Appendix A

Methods

Sample collection and RNA extraction.

All mice used were group housed. The details of the strain, age and sex of each sample can be found in Table B.1 in Appendix B. In the case of the VNO samples, each biological replicate was the pool of three animals. All WOM samples were obtained from a single animal, except the pup WOM samples, which were the pool of 3 or 4 individuals. Tissue was dissected and immediately homogenised in lysis RLT buffer (Qiagen) using a disposable RNAse free plastic grinder, except for the pup samples, which were stored in RNAlater. Total RNA was extracted using the RNeasy mini kit (Qiagen) with on-column DNAse digestion, following the manufacturer's protocol. Tissue homogenisation was performed on a QIAshredder column. All RNA was subsequently quantified with a spectrophotometer and visualised for quality by RNA integrity analysis.

Library preparation and sequencing.

mRNA was prepared for sequencing using the TruSeq RNA sample preparation kit (Illumina). All RNA sequencing was paired-end. The details of the specific Illumina platform used, read length and data strandedness are in Table B.1 in Appendix B. All raw sequencing data are available through the European Nucleotide Archive (ENA); the corresponding accession numbers for each sample can be found in Tables B.1 and B.4 in Appendix B.

RNAseq data processing and mapping.

BAM files were processed using SAMtools[346] and Picard tools version 1.64

(http://broadinstitute.github.io/picard).

Sequencing data were aligned with STAR 2.3[347]. Prior to mapping, the genome index was built with the GTF annotation file under *–sjdbGTFfile* and with option *–sjdbOverhang* 99. Mapping was performed to the GRCm38 mouse reference genome plus the ERCC spike-in sequences, with options *–outFilterMultimapNmax* 1000 *–outFilterMismatchNmax* 4 *–outFilterMatchNmin* 100 *–alignIntronMax* 50000 *–alignMatesGapMax* 50500 *–outSAMstrandField* intronMotif *–outFilterType* BySJout.

The annotation used for the first dataset presented in Chapter 2 was from the Ensembl mouse genome database, version 68 (http://jul2012.archive.ensembl.org/info/data/ftp/index.html). After reconstruction of full-length gene models for the VR and OR gene repertoires (*see below*), the GTF file from the Ensembl mouse genome database version 72 (http://jun2013.archive.ensembl.org/info/data/ftp/index.html) was modified to include all these reconstructed gene models. Additionally, the set of transcripts reported for *Trpc2* contain both short and long isoforms of the gene; the long isoforms represent a fusion with a different gene and were therefore removed[1]. All data was subsequently mapped and analysed using this annotation file (including the initial dataset which was reanalysed). In the case of the single-cell RNAseq data (Chapter 3), the gene *Gm20715* (a predicted gene that undergoes nonsense mediated decay) was also removed from the GTF file because it overlaps with *Olfr1344*; this overlap causes all the reads aligned to the OR to be deemed ambiguous.

Sequencing data was visualised using the Integrative Genomics Viewer (IGV)[348, 349].

Gene expression level estimation and data analysis.

The numbers of fragments uniquely aligned to each gene were obtained using the HTSeq 0.6.1 package, with the script htseq-count, mode *intersection-nonempty*[350]. All multi-mapped fragments were discarded. Data analysis, statistical testing and plotting was carried out in R (http://www.R-project.org). All the heatmaps were produced with the *gplots* package[351] using the $\log_{10}$ transformed normalised counts + 1.

[1]Transcripts removed: ENSMUST00000084843, ENSMUST00000094129, ENSMUST00000094130, ENSMUST00000106950, ENSMUST00000123372, ENSMUST00000125197, ENSMUST00000139104, ENSMUST00000140395, ENSMUST00000141646, ENSMUST00000142629, ENSMUST00000143839, ENSMUST00000146450, ENSMUST00000153176.

RNAseq data normalisation.

Raw counts were normalised to account for sequencing depth between samples, using the procedure implemented in the DESeq2 package[352]. Size factors were calculated with *estimateSizeFactorsForMatrix* and then used to divide the raw counts. For the single-cell data, ERCC spike-ins were not included for data normalisation.

To compare OR expression levels between datasets, normalisation to account for the number of OSNs present in the WOM samples was carried out subsequent to depth normalisation (data presented in Chapters 4 and 5). For this, a method proposed by Khan et al. [304] was used. Five different marker genes were considered, all of which are expressed exclusively in mature OSNs: *Adcy3*, *Ano2*, *Cnga2*, *Gnal* and *Omp*. Further, these have been shown to be expressed at stable levels[304]. To normalise for OSN number the following procedure was applied to the OR normalised counts. First, the correlation between the expression of each of the marker genes and the total number of counts in OR genes was calculated, and all those marker genes with strong correlation values were used. Second, the geometric mean of all marker genes was calculated for each sample. Then, the average of all means was obtained, and divided by each individual mean; this results in the generation of size factors. Third, the OR normalised counts were multiplied by the corresponding size factor.

Differential expression analysis.

To test for differential expression I used DESeq2 1.8.1 with standard parameters. When applied to the single-cell data, the parameter *minReplicatesForReplace* was set to *Inf* to turn off the automatic outlier replacement. Genes were considered differentially expressed if they had an adjusted p-value of 0.05 or less (equivalent to a false discovery rate of 5%). To test for differential expression on the OR repertoire (Chapters 4 and 5) the double normalised counts (accounting for OSN number per sample) were provided directly, and the *normalizationFactors* function was used with size factors of 1 to turn off further normalisation.

