**Supplemental Figures**

**Figure S1 (related to Figure 1). Expression of OMP-receptor transgenes in zebrafish embryos.** To determine whether OMP-OR transgenes are repressed in zebrafish olfactory sensory neurons, zebrafish embryos were co-injected at the one-cell stage with equivalent amounts of the following DNA plasmids: *OMP-unc76:GFP* + *OMP-unc76:mCherry* (control); or *OMP-MOR28:GFP* + *OMP-unc76:mCherry* (treatment); *MOR28* encodes a mouse OR. Embryos were harvested at 3 dpf and processed for double-label immunohistochemistry to localize GFP and mCherry expression. GFP was detected using an anti-chicken GFP antibody and Alexa488-conjugated goat anti-chicken IgG; mCherry was detected using a rabbit anti-DsRed antibody and Alexa568-conjugated goat anti-rabbit IgG. The number of cells expressing each transgene marker was determined (n = 10 fish per condition). The mean ratio of GFP:mCherry cells was 1.06 for the control condition (blue circles) and 0.372 for the treatment condition (red triangles); $p < 10^{-5}$ (two-tailed t-test).

**Figure S2 (related to Figures 3 and 6; Tables 1 and 2). Effects of perturbations of G protein signaling on OR gene expression and olfactory neuron maturation. (A)** Ectopic expression of constitutively active $G\alpha_s$ does not alter endogenous OR gene expression. The number of cells expressing endogenous odorant receptors OR111-6 (left column), OR111-1 (middle column) or OR103-2 (right column) was determined by RNA in situ hybridization on 3 dpf zebrafish embryos previously injected at the one-cell stage with a transgene encoding a constitutively active $G\alpha_s$ mutant under the control of the OMP promoter (*OMP-Gαs**). Plots show the number of receptor-positive cells per embryo injected with *OMP-Gαs** (red triangles) or *OMP-unc76:GFP* control (blue circles). Each pair of treatment and control plots represents an independent experiment for a particular transgene and probe; within an experiment, 15-20 embryos were analyzed for each treatment and control condition. Fitted cell counts (horizontal bars) and *p*-values for the test of transgene effects were obtained from generalized linear models with a Poisson distribution (see Experimental Procedures). The numbers of cells expressing any of the ORs tested were not significantly different between treatment and control conditions (p = 0.9  for OR111-6; p = 0.3 for OR111-1; p = 0.1 for OR103-2). n.s., not significant. **(B)** Inhibition of Gβγ signaling does not affect the appearance of mature olfactory sensory neurons. Stably transgenic zebrafish embryos (*TgOMP-Gal4;UAS-GCaMP1.6*) expressing GCaMP under the control of the OMP promoter were treated with gallein, BIX, PCPA or 0.2% DMSO (vehicle control) starting at 16-20 hpf, harvested at 72 hpf, processed for immunohistochemistry using an anti-GFP antibody and visualized by confocal microscopy. Confocal reconstructions reveal that none of the drug treatments causes discernible perturbations in the appearance of mature OMP transgene-expressing neurons or their innervation of the olfactory bulb (micrographs of one side of representative embryos are shown). OE, olfactory epithelium; OB, olfactory bulb; dorsal-ventral and medial-lateral axes are indicated; scale bar = 20 $\mu$m. **(C)** Quantitation of OMP transgene-positive olfactory sensory neurons in *TgOMP-Gal4;UAS-GCaMP1.6* transgenic fish reveals no statistically significant difference in GFP-positive cells between drug-treated and control embryos (two-tailed t-tests: gallein vs. control, p = 0.3; BIX vs. control, p = 0.7; PCPA vs. control, p = 0.9). Confocal stacks from individual fish were coded and scored blind. Histograms represent mean values ± SEM for 11-12 fish per conditions. **(D)** Inhibition of Gβγ signaling does not alter the number of mitotic or apoptotic cells in the olfactory placode. Embryos were treated with gallein starting at 16-20 hpf and harvested at 72 hpf for immunohistochemistry using antibodies specific for phospho-histone H3 (PH3), a marker of mitotic cells, or activated caspase, a marker of cells undergoing apoptosis; a subset of embryos was also processed for RNA in situ hybridization using a probe for OR111-1. Cells expressing these markers in the olfactory placode were counted from 11-15 embryos per condition and marker. The number of PH3-positive and activated caspase-positive cells was indistinguishable between control and gallein-treated fish ($p = 0.5$ and 0.9, respectively), whereas a highly significant difference was observed for the number of OR111-1-positive cells ($p < 10^{-4}$). Histograms summarizing this analysis are shown; error bars represent standard errors of the means. **(E)** Inhibition of Gβγ signaling by RACKnt alters OR gene expression. The number of cells expressing endogenous odorant receptors OR111-1 (left column) or OR103-2 (right column) was determined by RNA in situ hybridization on 3 dpf zebrafish embryos previously injected at the one-cell stage with a transgene encoding an N-terminal peptide of Receptor for Activated C-Kinase (RACKnt) under the control of the OMP promoter (*OMP-RACKnt*). Plots show the number of receptor-positive cells per embryo injected with *OMP-RACKnt* (red triangles) or *OMP-unc76:GFP* control (blue circles). Each pair of treatment and control plots represents an independent experiment for a particular transgene and probe; within an experiment, 15-20 embryos were analyzed for each treatment and control condition. Fitted cell counts (horizontal bars) and *p*-values for the test of transgene effects were obtained from generalized linear models with a Poisson distribution (see Experimental Procedures). Expression of the OMP-RACKnt transgene resulted in a 1.5-fold and 1.4-fold increase in the number of cells expressing OR111-1 and OR103-2, respectively, compared to controls (**** $p < 10^{-5}$).

