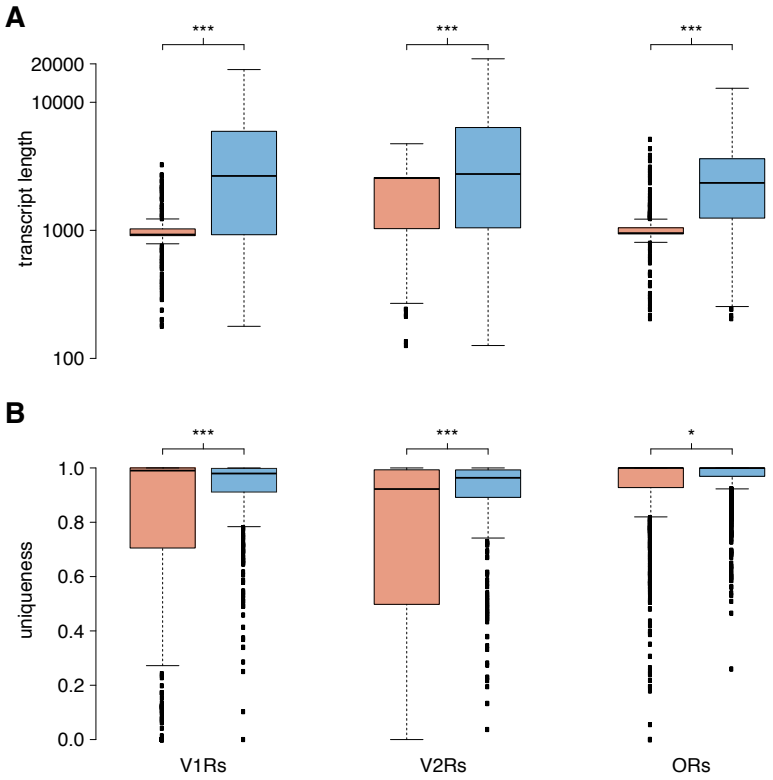

Figure 2.15 – Improved receptor gene models are more unique. **A)** Boxplots of the transcript length in Ensembl (orange) or the reconstructed gene models (blue) for the V1R, V2R and OR genes. The increase in transcript length is highly significant (*** P<0.0001, two-tailed Mann-Whitney test). **B)** As above, but quantifying the proportion of unique sequence (*** P<0.0001 and *P<0.01, two-tailed Mann-Whitney test). The uniqueness corresponds to the proportion of all 100 nucleotide long windows within the transcript that map uniquely to the genome.

Finally, I compared the 5' ends of the OR gene models reconstructed from the RNAseq data to the proposed TSS reported by Plessy et al. using nanoCAGE[193]. A third of the OR genes differ in less than 20 nucleotides, and almost 85% are within a 500 nucleotide window. However 34 OR genes have a discrepancy of more than 5 kb, where the 5' end proposed by nanoCAGE is upstream of the one found by Cufflinks. A close examination of the sequencing data for the 25 genes with the biggest 5' differences revealed that, for 24 genes, there were no sequencing fragments consistent with the TSS proposed by nanoCAGE. In 12 of these cases, the nanoCAGE TSS overlaps with the 3' UTR of an adjacent OR gene and 2 represent the TSS of a different gene. Only one TSS is correctly inferred by nanoCAGE where Cufflinks failed to reconstruct the full-length model. Clowney et al. also defined the 5' end of OR genes using tiling microarrays[194]. A similar comparison analysis revealed that a third of the receptor genes differ in less than 100 nucleotides and 80% of the data is contained within a 1.5 kb

window. Therefore, the reconstructed gene models based on the RNAseq data largely agree with previous reports, but provide a much better resolution and a detailed exon structure that has never been reported for such a large proportion of the receptor repertoire. The defined UTR sequences should aid in the design of more specific probes for experiments based on hybridisation strategies.

To assess how much additional information is provided by the improved annotation of the receptor repertoire and their increased uniqueness, I re-estimated the expression using the reconstructed gene models. The count data was normalised to account for depth of sequencing but not for gene length, since this is different for the new models and some receptors increased more than others. Overall, for those genes with an extended gene model, there is a clear increase in the number of sequencing fragments recovered. On average, expression estimates are 4.4, 1.4 and 1.9 fold more abundant for V1R, V2R and OR genes respectively (Figure 2.16). These additional data should improve the sensitivity and accuracy of the estimated expression levels, and some genes that were previously inaccessible now have unique sequencing fragments mapped to their UTR regions.