Fitting normal distributions to bimodal data.

To deconvolve bimodal distributions into two normal-like distributions I used Gaussian mixture models, through the expectation-maximisation algorithm of the *mixtools* Bioconductor package[353]. In all cases the algorithm converged to optimal values.

Gene enrichment analysis.

To find functional terms enriched in the lists of differentially expressed genes I used GeneTrail with 'Over-/Under-representation Analysis' with default parameters[354]. The background provided were all those genes tested for differential expression (those with an adjusted p-value different to NA).

Microarray profiling.

RNA was extracted from the VNO and WOM of six C57BL/6J males of 10 weeks of age as described above. Profiling was performed on the Illumina MouseWG-6 v2.0 Expression BeadChip following the manufacturer's instructions. Variance stabilising transformation was applied to the data obtained from BeadStudio, which was then quantile normalised using the Bioconductor R package, *lumi*[355].

Recovery of unannotated receptor genes

To recover the entirety of the VR gene repertoire, I took the cDNA sequences as reported[65, 136] and locally aligned them to the mouse genome with BLAST. Then I identified those alignments that overlap genes not annotated as VRs with 100% identity, and changed their name while preserving the Ensembl identifier. In all cases the coordinates obtained from the alignments were concordant with the annotation. A list detailing the gene names that were changed is reported in Table B.3 in Appendix B. Furthermore, 19 additional predicted genes have high identity alignments to other VR sequences. Similarly, I aligned with BLAST all the OR cDNA sequences present in Ensembl v68 and recovered four predicted genes that share high similarity to other ORs. Although these genes are most likely additional members of the VR and OR gene families, proper annotation with novel gene names is required; these were not included as part of the receptor repertoires.

Reconstruction of novel gene models.

To search for novel genes I performed Reference Annotation Based Transcript (RABT) Assembly, using Cufflinks v2.1.1[307] guided by the Ensembl annotation (version 68), with all six replicates of the VNO and WOM data presented in Chapter 2. Assembled transcripts from the different replicates were combined with Cuffmerge. In order to extract the candidates with greatest probability of encoding protein coding genes, I

cross-referenced all predicted loci to the Ensembl databases using the API[356]. *Ad hoc* perl scripts were used to further refine the gene models produced for VR and OR genes, deleting those predictions that fuse adjacent receptor genes or that are antisense to the annotated gene.

Estimation of gene uniqueness.

To calculate the proportion of sequence that is unique in the genome for each receptor gene, I used a perl script to produce all the 32, 76 and 100 nucleotide-long strings that cover the receptor transcripts, either using the Ensembl v68 annotation or the reconstructed gene models by Cufflinks. These were then aligned to the genome with bowtie version 0.12.8[357] and parameters -v 0 -m 1. The unmapped strings were subsequently aligned to the transcriptome, to account for those that span exon-exon junctions. Finally, *ad hoc* perl scripts were used to consolidate the data and count the number of strings that were unique for each gene. The *uniqueness* of a gene was defined as the number of unique strings over the total number of strings for that gene.

Coverage of OR genes.

To obtain the proportion of the OR gene models covered by the mapped sequencing fragments, the BEDtools 2.16.2[358] program coverageBed was used against a BED file containing the merged exonic regions for all isoforms of each OR gene (obtained with mergeBed). The output was then analysed in R to count all positions with at least one mapped fragment to them.

Allelic Discrimination of OR genes.

To determine the allele expressed for each OR in the single-cell data, the Mouse Genomes Project database release 1410 was queried (http://www.sanger.ac.uk/sanger/Mouse_SnpViewer/rel-1410)[321] to obtain all the SNPs for 129P2 that overlap OR gene models. These positions were visualised on IGV and the numbers of fragments containing each nucleotide were extracted.

Creation of pseudo-reference genomes.

To create psuedo-129 and pseudo-CAST genomes, I mined the Mouse Genomes Project data, release v3 (ftp://ftp-mouse.sanger.ac.uk/REL-1303-SNPs_Indels-GRCm38/) to

obtain all the high-quality SNPs and short indels for the 129S5SvEvBrd and CAST/EiJ strains, respectively. These were imputed into the GRCm38 mouse reference genome using Seqnature[359].

Proportional Venn diagrams.

Venn diagrams with areas proportional to the number of elements represented were created using the eulerAPE version 3 software[360].

Dissecting genetic from environmental effects experiment.

To dissect the influence of the genetic background from the olfactory environment between B6 and 129 animals, C57BL/6N and 129S5 4 to 8-cell stage embryos were transferred into F1 (C57BL/6J×CBA) pseudo-pregnant females, and allowed to develop in this equivalent *in utero* environment. One day after birth, the C57BL/6N and 129S5 litters were cross-fostered to C57BL/6N and 129S5 wild-type mothers, respectively. For this, the mothers were removed from their home cage, and the pups to be cross-fostered were introduced to the home-cage of the foster mother; each pup was gently rubbed with nesting material to transfer some of the odours. Then, the mother was introduced into the cage with the new litter, and observed for at least half an hour to ensure it did not reject the pups; those that did were separated from the litter. Then, a single pup from the other strain was transferred to the cross-fostered litter (the *alien*). At weaning, animals from the same sex as the alien animal were kept, always in a 4:1 ratio between strains. If not enough animals of the correct sex were available in the litter, surplus animals from other litters were used. At 10 weeks of age, the WOM was collected from the alien and a randomly selected cage-mate, and RNA was extracted as described previously.