**Figure S3 (related to Figure 6). Perturbations of histone methylation causes alterations in OR gene expression.** The effect of inhibiting H3K9 methylation was quantitated in 3 dpf zebrafish embryos treated with 20 $\mu$M UNC0638, a G9a/GLP histone methyltransferase inhibitor. The number of cells expressing endogenous odorant receptors OR111-1 (left column) or OR103-2 (right column) was determined by RNA in situ hybridization. Plots show the number of receptor-positive cells per embryo in control fish (blue circles) or fish treated with gallein (red triangles). Three independent experiments were performed, as indicated. Experimental design and GLM-based analysis are as described in Experimental Procedures. Fitted cell counts are indicated by horizontal bars. Treatment with UNC0638 resulted in a 1.7-fold increase in the number of cells expressing either receptor (**** $p < 10^{-5}$).

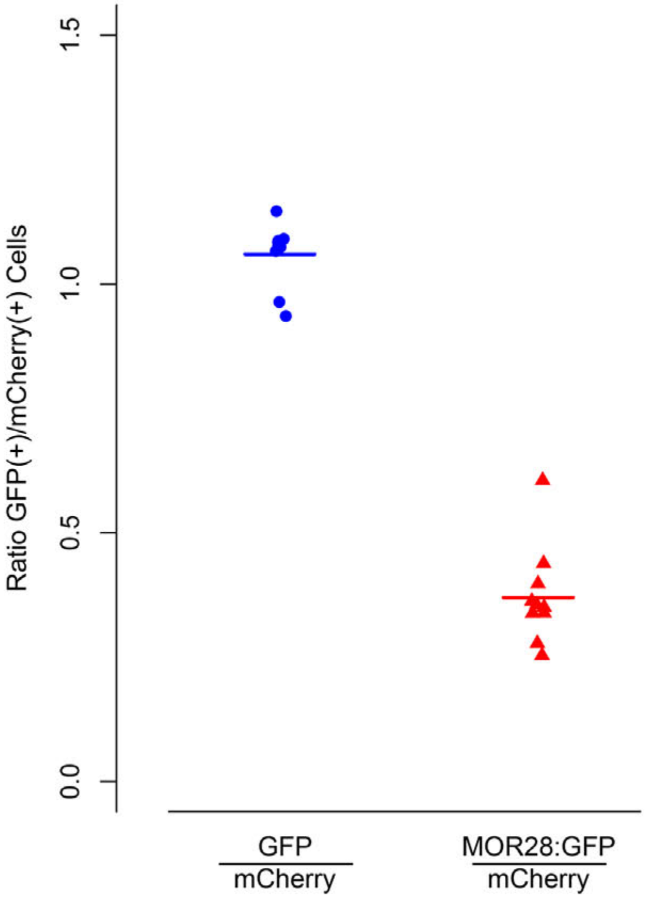
Ratio GFP(+)/mCherry(+) Cells
0.0
0.5
1.0
1.5
GFP
mCherry
MOR28:GFP
mCherry