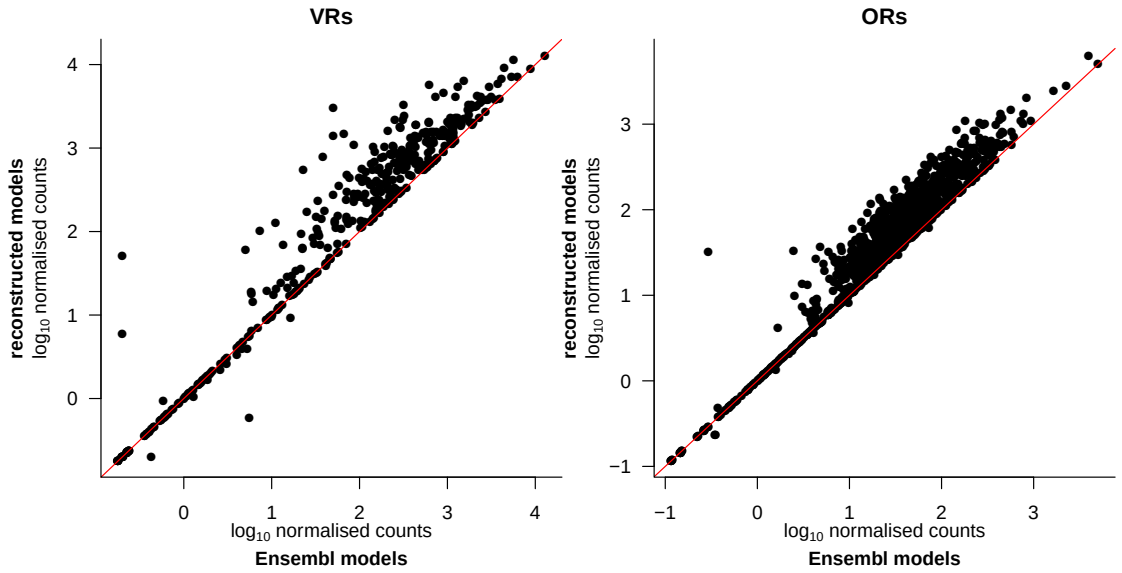

Figure 2.16 – Expression of the receptors with the new gene models. Scatter plot of the normalised expression estimates for each VR (left) and OR (right) gene using the Ensembl gene models (x-axis) versus the reconstructed gene models (y-axis). The red line indicates the 1:1 diagonal.

2.3 Identification of novel genes.

I then extended the analysis of the *de novo* assembly performed by Cufflinks to the full olfactory transcriptomes. This revealed 5,562 and 6,228 loci that have evidence of transcription in the VNO and WOM respectively, that do not overlap any annotated genes in the Ensembl database. 40% of these loci are found in both tissues. Many are located in close proximity to the start or end of annotated genes and are likely to represent unannotated UTRs. To search for new genes, I first excluded all those predictions that lie within 5 kb of cataloged genes. Of the remaining features, about 75% represent single-exon transcripts, leaving 756 and 847 putatively novel multi-exonic genes expressed in the VNO and WOM respectively (Table 2.1).

	VNO	**WOM**
Total identified genes	**5,562**	**6,228**
Shared between tissues	2,331	2,519
Within 5kb of annotated genes	2,564	2,889
Putative novel gene models	**2,998**	**3,339**
Single-exon predictions	2,242	2,492
Multi-exon predictions	**756**	**847**
Aligned to a protein domain	229	258
Overlap with a predicted transcript	625	694

Table 2.1 – Putative novel genes. Number of novel genes predicted from the RNAseq data (that do not overlap any annotated gene in Ensembl). Predictions within 5 kb of annotated genes were excluded and the remaining are considered putative novel gene models. Putative genes are considered shared between tissues if 50% or more of the gene length is found in both the VNO and WOM. In some cases two predicted genes in one tissue can overlap with a single prediction in the other, leading to a different number shared in each.