The details on the strain of the alien and cage-mate for each sequenced sample are as follows:

sample	sex	alien	cage-mate
1	female	B6	129
2	female	129	B6
3	female	129	B6
4	female	B6	129
5	male	129	B6
6	male	B6	129

To test for the effect of the environment on gene expression, I used a likelihood ratio test with DESeq2, to test the model genetics+environment+genetics:environment versus accounting only for the genetics; this revealed two significant genes, both ORs. If, instead, the data from each strain was tested separately for the effect of the environment, one of the genes previously identified was again recovered for the B6 data, and a new gene (ENSMUSG00000063779) was significant for the 129 data.

Allelic discrimination of the F1 RNAseq data.

RNAseq data was processed as described above. Total expression estimates were obtained by mapping the RNAseq data to the B6 or pseudo-CAST genomes, with standard parameters. The expression estimates obtained with each genome were very highly correlated. For the OR repertoire, nearly all the genes (96.23%) differed in less than 10 counts and were almost perfectly correlated (rho = 0.9991006, p-value < 2.2e-16). Thus, by allowing 4 mismatches per paired-end fragment, nearly all reads were able to be mapped regardless of the reference used. Therefore, the data mapped to the B6 reference was used in downstream analyses.

To obtain allele-specific expression estimates, the RNAseq data was mapped to both the B6 and the pseudo-CAST genomes, without mismatches. In this way, those reads that span SNPs, could only map to the genome corresponding to the allele they come from. Subsequent analyses were performed on the OR repertoire only. All reads mapped across each SNP were retrieved with SAMtools[346]. In cases where different transcripts exist, and one of them splices across the SNP, SAMtools reports both the reads that map and splice across the SNP. *Ad hoc* perl scripts were used to retain only reads that contained the SNP (using the cigar string) and that were uniquely mapped. Finally, the number of different reads mapping across all SNPs of each gene was obtained. The results using the data mapped to either the B6 or CAST genomes then provide the number of reads that are specific for each allele.

To normalise for depth of sequencing, the total expression raw data was combined with the estimates from the parental strains, and normalised all together. The OR data was then further normalised to account for the number of OSNs, as described above. The same size factors were used to normalise the expression estimates from SNP positions.

To deconvolve the total expression into allele-specific expression, a ratio of the expression of each allele was obtained from the counts in SNP positions with:

$$\frac{\text{counts in B6}}{\text{counts in B6 + counts in CAST}}$$

Then, the total expression normalised counts were multiplied by the ratio to obtain the B6 expression, and to the inverse of the ratio for the CAST-specific expression. Finally, since those genes with very low number of SNPs and/or very low expression have very few reads spanning SNPs, the information is very limited and the estimated ratio is not robust. Thus, only those genes with normalised counts in SNP positions above the lowest quartile were used (82.5%).

Odour-exposure experiments.

To test the effects of enriching the environment with specific odorants, I selected heptanal, (R)-carvone, eugenol and acetophenone because they all have been shown to activate at least one specific OR gene. All odorants were from Sigma, except for acetophenone which was from Alfa Aesar. The mixture of all four consisted of equimolar proportions of each, diluted in mineral oil (Sigma) for a final concentration of 1mM each.

For the chronic exposure experiments, a couple drops of the odour mixture, or mineral oil only, were applied to a cotton ball with a plastic pasteur pipette, for the *exposed* and *control* groups respectively; these were put into metal tea strainers that were then introduced into the cage of the animals. The cotton ball was replaced fresh daily. The odour mix was changed twice a week for a freshly prepared stock. The exposure started from birth and the WOM was collected from age-matched exposed and control groups at different time-points after the start of the treatment.

For the acute exposure experiments, the odour mix or mineral oil was added to the water bottles of the animals. Water bottles were replaced twice a week with freshly prepared ones. The exposure started from at least E14.5 and the WOM was collected from age-matched exposed and control groups at different time-points after the start of the treatment.

The number of animals analysed in each group were as follows:

CHRONIC						
time-point	**control**		**exposed**		**total**	
	males	**females**	**males**	**females**	**control**	**exposed**
4	4	0	5	0	4	5
10	3	0	4	0	3	4
24	5	5	4	5	10	9

ACUTE						
time-point	**control**		**exposed**		**total**	
	males	**females**	**males**	**females**	**control**	**exposed**
1	8		8		8	8
4	5	3	5	5	8	10
10	6	3	6	4	9	10
24	8	5	4	5	13	9
4+6 *	4	4	4	5	8	9

All time-points are in weeks.

*Animals exposed during 4 weeks and then left to recover for 6 weeks.

For the follow-up experiments, animals were acutely exposed only to (R)-carvone, to heptanal alone, or to the combination of both. The final concentration of each odorant was 1mM. The odorants were directly added to the water bottles, without dilution in mineral oil. Therefore, the controls were kept with pure water. The water bottles were changed twice a week. The exposure started from at least E16.5 and the WOM was collected at 10 weeks of age. For each group, 3 males and 3 females were used.

qRT-PCR expression estimation.

For qRT-PCR experiments, RNA from WOM was extracted as previously described. 1 μg of RNA was reversed-transcribed into cDNA using the High-Capacity RNA-to-cDNA kit (Applied Biosystems) with the manufacturer's protocol. Predesigned TaqMan gene expression assays were used on a 7900HT Fast Real-Time PCR System (Life Technologies) following the manufacturer's instructions. Mean cycle threshold (Ct) values were obtained from two technical replicates, each normalised to *Actb* using the ΔCt method. Relative quantity (RQ) values were calculated using the formula RQ = $2^{\Delta Ct}$. Differential expression between groups was assessed in R, by a t-test, with multiple-testing correction by the Benjamini & Hochberg (FDR) method.