**Figure S1**

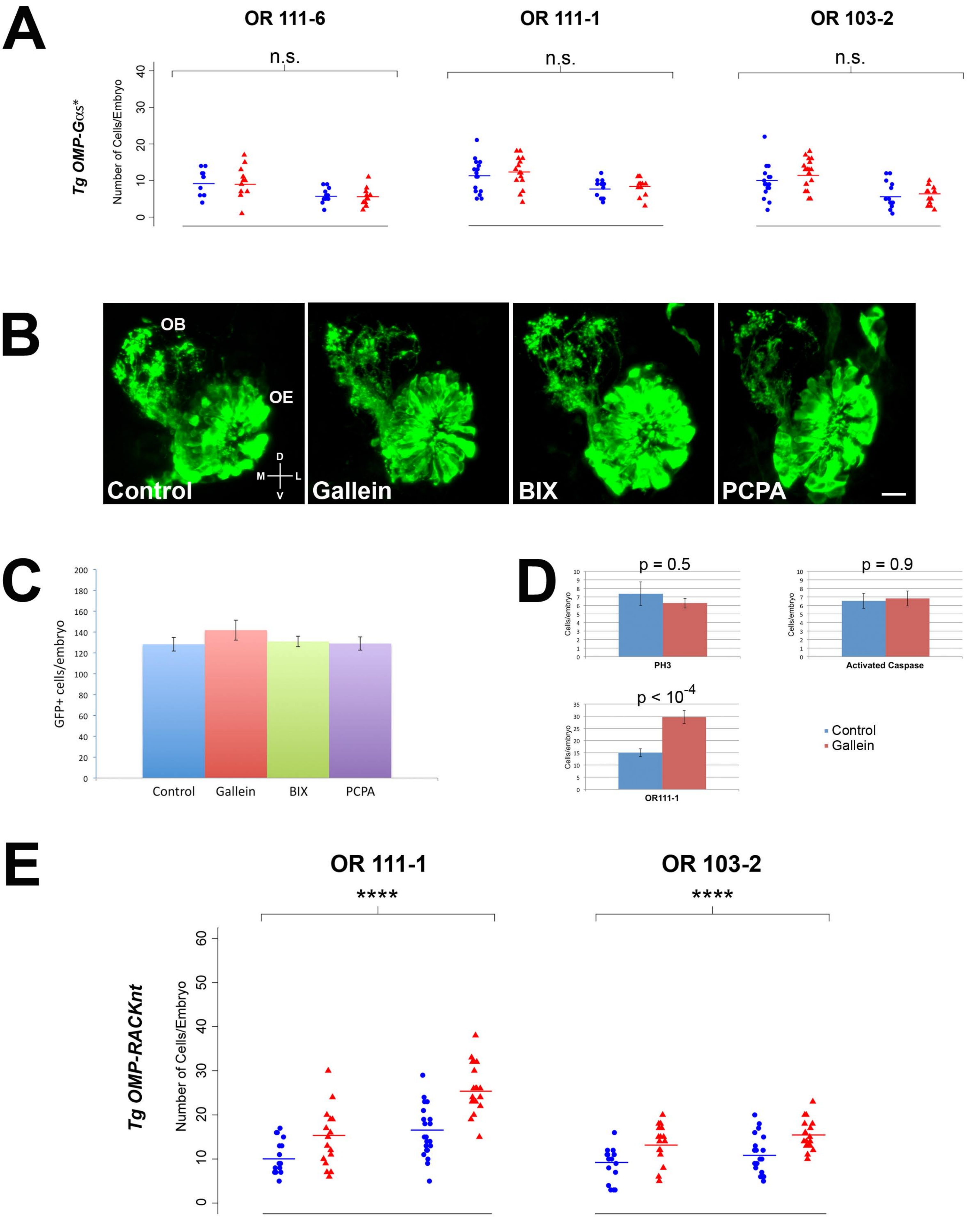
A
OR 111-6
OR 111-1
OR 103-2
n.s.
n.s.
n.s.
Tg OMP-Gαs*
Number of Cells/Embryo
0
10
20
30
40
B
OB
OE
D
M
L
V
Control
Gallein
BIX
PCPA
C
GFP+ cells/embryo
0
20
40
60
80
100
120
140
160
180
200
Control
Gallein
BIX
PCPA
D
p = 0.5
p = 0.9
p < 10^-4
Cells/embryo
PH3
Activated Caspase
OR111-1
0
1
2
3
4
5
6
7
8
9
10
0
5
10
15
20
25
30
35
Control
Gallein
E
OR 111-1
OR 103-2
****
****
Tg OMP-RACKnt
Number of Cells/Embryo
0
10
20
30
40
50
60

Figure S2

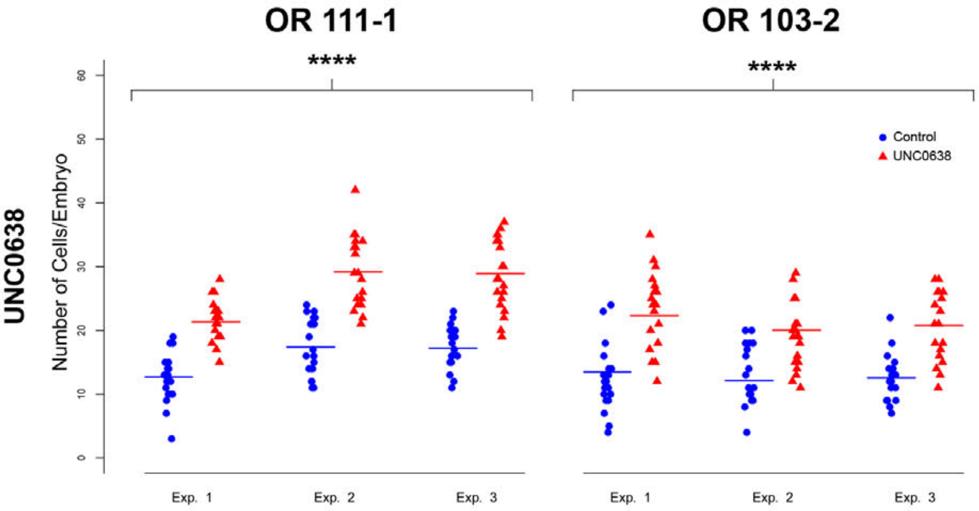
OR 111-1
OR 103-2
****
****
UNC0638
Number of Cells/Embryo
0
10
20
30
40
50
60
● Control
▲ UNC0638
Exp. 1
Exp. 2
Exp. 3
Exp. 1
Exp. 2
Exp. 3