Ensembl provides databases of annotated genomic regions that match protein domains or computationally predicted genes. When cross-referenced to the putative novel genes, about 30% had matches to known protein domains and 80% overlapped with *in silico* predicted transcripts, suggesting that these might be true genes. Some of them are surprisingly abundantly expressed. One of the genes reconstructed from the VNO data is the 6th most abundant gene in the transcriptome, and it is also present in the WOM samples though at much lower levels. Interestingly, there is a second predicted gene in close proximity with a similar exon-intron structure (Figure 2.17A). These two genes were validated in collaboration with members of the lab. Full-length transcripts were cloned for both of these genes and we could identify ORFs on opposite strands that encode two closely related proteins; by homology, these could be classified as novel

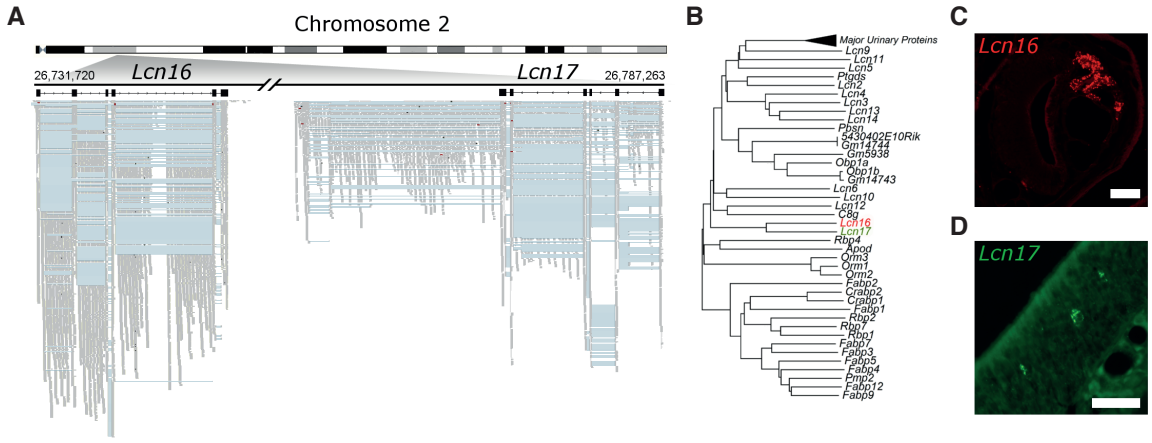

Figure 2.17 – Novel genes expressed in the olfactory system. A) Chromosome 2 is schematized on the top and the locus where two previously unidentified genes were found is amplified below. In black are *Lcn16* and *Lcn17* gene models; boxes correspond to the exons and lines joining them represent introns. The mapped RNAseq reads are below: each read is drawn in grey and blue lines join read fragments that span exon-exon junctions. **B)** Phylogenetic reconstruction of the novel genes with other members of the mouse lipocalin gene family. **C)** *In situ* hybridisation reveals *Lcn16* is expressed in glandular tissue of the VNO and **D)** *Lcn17* is expressed in cells within the MOE. Scale bars: (B) 100 mm, (C) 50 mm.

members of the lipocalin gene family, and were named *Lcn16* and *Lcn17*[4]. A phylogeny of all mouse lipocalins revealed these genes form a distinct sub-clade (Figure 2.17B). *In situ* hybridisation experiments confirmed that *Lcn16* is expressed abundantly in glandular tissues of the VNO, while *Lcn17* is expressed in a small number of cells in the MOE (Figure 2.17C-D)[5]. Despite the mouse genome being one of the best annotated genomes available, this data suggests that there still exists a significant number of unannotated genes. Since the olfactory system is not a tissue routinely used for gene identification, many of these novel genes are likely to have olfactory-related functions and, therefore, would have been missed in other tissues. The validation of two putative genes suggests that many more might be true uncharacterised genes.

Overall, I have shown that RNAseq not only is the best technology available to study the transcriptome of the olfactory system, but it provides reproducible, accurate information about its transcriptome. I have produced a comprehensive profile of the genes expressed in both the VNO and the WOM, that represent the average expression across all the different cell types present therein. Furthermore, I have identified additional genes that have never been characterised before. Most importantly, these data has demonstrated great specificity in the detection of the receptor repertoires which,

[4]The sequences for *Lcn16* and *Lcn17* are available in GenBank under accessions KJ004569 and KJ004570.

[5]The full-length transcripts were obtained by Maria Levitin; the phylogeny was produced by Darren Logan; the *in situ* hybridisation was performed by Luis Saraiva.

until now, have been very difficult to study as a whole. I have demonstrated that the reconstruction of full-length gene models for a significant proportion of the receptor genes allows better estimation of their expression values and increases significantly the amount of unique sequence available for each of them. Some subfamilies of VR genes still represent a challenge, because they are composed of nearly identical genes. These will be very difficult to accurately differentiate with nearly every technique available to date. Nonetheless, a good proportion of the VR repertoire, and nearly the complete OR repertoire are divergent enough across their full length, for accurate discrimination and further study.