Appendix B

Supplementary tables

sample	strain	tissue	age	sex	Illumina platform	read length	stranded	ENA ID
Transcriptome analysis of the WOM and VNO of male and female mice – Chapter 2.								
male1	C57BL/6J	VNO	8-10 weeks	male	Genome Analyzer II	76	no	ERS092040
male2	C57BL/6J	VNO	8-10 weeks	male	Genome Analyzer II	76	no	ERS092041
male3	C57BL/6J	VNO	8-10 weeks	male	Genome Analyzer II	76	no	ERS092042
female1	C57BL/6J	VNO	8-10 weeks	female	Genome Analyzer II	76	no	ERS092043
female2	C57BL/6J	VNO	8-10 weeks	female	Genome Analyzer II	76	no	ERS092044
female3	C57BL/6J	VNO	8-10 weeks	female	Genome Analyzer II	76	no	ERS092045
male1	C57BL/6J	WOM	8-10 weeks	male	HiSeq 2000	76	no	ERS092545
male2	C57BL/6J	WOM	8-10 weeks	male	HiSeq 2000	76	no	ERS092547
male3	C57BL/6J	WOM	8-10 weeks	male	HiSeq 2000	76	no	ERS092549
female1	C57BL/6J	WOM	8-10 weeks	female	HiSeq 2000	76	no	ERS092546
female2	C57BL/6J	WOM	8-10 weeks	female	HiSeq 2000	76	no	ERS092548
female3	C57BL/6J	WOM	8-10 weeks	female	HiSeq 2000	76	no	ERS092550

sample	strain	tissue	age	sex	Illumina platform	read length	stranded	ENA ID
RNAseq of mice lacking a cluster of OR genes in chromosome 9 – Chapter 2.								
delta1	129/SvEv-ΔOlfr7Δ	WOM	9 weeks	male	HiSeq 2500	100	yes	ERS473426
delta2	129/SvEv-ΔOlfr7Δ	WOM	9 weeks	male	HiSeq 2500	100	yes	ERS473427
delta3	129/SvEv-ΔOlfr7Δ	WOM	9 weeks	male	HiSeq 2500	100	yes	ERS473428
Comparison of the transcriptome of the OSNs versus the WOM – Chapter 3.								
WOM1	OMP-GFP	WOM	21 days	male	HiSeq 2000	100	no	ERS252155
WOM2	OMP-GFP	WOM	21 days	male	HiSeq 2000	100	no	ERS252156
WOM3	OMP-GFP	WOM	21 days	female	HiSeq 2000	100	no	ERS252157
OSN1	OMP-GFP	FACS OSNs	25 days	mixed	HiSeq 2000	100	no	ERS252158
OSN2	OMP-GFP	FACS OSNs	25 days	mixed	HiSeq 2000	100	no	ERS252159
OSN3	OMP-GFP	FACS OSNs	25 days	mixed	HiSeq 2000	100	no	ERS252160
Characterisation of two subpopulations of OMP$^+$ OSNs – Chapter 3.								
GFP^{low}1	OMP-GFP	FACS OSNs	25 weeks	male	HiSeq 2500	100	no	ERS715983
GFP^{low}2	OMP-GFP	FACS OSNs	25 weeks	male	HiSeq 2500	100	no	ERS715985
GFP^{low}3	OMP-GFP	FACS OSNs	25 weeks	male	HiSeq 2500	100	no	ERS715987
GFP^{high}1	OMP-GFP	FACS OSNs	25 weeks	male	HiSeq 2500	100	no	ERS715984
GFP^{high}2	OMP-GFP	FACS OSNs	25 weeks	male	HiSeq 2500	100	no	ERS715986
GFP^{high}3	OMP-GFP	FACS OSNs	25 weeks	male	HiSeq 2500	100	no	ERS715988

sample	strain	tissue	age	sex	Illumina platform	read length	stranded	ENA ID
The transcriptome of single OSNs – Chapter 3.								
single OSNs	OMP-GFP	single OSNs	23 weeks	male	HiSeq 2500	100	no	See Table B.4 for details
Comparison of the OR expression profile in different strains of mice – Chapter 4.								
B6_1	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658588
B6_2	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658589
B6_3	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658590
B6_4	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658591
B6_5	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658592
B6_6	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658593
129_1	129S5/SvEv	WOM	11 weeks	male	HiSeq 2000	100	no	ERS215497
129_2	129S5/SvEv	WOM	11 weeks	male	HiSeq 2000	100	no	ERS215498
129_3	129S5/SvEv	WOM	11 weeks	male	HiSeq 2000	100	no	ERS215499
cast1	CAST/Ei	WOM	12 weeks	female	HiSeq 2500	100	yes	ERS473423
cast2	CAST/Ei	WOM	12 weeks	female	HiSeq 2500	100	yes	ERS473424
cast3	CAST/Ei	WOM	12 weeks	female	HiSeq 2500	100	yes	ERS473425
Dissecting the genetic from the environmental effects on OR gene expression – Chapter 4.								
black1	C57BL/6NTac	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373470
black2	C57BL/6NTac	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373471
black3	C57BL/6NTac	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373472
black4	C57BL/6NTac	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373473
black5	C57BL/6NTac	WOM	10 weeks	male	HiSeq 2000	100	no	ERS373474
black6	C57BL/6NTac	WOM	10 weeks	male	HiSeq 2000	100	no	ERS373475
agouti1	129S5/SvEv	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373476
agouti2	129S5/SvEv	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373477
agouti3	129S5/SvEv	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373478
agouti4	129S5/SvEv	WOM	10 weeks	female	HiSeq 2000	100	no	ERS373479
agouti5	129S5/SvEv	WOM	10 weeks	male	HiSeq 2000	100	no	ERS373480
agouti6	129S5/SvEv	WOM	10 weeks	male	HiSeq 2000	100	no	ERS373481