**Figure S3**

## Supplemental Experimental Procedures

**Transgene expression constructs.** DNA sequences were subcloned into a plasmid vector containing Tol2 transposon recombination sites, as previously described (Kwan et al., 2007). Each transgene expression vector contains as its promoter a 1.4 kb fragment upstream of the translational start site of the zebrafish OMP gene (Celik et al., 2002) (accession number NM_173281). Receptor expression vectors were generated by removing the stop codon of each receptor's coding sequence and replacing it with a short linker (5′-GCCACCGCA-3′) followed in-frame by the coding sequence of either eGFP or mCherry, to generate a C-terminal fusion. The receptor:fluorescent protein fusion sequences were then inserted between the OMP promoter and a genomic DNA fragment containing 1 kb genomic sequence downstream of the OR111-1 stop codon, which provides a transcriptional termination/poly adenylation signal. The receptor coding sequences used in the present study were: OR111-1 (accession number DQ306093, nt 1-978), human $\beta_2$AR (accession number NM_000024, nt 240-1478), and MOR28 (accession number NM_001170918, nt 578-1516). A constitutively active $\beta_2$AR NRY mutant (Rasmussen et al., 1999) ($\beta_2$AR*) was generated by site-directed mutagenesis using the following primers: 5′-CCCTGTGCGTGATCGCAGTGAACCGCTACTTTGCCATTACTTCAC-3′ (forward) and 5′-GTGAAGTAATGGCAAAGTAGCGGTTCACTGCGATCACGCACAGGG-3′ (reverse). Control plasmids contained the OMP promoter driving a fusion of the cytoskeleton-associated protein unc76 to eGFP or mCherry (Dynes and Ngai, 1998), linked to 1 kb OR111-1 3′ flanking sequence. Zebrafish G$\beta_1$ (nt 345-1367, accession number NM_212609); G$\gamma_{13}$ (nt 29-232, accession number NM_001166125); G$\alpha$s* (nt 208-1392, accession number NM_201616; the constitutively active variant was generated by introducing a Q227L substitution); the segment encoding the C-terminal 195 amino acids of mouse GRK2 (nt 1496-2083, accession number NM_130863); and the segment encoding the N-terminal 102 amino acids of RACK1 (nt 1-306, accession number NM_131444.1) were each inserted into a Tol2 vector containing the OMP promoter and SV40 polyadenylation sequence.

**Zebrafish.** Zebrafish (Danio rerio, strain AB*) were maintained at 28.5°C and bred naturally using a timed light cycle. For transient transgenesis, plasmid DNA (80ng/$\mu$l) was co-injected with Tol2 messenger RNA (25ng/$\mu$l) in 100mM KCl, 0.1% phenol red into one-cell embryos. Injection volumes were approximately 1/10 the volume of the cell, as judged by visualization of the phenol red dye. Embryos were raised in embryo medium (EM) and transferred at 24 hours post-fertilization (hpf) to EM containing 200 $\mu$M phenylthiourea to prevent pigmentation. For embryos injected with *OMP-GRKct*, *OMP-G*$\beta_1$ + *OMP-G*$\gamma_{13}$, *OMP-G*$\alpha$*s** or *OMP-RACKnt*, the plasmids were co-injected with *OMP-unc76:GFP* to allow the identification of transgenic embryos; control embryos were co-injected with *OMP-unc76:GFP* and an empty vector containing only the OMP promoter (80ng/$\mu$l total DNA concentration). At 3 dpf, embryos were sorted for GFP fluorescence (keeping only embryos in which 50-75% of the olfactory placodes exhibited fluorescent cells), fixed in 4% paraformaldehyde for 2 hr at room temperature, dehydrated through a graded series of increasing methanol concentrations, and stored at -20°C in 100% methanol. For drug treatments, wild type embryos were transferred at 16-20 hpf to EM containing either 100 $\mu$M gallein, 20 $\mu$M BIX, 20 $\mu$M UNC0638, 75 $\mu$M PCPA, or a control solution of EM containing DMSO to match the amount used to deliver drug in the corresponding experimental treatment (typically 0.2%). Embryos were harvested at 3 dpf, fixed, and processed as described above for transgenic fish.