sample	strain	tissue	age	sex	Illumina platform	read length	stranded	ENA ID
Transcriptome of the WOM of newborn mice – Chapter 4.								
pups1	C57BL/6J	WOM	E19.5	mixed	HiSeq 2000	100	no	ERS223116
pups2	C57BL/6J	WOM	E19.5	mixed	HiSeq 2000	100	no	ERS223117
pups3	C57BL/6J	WOM	E19.5	mixed	HiSeq 2000	100	no	ERS223118
OR expression after exposure to a mix of odorants – Chapter 5.								
control1	C57BL/6J	WOM	24 weeks	male	HiSeq 2500	100	yes	ERS427453
control2	C57BL/6J	WOM	24 weeks	male	HiSeq 2500	100	yes	ERS427454
control3	C57BL/6J	WOM	24 weeks	male	HiSeq 2500	100	yes	ERS427455
control4	C57BL/6J	WOM	24 weeks	female	HiSeq 2500	100	yes	ERS427456
control5	C57BL/6J	WOM	24 weeks	female	HiSeq 2500	100	yes	ERS427457
control6	C57BL/6J	WOM	24 weeks	female	HiSeq 2500	100	yes	ERS427458
odour1	C57BL/6J	WOM	24 weeks	male	HiSeq 2500	100	yes	ERS427447
odour2	C57BL/6J	WOM	24 weeks	male	HiSeq 2500	100	yes	ERS427448
odour3	C57BL/6J	WOM	24 weeks	male	HiSeq 2500	100	yes	ERS427449
odour4	C57BL/6J	WOM	24 weeks	female	HiSeq 2500	100	yes	ERS427450
odour5	C57BL/6J	WOM	24 weeks	female	HiSeq 2500	100	yes	ERS427451
odour6	C57BL/6J	WOM	24 weeks	female	HiSeq 2500	100	yes	ERS427452
OR expression after exposure to particular odorants – Chapter 5.								
carvone1	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658594
carvone2	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658595
carvone3	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658596
carvone4	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658597
carvone5	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658598
carvone6	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658599
heptanal1	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658600
heptanal2	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658601
heptanal3	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658602
heptanal4	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658603
heptanal5	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658604
heptanal6	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658605

sample	strain	tissue	age	sex	Illumina platform	read length	stranded	ENA ID
both1	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658606
both2	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658607
both3	C57BL/6J	WOM	10 weeks	male	HiSeq 2500	100	yes	ERS658608
both4	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658609
both5	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658610
both6	C57BL/6J	WOM	10 weeks	female	HiSeq 2500	100	yes	ERS658611

Table B.1 – Sequenced samples presented in this dissertation. Details about each of the samples used for RNAseq. All sequencing was paired-end; the read length is indicated, in basepairs. *Stranded* indicates whether the library preparation method was strand-specific or not. All raw data are available through the European Nucleotide Archive (ENA).

sample	total fragments	uniquely mapped %		multimapped %		unmapped %	
Transcriptome analysis of the VNO of male and female mice – Chapter 2.							
male1	33,829,828	27,791,186	82.15	2,356,980	6.97	3,681,662	10.88
male2	34,334,069	27,943,814	81.39	2,300,990	6.70	4,089,265	11.91
male3	33,452,308	26,979,727	80.65	2,259,065	6.75	4,213,516	12.60
female1	38,989,649	30,690,761	78.72	2,517,113	6.46	5,781,775	14.83
female2	41,267,287	33,471,377	81.11	1,828,650	4.43	5,967,260	14.46
female3	40,783,743	33,330,682	81.73	2,907,635	7.13	4,545,426	11.15
Transcriptome analysis of the WOM of male and female mice – Chapter 2.							
male1	47,449,378	43,428,430	91.53	2,422,702	5.11	1,598,246	3.37
male2	45,919,675	41,968,773	91.40	2,815,735	6.13	1,135,167	2.47
male3	45,436,958	38,304,453	84.30	5,906,163	13.00	1,226,342	2.70
female1	41,096,169	35,868,924	87.28	4,003,075	9.74	1,224,170	2.98
female2	53,985,044	46,021,315	85.25	6,361,505	11.78	1,602,224	2.97
female3	44,548,659	38,716,506	86.91	4,838,407	10.86	993,746	2.23
RNAseq of mice lacking a cluster of OR genes in chromosome 9 – Chapter 2.							
delta1	40,815,069	37,230,201	91.22	2,428,922	5.95	1,155,946	2.83
delta2	41,774,414	38,165,713	91.36	2,444,901	5.85	1,163,800	2.79
delta3	48,779,436	44,509,100	91.25	2,975,024	6.10	1,295,312	2.66
Comparison of the transcriptome of the OSNs versus the WOM – Chapter 3.							
WOM1	43,534,928	38,820,863	89.17	1,723,998	3.96	2,990,067	6.87
WOM2	75,289,455	67,690,537	89.91	3,018,346	4.01	4,580,572	6.08
WOM3	54,231,767	49,952,440	92.11	2,316,878	4.27	1,962,449	3.62
OSN1	48,523,309	45,373,409	93.51	1,764,633	3.64	1,385,267	2.85
OSN2	57,565,818	46,820,001	81.33	2,142,656	3.72	8,603,161	14.94
OSN3	75,288,647	69,454,921	92.25	2,506,461	3.33	3,327,265	4.42
Characterisation of two subpopulations of OMP$^+$ OSNs – Chapter 3.							
GFP^{low}1	66,274,523	58,952,254	88.95	3,619,369	5.46	3,702,900	5.59
GFP^{low}2	66,293,232	59,198,484	89.30	3,343,118	5.04	3,751,630	5.66
GFP^{low}3	80,748,448	72,241,026	89.46	4,113,321	5.09	4,394,101	5.44
GFP^{high}1	17,734,782	15,059,594	84.92	908,067	5.12	1,767,121	9.96
GFP^{high}2	72,410,349	64,410,818	88.95	3,528,935	4.87	4,470,596	6.17
GFP^{high}3	73,843,358	65,957,173	89.32	3,764,830	5.10	4,121,355	5.58