**RNA in situ hybridization and immunohistochemistry.** Whole mount RNA in situ hybridization using digoxigenin and fluorescein-labeled probes and peroxidase detection with FITC-labeled tyramide was performed as described previously (Welten et al., 2006). Simultaneous detection of OR expression by RNA in situ hybridization and GFP by immunohistochemistry was carried out through a modification of the RNA in situ hybridization protocol, using a chicken anti-GFP primary antibody and incubation with an Alexa568-labeled goat anti-chicken secondary antibody following the tyramide detection step. For double label RNA in situ hybridizations, fluorescein-labeled probe was first localized with a peroxidase-conjugated anti-fluorescein-antibody and reacted with Cy3-labeled tyramide. Embryos were washed twice for 30 minutes each in 30% $H_2O_2$ at room temperature to inactivate peroxidase, followed by subsequent washes in phosphate-buffered saline. Embryos were then incubated with peroxidase-conjugated anti-digoxigenin and reacted with FITC-labeled tyramide to detect digoxigenin-labeled probe.  Antisense RNA probes were generated for the following OR sequences, which represent 3 OR gene families located within the same ~100 kb cluster in the zebrafish genome (Alioto and Ngai, 2005; Dugas and Ngai, 2001): OR111-1 (accession number DQ306093, nt 1-981), OR111-6 (accession number DQ306098, nt 1-978), OR103-1 (accession number DQ306104, nt 1-987), OR103-2 (accession number DQ306106, nt 1-945), and OR119-2 (accession number DQ306091, nt 1-945). Templates for these probes were synthesized by PCR, with a T3 RNA polymerase promoter incorporated into the reverse primer to allow probe synthesis using T3 RNA polymerase.

**Microscopy and image analysis.** Fixed embryos were embedded in 1.2% low melting point agarose gel and imaged head-on using a Nikon PCM-2000 scanning laser confocal microscope with argon and helium-neon lasers using a Nikon 20X Fluor (0.75 N.A.)

objective with a 2X digital zoom. Optical sections spanning the entire olfactory placodes of each embryo were acquired at 2 $\mu$m intervals and saved as a stack of high resolution TIFF files. TIFF stacks were imported into NIH ImageJ for analysis. File names were randomized and individual stacks were scored blind. The number of cells positive for a given RNA probe was tabulated for each olfactory placode; cell counts for the left and right olfactory placodes were combined to yield the total number of positive cells per individual fish. Optical sections in ImageJ that showed potential co-localization of GFP signal with OR expression or overlapping expression between two RNA probes were identified and then exported to Adobe Photoshop for verification. The number of cells showing co-localization of GFP and OR expression was tabulated and expressed as the proportion of OR positive cells that are GFP positive. For double label RNA in situ hybridization experiments aimed at detecting instances of OR co-localization, cells were deemed to be co-expressing two ORs if the signals from each of the two OR-specific probes were clearly above background levels and comparable to other strongly positive cells in the field. The proportion of cells exhibiting co-expression of OR111-1 and OR119-2 was defined as follows: proportion = $N_{co} / [N_{OR111\text{-}1} + N_{OR119\text{-}2} - N_{co}]$, where $N_{co}$ = number of cells co-labeled with both receptor probes, $N_{OR111\text{-}1}$ = number of cells labeled with OR111-1, and $N_{OR119\text{-}2}$ = number of cells labeled with OR119-2.

Generalized linear models (GLM) with logarithmic link function for a Poisson distribution (i.e., Poisson or log-linear regression) were used for comparing the number of cells expressing endogenous ORs under treatment and control conditions (McCullagh and Nelder, 1989). Specifically, let $X_{ijk}$ denote the total number of cells expressing a given OR under condition $i$ (treatment or control), in experiment $j$ (3 to 4), for zebrafish embryo $k$ (15 to 20). The model assumes that $X_{ijk}$ has a Poisson distribution with mean exp($\mu + \alpha_i + \beta_j$), where $\mu$ reflects baseline (i.e., control) expression of the OR, $\alpha$ corresponds to the treatment effect, and $\beta$ represents nuisance experiment effects. Identifying transgenes that affect endogenous OR expression corresponds to testing the null hypotheses that the corresponding $\alpha$ parameters are zero. Standard GLM-based *t*-tests were used for this purpose. Goodness-of-fit diagnostics (e.g., residual plots and quantile-quantile plots) indicate that the posited GLM are appropriate for the cell count data. In the case of double label experiments, GLM with logit link function for a binomial distribution (i.e., logistic regression) were used for the proportion of double-positive cells, i.e., the number of cells expressing both the endogenous OR and the GFP-containing transgene divided by the total number of OR-positive cells (McCullagh and Nelder, 1989). All statistical analyses were performed using the R language and environment (http://www.R-project.org) (R_Development_Core_Team, 2010) and, in particular, the glm function.

**ChIP-qPCR.** Olfactory epithelia and liver of 3-6 month old zebrafish adults were dissected and native chromatin was prepared essentially as described (Magklara et al., 2011). The following antibodies were used for immunoprecipitation: H3 dimethyl lysine-9 (Abcam, mAbcam 1220), H3 trimethyl lysine-9 (Abcam, ab8898), and H3 mono-di-trimethyl lysine-4 (Millipore, 04-791). Immunoprecipitated DNA was purified using a MinElute PCR Purification Kit (Qiagen) and amplified using the WGA4 Whole Genome Amplification kit (SIGMA). Duplicate or triplicate aliquot of each amplified reaction was then subjected to quantitative PCR. Enrichment over input is expressed as $2^{-\Delta Ct}$, where Ct = the PCR cycle number at which detection crossed threshold in the qPCR reaction and $\Delta Ct = (Ct_{ChiP} - Ct_{input})$. Primer sequences are as follows:

| | |
|---|---|
| β-actin | 5′-CGAGCTGTCTTCCCATCCA-3′ |
| | 5′-TCACCAACGTAGCTGTCTTTCTG-3′ |
| OR107-1 | 5′-GGGTGAGCTGTAGATTTTAGCTTCAGG-3′ |
| | 5′-GGTGTCCAAGCCTGCTACCCA-3′ |
| OR103-1 | 5′-GCACTACTACCCTGCCAAAAA-3′ |
| | 5′-TGCCAGGATAAAAGCATTGAC-3′ |
| OR103-2 | 5′-CAGTGGGAAACACCAGCTTT-3′ |
| | 5′-TACAGGCTTGTGGAGGCTCT-3′ |
| OR111-1 | 5′-TTGGGCTCACAGGTATAGGG-3′ |
| | 5′-CAGCCTCCGGTCAAGAACTA-3′ |
| OR111-6 | 5′-TTCTTGCCTTTGATCGCTTT-3′ |
| | 5′-AAGGCCACCACAAAGGAGTT-3′ |
| OR101-1 | 5′-GCCGGGGCGACTTGTGTGTC-3′ |
| | 5′-AGCAGGGGTGTCAGGACGGA-3′ |
| OR119-2 | 5′-ATCTGCATTTGCCCTGTTTT-3′ |
| | 5′-AGCAAACAATATGTGCAAACAA-3′ |
| OR108-1 | 5′-TGCATCCTGTAATGAAGGCCTGTGT-3′ |
| | 5′-GCCTTCGCTGAGGTGTGCCT-3′ |
| OMP | 5′-TCCAACGAGCACGTCTACAG-3′ |
| | 5′-AGTGGCGATGATGTTTAGGC-3′ |
| Satellite | 5′-CATGTTAAAGCAAGTTGCAAGTGA-3′ |
| | 5′-CAGCCAGCAGAGAGGTCAAAT-3′ |

**RNA-Seq of FACS-purified olfactory sensory neurons.** For fluorescent labeling and FACS purification of olfactory sensory neurons, a stable *OMP-Gal4* transgenic zebrafish line transgenic line was generated using a plasmid construct containing 1.4 kb *OMP* 5′ promoter sequence (Celik et al., 2002) inserted upstream of the *Gal4* transactivator and crossed with *UAS-GCaMP1.6* transgenic fish (Del Bene et al., 2010). *OMP-Gal4;UAS-GCaMP1.6* transgenic zebrafish embryos were collected at 16-20 hpf and treated with 100 $\mu$M gallein or DMSO (control). Heads of 5 dpf embryos were dissected and dissociated with trypsin and collagenase essentially as described (http://lawsonlab.umassmed.edu/PDFs/dissociationforfacs.pdf). The resulting cell suspension was passed through a 70 $\mu$m mesh, sorted for GFP fluorescence on a Cytopeia INFLUX fluorescence activated cell sorter, collected in 1ml of Trizol and stored at -80$^{o}$C.

Three pairs of FACS-purified gallein-treated and control cells were analyzed by RNA-Seq. Samples comprising each pair were isolated on FACS runs performed on the same day. For each RNA sample, poly(A)+ RNA was enriched from 10-30 ng total RNA using oligo(dT)25 magnetic beads (Life Technologies). cDNA libraries were then prepared using PrepX SPIA RNA-Seq and ILM DNA Library kits (IntegenX, Inc.) according to the manufacturer's protocols. The six libraries were sequenced in two multiplexed runs on an Illumina HiSeq2000 sequencer, yielding approximately 50 million 100 bp paired-end reads per library. Reads were mapped to the Zv9 reference zebrafish genome downloaded from Ensembl v. 67 (Flicek et al., 2012) using TopHat (Trapnell et al., 2009). Gene-level counts were obtained using the htseq-count python script (http://www-huber.embl.de/users/anders/HTSeq/) in the "union" mode and Ensembl v. 67 gene annotation. After verifying that there was no run-specific bias (data not shown), we used the sums of the counts of the two runs as the expression estimates for each library. Genes with fewer than five reads in four or more libraries were discarded, resulting in a total of 20,773 (out of 32,469) expressed genes. Gene-level counts were normalized prior to differential expression analysis by extending the RUV-2 method (Gagnon-Bartsch and Speed, 2012) to RNA-Seq data. This normalization method removes unwanted variation by performing factor analysis for a suitably chosen subset of control genes (for details see Supplemental Statistical Procedures, below).