sample	total fragments	uniquely mapped %		multimapped %		unmapped %	
Comparison of the OR expression profile in different strains of mice – Chapter 4.							
B6_1	37,332,765	32,250,593	86.39	2,121,241	5.68	5,082,164	7.93
B6_2	58,184,940	51,710,263	88.87	3,191,347	5.48	6,474,671	5.65
B6_3	46,459,677	40,920,392	88.08	3,338,396	7.19	5,539,280	4.74
B6_4	39,423,479	34,897,650	88.52	2,466,335	6.26	4,525,824	5.22
B6_5	24,228,124	20,190,426	83.33	1,468,759	6.06	4,037,687	10.61
B6_6	36,811,932	32,155,493	87.35	2„712,152	7.37	4,656,434	5.28
129_1	51,360,567	41,641,267	81.08	6,228,439	12.13	3,490,861	6.80
129_2	56,018,117	49,911,723	89.10	2,897,953	5.17	3,208,441	5.73
129_3	75,872,597	68,102,703	89.76	4,015,197	5.29	3,754,697	4.95
cast1	42,193,697	38,185,627	90.50	2,144,719	5.08	1,863,351	4.42
cast2	35,534,499	31,307,518	88.10	2,553,077	7.18	1,673,904	4.71
cast3	46,273,696	41,504,133	89.69	2,618,676	5.66	21,50,887	4.65
Dissecting the genetic from the environmental effects on OR gene expression – Chapter 4.							
black1	53,532,994	45,079,303	84.21	5,431,071	10.15	3,022,620	5.65
black2	74,253,096	64,338,991	86.65	5,185,893	6.98	4,728,212	6.37
black3	41,608,225	37,543,713	90.23	1,937,432	4.66	2,127,080	5.11
black4	71,212,832	63,952,008	89.80	3,387,536	4.76	3,873,288	5.44
black5	51,920,894	45,650,480	87.92	3,106,416	5.98	3,163,998	6.09
black6	90,279,406	77,481,328	85.82	7,415,649	8.21	5,382,429	5.96
agouti1	60,959,853	53,430,862	87.65	3,088,901	5.07	4,440,090	7.28
agouti2	26,709,804	23,547,335	88.16	1,286,307	4.82	1,876,162	7.02
agouti3	30,791,098	26,913,940	87.41	1,483,789	4.82	2,393,369	7.77
agouti4	49,844,784	43,363,086	87.00	3,638,676	7.30	2,843,022	5.70
agouti5	34,387,223	29,030,014	84.42	3,412,419	9.92	1,944,790	5.66
agouti6	41,895,931	37,528,206	89.57	1,927,666	4.60	2,440,059	5.82
Transcriptome of the WOM of newborn mice – Chapter 4.							
pups1	49,335,880	41,680,801	84.48	3,860,509	7.82	3,794,570	7.69
pups2	40,532,710	35,680,710	88.03	2,830,765	6.98	2,021,235	4.99
pups3	64,224,553	57,477,476	89.49	4,100,707	6.38	2,646,370	4.12
OR expression after exposure to a mix of odorants – Chapter 5.							
control1	52,160,507	46,117,696	88.41	4,374,645	8.39	1,668,166	3.20