Differential expression (DE) analysis was carried out within the framework of generalized linear models (GLM), as implemented in the Bioconductor R package edgeR (Robinson et al., 2010). Since there are three treated-control library pairs (corresponding to individual cell sorting runs), we applied the negative binomial log-linear model of edgeR with a treatment main effect as well as pairing effects (see Supplemental Statistical Procedures, below). A distinct dispersion parameter was estimated for each gene, with no "shrinkage" to the common dispersion (i.e., setting prior.df=0 in edgeR's estimateGLMTagwiseDisp() function). A likelihood ratio test of differential expression between treated and control libraries identified 5,094 differentially expressed genes at a false discovery rate (Benjamini and Hochberg, 1995) $< 0.05$. See Table S1 for the list of DE genes.

## References

Alioto, T.S., and Ngai, J. (2005). The odorant receptor repertoire of teleost fish. BMC Genomics *6*, 173.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Series B (Methodological) *57*, 289-300.

Celik, A., Fuss, S.H., and Korsching, S.I. (2002). Selective targeting of zebrafish olfactory receptor neurons by the endogenous OMP promoter. Eur J Neurosci *15*, 798-806.

Del Bene, F., Wyart, C., Robles, E., Tran, A., Looger, L., Scott, E.K., Isacoff, E.Y., and Baier, H. (2010). Filtering of visual information in the tectum by an identified neural circuit. Science *330*, 669-673.

Dugas, J.C., and Ngai, J. (2001). Analysis and characterization of an odorant receptor gene cluster in the zebrafish genome. Genomics *71*, 53-65.

Dynes, J.L., and Ngai, J. (1998). Pathfinding of olfactory neuron axons to stereotyped glomerular targets revealed by dynamic imaging in living zebrafish embryos. Neuron *20*, 1081-1091.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., *et al.* (2012). Ensembl 2012. Nucleic Acids Res *40*, D84-90.

Gagnon-Bartsch, J.A., and Speed, T.P. (2012). Using control genes to correct for unwanted variation in microarray data. Biostatistics *13*, 539-552.

Kwan, K.M., Fujimoto, E., Grabher, C., Mangum, B.D., Hardy, M.E., Campbell, D.S., Parant, J.M., Yost, H.J., Kanki, J.P., and Chien, C.B. (2007). The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs. Dev Dyn *236*, 3088-3099.

Magklara, A., Yen, A., Colquitt, B.M., Clowney, E.J., Allen, W., Markenscoff-Papadimitriou, E., Evans, Z.A., Kheradpour, P., Mountoufaris, G., Carey, C., *et al.* (2011). An epigenetic signature for monoallelic olfactory receptor expression. Cell *145*, 555-570.

McCullagh, P., and Nelder, J.A. (1989). Generalized Linear Models, 2nd edn (Chapman & Hall/CRC Press).

R_Development_Core_Team (2010). R: A Language and Environment for Statistical Computing (Vienna, Austria, R Foundation for Statistical Computing).

Rasmussen, S.G., Jensen, A.D., Liapakis, G., Ghanouni, P., Javitch, J.A., and Gether, U. (1999). Mutation of a highly conserved aspartic acid in the beta2 adrenergic receptor: constitutive activation, structural instability, and conformational rearrangement of transmembrane segment 6. Mol Pharmacol *56*, 175-184.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Welten, M.C., de Haan, S.B., van den Boogert, N., Noordermeer, J.N., Lamers, G.E., Spaink, H.P., Meijer, A.H., and Verbeek, F.J. (2006). ZebraFISH: fluorescent in situ hybridization protocol and three-dimensional imaging of gene expression patterns. Zebrafish *3*, 465-476.

# Supplementary information: Applying RUV-2 to RNA-Seq data

## 1 General idea

Gagnon-Bartsch and Speed (2012) proposed a method for removing batch effects from microarray data coined RUV-2, for Remove Unwanted Variation in 2 steps. Here, we apply the same idea to the normalization of RNA-Seq data in the following way. Consider the log-linear model

$$\log E[Y] = W\alpha + X\beta, \tag{1}$$

where $Y$ is a $n \times G$ matrix containing the observed read counts for the $n$ samples and $G$ genes, $X$ is the $n \times p$ design matrix, and $\beta$ is the $p \times G$ matrix of parameters of interest. Moreover, $W$ is an $n \times k$ matrix containing the $k$ factors of unwanted variation and $\alpha$ is a $k \times G$ matrix with the corresponding nuisance parameters. While the $X$ matrix is fixed and known from the problem at hand, $W$, $\alpha$, and $\beta$ are unknown and need to be estimated. Also, $k$ is unknown, in the sense that the number of factors of unwanted variation is not known *a priori*. Here, rather than estimate $k$, we will fix $k = 1$ in advance.

The simultaneous estimation of $W$, $\alpha$, and $\beta$ in Equation (1) is unfeasible. However, imagine we have a subset of genes for which we know that $\beta = 0$, i.e., for which the expression levels do not change with the variables of interest encoded in the design matrix $X$. We will refer to these non-differentially expressed genes as *negative controls*. For the negative controls, the model in Equation (1) reduces to

$$\log E[Y_C] = W\,\alpha_C, \tag{2}$$

where the subscript $C$ indicates the set of $c$ negative control genes, $Y_C$ is a $n \times c$ matrix containing the observed read counts for the $c$ negative controls and $n$ samples, $W$ is an $n \times k$ matrix containing the $k$ factors of unwanted variation and $\alpha_C$ is a $k \times c$ matrix with the corresponding nuisance parameters for the negative control genes.

In a linear model setting as in Gagnon-Bartsch and Speed (2012), a standard factor analysis can be applied to estimate $W$. Following a similar approach, we estimate $W$ using a Singular Value Decomposition (SVD) of the logarithm of the RNA-Seq counts, which provides us with an approximate solution. Specifically, we consider

$$\log Y_C = U \Lambda V^T$$

where $U$ is a $n \times n$ matrix containing the left singular vectors, $V$ is a $c \times c$ matrix containing the right singular vectors and $\Lambda$ is a $n \times c$ diagonal matrix containing the singular values. We estimate $W$ by $\hat{W} = U\Lambda_k$, where $\Lambda_k$ is a $n \times k$ diagonal matrix with the $k$ largest singular values.

Once we have the estimate $\hat{W}$, we plug it back into Equation (1) and estimate $\alpha$ and $\beta$ using a generalized linear modeling (GLM) approach. If we assume a negative binomial distribution for $Y$, we can use *edgeR* (Robinson et al, 2010) in the usual way to test the null hypothesis $\beta = 0$ (no differential expression) using a likelihood ratio test, considering as design matrix the concatenation of $X$ and $W$.

# 2 Choice of negative controls

For the approach to work in practice, one needs to define a suitable set of negative control genes. Housekeeping genes and/or spike-in standards can be used to this end (see Gagnon-Bartsch and Speed, 2012, for a detailed discussion). Here, we use the following empirical approach: (i) perform a differential expression analysis without any RUV correction; (ii) rank the genes according to their differential expression $p$-values; (iii) select all but the first 1,000 genes with the lowest $p$-values as negative controls; (iv) apply the RUV-2 approach as described in Section 1.

# 3 Application to the data

In our application, we have a paired design, in the sense that there are three treated-control library pairs corresponding to cell-sorting day. This pairing can be accounted for as follows through the design matrix of the generalized linear model. Consider a categorical variable defining the library pairs and one defining the treatment. One strategy for regression on categorical variables (known as factors) is the creation of *dummy variables* (see McCullagh and Nelder (1989) for an introduction to dummy variables in the context of GLMs). In our case, we need two dummy variables for the pairing and one for the treatment. Hence, the design matrix is

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix},$$

where each row represents a sample, the first column is the intercept, columns 2 and 3 are the dummy variables for the pairing, and column 4 is the indicator for the treatment.

For the RUV normalization, the design matrix that we will pass to *edgeR* contains an extra column that represents $W$ from Equation (1), hence

$$(X, W) = \begin{bmatrix} 1 & 0 & 0 & 0 & w_1 \\ 1 & 1 & 0 & 0 & w_2 \\ 1 & 0 & 1 & 0 & w_3 \\ 1 & 0 & 0 & 1 & w_4 \\ 1 & 1 & 0 & 1 & w_5 \\ 1 & 0 & 1 & 1 & w_6 \end{bmatrix}. \tag{3}$$

To perform the likelihood ratio test for differential expression, we compare the full model represented by the design matrix in (3) with the reduced model, defined by the design matrix without the column for the treatment effect (column 4). This can be done in *edgeR* with the functions `glmFit` and `glmLRT`.

# References

Gagnon-Bartsch J, Speed T (2012) Using control genes to correct for unwanted variation in microarray data. Biostatistics 13(3):539–552

McCullagh P, Nelder J (1989) Generalized Linear Models. Chapman and Hall, New York

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140