sample	total fragments	uniquely mapped %		multimapped %		unmapped %	
control2	45,667,031	41,594,467	91.08	2,571,602	5.63	1,500,962	3.29
control3	45,665,776	41,565,777	91.02	2,501,129	5.48	1,598,870	3.50
control4	54,725,715	49,969,350	91.31	3,109,856	5.68	1,646,509	3.01
control5	46,906,572	42,584,649	90.79	2,784,963	5.94	1,536,960	3.28
control6	51,235,209	45,397,000	88.61	4,063,968	7.93	1,774,241	3.46
odour1	53,005,866	48,477,141	91.46	2,796,147	5.28	1,732,578	3.27
odour2	44,239,992	39,344,128	88.93	3,379,210	7.64	1,516,654	3.43
odour3	50,470,024	45,624,241	90.40	3,155,133	6.25	1,690,650	3.35
odour4	48,495,642	43,996,672	90.72	2,870,256	5.92	1,628,714	3.36
odour5	50,189,225	43,976,059	87.62	4,345,171	8.66	1,867,995	3.72
odour6	50,338,265	45,770,319	90.93	2,824,275	5.61	1,743,671	3.46
OR expression after exposure to particular odorants – Chapter 5.							
carvone1	34,538,910	30,563,297	88.49	1,892,137	5.48	2,083,476	6.03
carvone2	35,531,736	31,117,642	87.58	1,800,586	5.07	2,613,508	7.36
carvone3	32,064,615	27,022,540	84.28	3,134,406	9.78	1,907,669	5.95
carvone4	33,834,834	29,883,589	88.32	1,984,988	5.87	1,966,257	5.81
carvone5	41,188,840	36,500,265	88.62	2,551,702	6.20	2,136,873	5.19
carvone6	33,179,966	28,071,868	84.60	3,700,369	11.15	1,407,729	4.24
heptanal1	36,368,912	30,577,330	84.08	3,670,436	10.09	2,121,146	5.83
heptanal2	47,799,810	42,578,959	89.08	2,691,320	5.63	2,529,531	5.29
heptanal3	71,528,814	63,800,289	89.20	3,801,877	5.32	3,926,648	5.49
heptanal4	34,423,907	31,050,666	90.20	1,816,408	5.28	1,556,833	4.52
heptanal5	39,095,683	34,761,037	88.91	2,190,893	5.60	2,143,753	5.48
heptanal6	14,901,039	12,886,192	86.48	1,105,993	7.42	908,854	6.10
both1	44,818,699	39,855,390	88.93	2,429,371	5.42	2,533,938	5.65
both2	37,287,128	33,330,936	89.39	2,084,171	5.59	1,872,021	5.02
both3	43,729,818	31,058,026	71.02	3,500,396	8.00	9,171,396	20.97
both4	50,876,569	45,689,933	89.81	2,696,457	5.30	2,490,179	4.89
both5	35,849,711	32,049,662	89.40	1,908,123	5.32	1,891,926	5.28
both6	41,060,091	35,784,547	87.15	3,004,507	7.32	2,271,037	5.53

Table B.2 – Mapping statistics of RNAseq samples. Mapping statistics of the samples sequenced (see also Table B.1).

Ensembl ID	Ensembl gene name	Matched cDNA from [65,136]
ENSMUSG00000096294	*Gm10302*	*Vmn2r47*
ENSMUSG00000096871	*Gm10665*	*Vmn1r102*
ENSMUSG00000096348	*Gm10666*	*Vmn1r141.Vmn1r93*
ENSMUSG00000094762	*Gm10670*	*Vmn1r150*
ENSMUSG00000087688	*Gm11300*	*Vmn1r203*
ENSMUSG00000087643	*Gm11314*	*Vmn1r208*
ENSMUSG00000096152	*Gm16442*	*Vmn1r140*
ENSMUSG00000095745	*Gm4133*	*Vmn1r146*
ENSMUSG00000095837	*Gm4141*	*Vmn1r106*
ENSMUSG00000093941	*Gm4172*	*Vmn1r131*
ENSMUSG00000096513	*Gm4175*	*Vmn1r133*
ENSMUSG00000096760	*Gm4177*	*Vmn1r134*
ENSMUSG00000095163	*Gm4179*	*Vmn1r138*
ENSMUSG00000093871	*Gm4187*	*Vmn1r98*
ENSMUSG00000095984	*Gm4201*	*Vmn1r154*
ENSMUSG00000092297	*Gm4214*	*Vmn1r161*
ENSMUSG00000094532	*Gm4216*	*Vmn1r162*
ENSMUSG00000096073	*Gm4220*	*Vmn1r166*
ENSMUSG00000094757	*Gm4498*	*Vmn1r145*
ENSMUSG00000095191	*Gm5725*	*Vmn1r136*
ENSMUSG00000096761	*Gm5726*	*Vmn1r105*
ENSMUSG00000095806	*Gm5728*	*Vmn1r147*
ENSMUSG00000094298	*Gm6164*	*Vmn1r144*
ENSMUSG00000094149	*Gm8453*	*Vmn1r97*
ENSMUSG00000094981	*Gm8653*	*Vmn1r96*
ENSMUSG00000093917	*Gm8660*	*Vmn1r99*
ENSMUSG00000094748	*Gm8677*	*Vmn1r153*
ENSMUSG00000095081	*Gm8693*	*Vmn1r108.Vmn1r156*
ENSMUSG00000096601	*Gm8720*	*Vmn1r164*
ENSMUSG00000091528	*Gm9268*	*Vmn2r64*
ENSMUSG00000096304	*RP23-331M13.5*	*Vmn1r92*
ENSMUSG00000092456	*V1rd19*	*Vmn1r182*

Table B.3 – VR genes not properly annotated in Ensembl. The matched cDNA sequences are those that aligned with 100% coverage and 100% identity, indicating that they represent the same gene but haven't been properly annotated in Ensembl. Other genes matched VR sequences with lower identity and most likely represent unannotated paralogs, but were not included in the analyses since there is a lack of annotation for them.

sample	total fragments	unique	%	multimapped	%	unmapped	%	included	ENA ID
OSN_171	5,403,186	4,587,828	84.91	296,854	5.49	518,504	9.59	yes	ERS361292
OSN_177	3,416,492	2,962,593	86.71	171,355	5.02	282,544	8.27	yes	ERS361298
OSN_183	4,884,518	4,139,987	84.76	259,317	5.31	485,214	9.93	yes	ERS361304
OSN_188	4,087,523	3,468,899	84.87	235,054	5.75	383,570	9.38	yes	ERS361309
OSN_193	3,490,923	2,908,510	83.32	164,005	4.70	418,408	11.98	yes	ERS361314
OSN_195	3,376,521	2,496,006	73.92	390,214	11.56	490,301	14.52	yes	ERS361316
OSN_201	4,604,541	4,055,065	88.07	185,030	4.02	364,446	7.91	yes	ERS361322
OSN_204	4,187,094	3,587,283	85.67	204,236	4.88	395,575	9.45	yes	ERS361325
OSN_205	5,487,975	4,787,155	87.23	260,788	4.75	440,032	8.02	yes	ERS361326
OSN_216	4,805,114	4,155,706	86.49	257,473	5.36	391,935	8.16	yes	ERS361337
OSN_222	4,080,624	3,481,018	85.31	241,642	5.92	357,964	8.77	yes	ERS361343
OSN_224	3,370,232	2,723,513	80.81	149,548	4.44	497,171	14.75	yes	ERS361345
OSN_230	4,138,735	3,379,738	81.66	247,198	5.97	511,799	12.37	yes	ERS361351
OSN_236	2,962,912	2,467,855	83.29	104,321	3.52	390,736	13.19	yes	ERS361357
OSN_238	3,633,203	3,058,970	84.19	214,892	5.91	359,341	9.89	yes	ERS361359
OSN_243	5,146,808	4,440,478	86.28	269,240	5.23	437,090	8.5	yes	ERS361364
OSN_251	5,069,051	4,216,080	83.17	217,145	4.28	635,826	12.54	yes	ERS361372
OSN_259	4,997,202	4,331,122	86.67	197,519	3.95	468,561	9.38	yes	ERS361380
OSN_261	6,936,460	6,092,355	87.83	297,816	4.29	546,289	7.87	yes	ERS361382
OSN_262	4,420,237	3,770,751	85.31	277,302	6.27	372,184	8.42	yes	ERS361383
OSN_263	5,688,875	4,560,119	80.16	399,782	7.03	728,974	12.81	yes	ERS361384
OSN_178	3,078,169	2,644,556	85.91	165,023	5.36	268,590	8.73	no	ERS361299
OSN_185	4,404,713	3,789,209	86.03	240,960	5.47	374,544	8.51	no	ERS361306
OSN_191	4,136,140	3,487,140	84.31	285,589	6.90	363,411	8.79	no	ERS361312
OSN_207	4,378,400	3,830,209	87.48	175,419	4.01	372,772	8.52	no	ERS361328
OSN_214	4,120,952	3,398,111	82.46	270,927	6.57	451,914	10.96	no	ERS361335
OSN_218	4,693,293	3,890,545	82.90	242,345	5.16	560,403	11.94	no	ERS361339
OSN_223	3,897,470	3,319,616	85.17	185,317	4.75	392,537	10.07	no	ERS361344
OSN_255	4,790,253	3,705,418	77.35	449,852	9.39	634,983	13.25	no	ERS361376
OSN_257	4,911,385	4,252,816	86.59	172,097	3.50	486,472	9.91	no	ERS361378

Table B.4 – Mapping statistics of RNAseq single-OSN samples. Mapping statistics of the single-OSN samples sequenced. Column *included* indicates whether the sample was included in downstream analyses after the QC stage. Excluded samples showed expression of more than a single abundant OR gene and represent carry-over from adjacent wells or could contain two cells. All raw data is available through the European Nucleotide Archive (ENA).

Appendix C

Papers produced during my PhD.

C.1 Papers associated with this dissertation.

- Ibarra-Soria, X., Levitin, M. O. & Logan, D. W. The genomic basis of vomeronasal-mediated behaviour. *Mamm Genome* **25**, 75–86 (2014). DOI: 10.1007/s00335-013-9463-1.

- Ibarra-Soria, X., Levitin, M. O., Saraiva, L. R. & Logan, D. W. The olfactory transcriptomes of mice. *PLoS Genet* **10**, e1004593 (2014). DOI: 10.1371/journal.pgen.1004593.

- Saraiva, L. R.*, Ibarra-Soria, X.*, Khan, M., Omura, M., Scialdone, A., Mombaerts, P., Marioni, J. C. & Logan, D. W. Hierarchical deconstruction of mouse olfactory sensory neurons: from whole mucosa to single-cell RNA-seq. *Sci Rep* **5**, 18178 (2015). DOI: 10.1038/srep18178.

  * Contributed equally.

C.2 Other papers.

- Dey, S., Chamero, P., Pru, J. K., Chien, M.-S., Ibarra-Soria, X., Spencer, K. R., Logan, D. W., Matsunami, H., Peluso, J. J., & Stowers, L. Cyclic regulation of sensory perception by a female hormone alters behavior. *Cell* **161**, 1334–1344 (2015). DOI: 10.1016/j.cell.2015.04.052.

- Oboti, L., Ibarra-Soria, X., Pérez-Gómez, A., Schmid, A., Pyrski, M., Paschek, N., Kircher, S., Logan, D. W., Leinders-Zufall, T., Zufall, F. & Chamero, P. Pregnancy

and estrogen enhance neural progenitor-cell proliferation in the vomeronasal sensory epithelium. *BMC Biology* **13**, 104 (2015). DOI: 10.1186/s12915-015-0211-8.

- Nakahara, T. S., Cardozo, L. M., Ibarra-Soria, X., Bard, A., Carvalho, V. M. A., Trintinalia, G. Z. , Logan, D. W., & Papes, F., Detection of pup odors by adult vomeronasal neurons non-canonically expressing an odorant receptor gene is influenced by sex and parenting status, *BMC Biology*, **In press**.