Predicting Ligand Selectivity of Mammalian Odorant Receptors

by

Yue Jiang

Department of Statistical Science
Duke University

Date:_____
Approved:

_____
Sayan Mukherjee, Supervisor

_____
Li Ma

_____
Hiroaki Matsunami

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Department of
Statistical Science in the Graduate School
of Duke University

2015

ABSTRACT

Predicting Ligand Selectivity of Mammalian Odorant Receptors

by

Yue Jiang

Department of Statistical Science
Duke University

Date:_____
Approved:

_____
Sayan Mukherjee, Supervisor

_____
Li Ma

_____
Hiroaki Matsunami

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the Department of
Statistical Science in the Graduate School of
Duke University

2015

Copyright by
Yue Jiang
2015

Abstract

The mammalian olfactory system uses a large family of odorant receptors to detect and discriminate amongst a myriad of volatile odor molecules. The odorant receptors are similar in protein sequence, but their ligand selectivities dramatically differ. It is not clear how the protein sequences determine the responsiveness of odorant receptors. In this study, I attempt to establish the link between the protein sequences of odorant receptors and their ligand selectivity.

Starting from the response profiles of hundreds of mouse odorant receptors to an odorant generated from my previous work, I used machine learning and variable selection methods to identify properties of amino acid residues that predict receptor response. This leads to protein sequence-based models for odorant receptor response prediction. The models trained with mouse odorant receptor data can predict human odorant receptor responses.

Contents

Abstract iv

List of Tables vii

List of Figures viii

1. Introduction 1

    1.1 Mammalian Odorant Receptors 1

    1.2 'De-orphanization' of ORs 3

    1.3 Functional evolution of ORs 5

    1.4 Links between OR sequence and ligand selectivity 8

2. Data description 10

    2.1 In vivo identification of acetophenone receptors 10

        2.1.1 Mapping of RNA-Seq reads and handling multiple alignment 10

        2.1.2 Differential expression analysis 12

    2.2 In vitro validation for ORs 15

3. Data exploration 21

    3.1 Sequence diversity of acetophenone ORs 22

    3.2 Clustering ORs based on amino acid properties 23

    3.3 Some amino acid residues are more conserved within acetophenone ORs 25

4. Modeling OR responses 28

    4.1 Overall sequence similarity based approach 28

    4.2 Support Vector Machine approach 29

        4.2.1 Linear SVM 29

4.2.2 Kernel method 30

4.3 Semi-Supervised SVM (S3VM) 31

4.4 Logistic regression and variable selection 32

5. External Validation of the model 35

5.1 Overall sequence similarity based approach 36

5.2 SVM and Logistic regression 37

5.3 Conclusion 38

References 40

List of Tables

Table 1: Numbers of OR genes in the genomes of several mammal species as reported in (Niimura et al., 2014). The truncated OR was defined as a partial, intact sequence located at a contig end, and may be either a functional OR or a pseudogene. 3

Table 2: Summary of model performance. 39

List of Figures

Figure 1: Olfaction and ORs. ORs are individually expressed in on olfactory sensory neurons (OSNs) that line the olfactory epithelium. Volatile chemicals in the air activate certain combination of ORs to initiate the perception of smell. From http://www.nobelprize.org 2

Figure 2: Functional evolution and variation of ORs. (A) Expansion of the OR repertoire. During evolution, an ancestral OR gene is duplicated. The gained copy of this OR can become a pseudogene (pseudogenization), or obtain new functions such as responsiveness to a different odor (neofunctionalization). In addition, the two copies of the ancestral OR can each maintain a subset of the original function, each responding to subsets of the ligands of the ancestral OR (subfunctionalization). (B) OR orthologs and paralogs. Orthologs are defined as genes related via a speciation event (for example, human OR 1 and chimpanzee OR 1). Paralogs are defined as genes related via a gene duplication event (for example, mouse OR 1 and mouse OR 2). (C) ORs exhibit high degrees of variation within human populations. A given OR can have several variants with altered protein sequences, and the functions of these variants are sometimes also altered. 7

Figure 3: Distribution of cycle numbers when maximum read counts are achieved. Note that for most of the genes, maximum read counts are reached within the first few cycles. 12

Figure 4: Enrichment of OR transcripts following odor stimulation. Red and blue dots represent ORs enriched in acetophenone-stimulated group (n=3 pairs). A,B: 100% acetophenone; C,D: 1% acetophenone. 14

Figure 5: High concentraion of odor stimulation recruits additional odorant receptors (top left corner). Note the lack of points in the bottom right corner. 15

Figure 6: ORs that are selected for *in vitro* validation. This pool covers almost all the enriched ORs and a large fraction of the ORs not enriched. A known mouse OR that responds to acetophenone, Olfr160 (M72), is also enriched in our experiment..... 16

Figure 7: *In vitro* activation of 71 enriched and 449 un-enriched ORs. Y-axis: normalized fold of increase in luciferase signals. 100% is determined by the fold of increase of Olfr1126 stimulated with 300μM acetophenone. 0% is determined by the fold of increase of empty vector stimulated with 3μM acetophenone. Red: ORs enriched at $p<0.05$. Gray: ORs not called enriched at $p=0.05$. Black bar: median; Box: 25 and 75 percentile; Whisker:

5 and 95 percentile. Wilcoxon rank sum test for difference of *in vitro* responses between enriched and not enriched: 3μM p=0.1, 30μM p=0.005 300μM $p=7\times10^{-21}$. 17

Figure 8: *In vivo* enrichment of 107 activated and 413 not activated ORs. Y-axis: fold of enrichment of transcripts by pS6 IP. Red: ORs activated *in vitro*. Gray: ORs not activated *in vitro*. Black bar: median; Box: 25 and 75 percentile; Whisker: 5 and 95 percentile. Wilcoxon rank sum test for difference of *in vivo* enrichment between activated and not activated: $p=1\times10^{-14}$. 18

Figure 9: *In vivo* and *in vitro* responses predict one another. Left: ROC curves illustrating performance of classifiers using *in vivo* enrichment p-values to predict whether the OR responds to acetophenone *in vitro*. Area Under Curve: 0.754, $p=6\times10^{-16}$, Wilcoxon rank-sum test. Right: ROC curves illustrating performance of classifiers using *in vitro* responses to predict whether the OR is enriched at p<=0.05. Area Under Curve: 0.845, $p=7\times10^{-21}$, Wilcoxon rank-sum test. 19

Figure 10: Identification of acetophenone receptors. (A) Read counts of acetophenone receptors in control and 1% acetophenone stimulated samples. (B) Read counts of acetophenone receptors in control and 100% acetophenone stimulated samples. (C) *In vitro* responses of the acetophenone receptors. Responses are scaled to Olfr160 (M72). The maximum response of Olfr160 is defined as 1. 20

Figure 11: A fraction of aligned ORs. Different colors indicate different amino acids. Gray bars on the bottom panel indicate conservation at each position. 22

Figure 12: Distance tree of OR protein sequences. Red, ORs enriched by 1% acetophenone and confirmed *in vitro*. Orange, ORs additionally enriched by 100% acetophenone and confirmed *in vitro*. 23

Figure 13: Principle component analysis of ORs. The first three PC's are used to separate ORs into four clusters. The acetophenone ORs are mainly present in two of the clusters (cyan). 25

Figure 14: Amino acid residues that are more conserved in the acetophenone receptors than random OR sets of the same size. Red, p<0.01. Orange, p<0.05. 27

Figure 15: ROC of overall similarity based classifier, 10-fold cross-validation. AUC=0.528, p=0.01 29

Figure 16: ROC of the linear SVM classifier, 10-fold cross-validation. AUC=0.798, $p=3\times10^{-112}$, Wilcoxon rank-sum test. 30

Figure 17: ROC of the RBF kernel SVM classifier, 10-fold cross-validation. AUC=0.800, $p=7\times10^{-114}$, Wilcoxon rank-sum test. 31

Figure 18: ROC of the S3VM classifier, 10-fold cross-validation. AUC=0.736, $p=2\times10^{-181}$, Wilcoxon rank-sum test. 32

Figure 19: ROC of logistic regression with 47 predictors selected by elastic-net penalty, 10-fold cross-validation. AUC=0.834, $p=2\times10^{-135}$, Wilcoxon rank-sum test. 33

Figure 20: Distribution of selected residues on a hypothetic OR (magenta), and compared with residues that are conserved across all ORs, regardless of the response profile (green circles). 34

Figure 21: Dose response curves of 27 human ORs to acetophenone, used for external validation of the model. 36

Figure 22: ROC of overall sequence similarity based classifier, external validation on 27 human ORs. AUC=0.660, p=0.1, Wilcoxon rank-sum test. 37

Figure 23: ROC of linear SVM (AUC=0.696, p=0.05), RBF kernel SVM (AUC=0.713, p=0.04), S3VM (AUC=0.684, p=0.06) and logistic regression with 47 predictors selected by elastic-net penalty (AUC=0.728,p=0.03). External validation on 27 human ORs. P-values are from Wilcoxon rank-sum test. 38

1. Introduction

In mammals, olfactory sensation starts with the detection of odor ligands by odorant receptors (ORs), a family of seven transmembrane G protein-coupled receptors (GPCRs) (Buck and Axel, 1991). Mammalian genomes encode a large number of ORs evolved through numerous gene duplication and loss events (Niimura et al., 2014). To date, it is not clear how the sequence of ORs determine their ligand selectivity.

1.1 Mammalian Odorant Receptors

ORs are individually expressed in the olfactory sensory neurons (OSNs) located in the olfactory epithelium (OE) (Figure 1). The mouse genome encodes over 1000 intact ORs, although it is unclear how many are functionally expressed in the OSNs (Healy et al., 1997; Niimura and Nei, 2005; Zhang and Firestein, 2002). Odor recognition follows a combinatorial coding scheme, where one OR can be activated by a set of odors, and one odor can activate a combination of ORs (Buck, 2004; Malnic et al., 1999; Nara et al., 2011; Saito et al., 2009). Through such combinatorial coding, mammals can detect and discriminate a large number of odors (Bushdid et al., 2014).

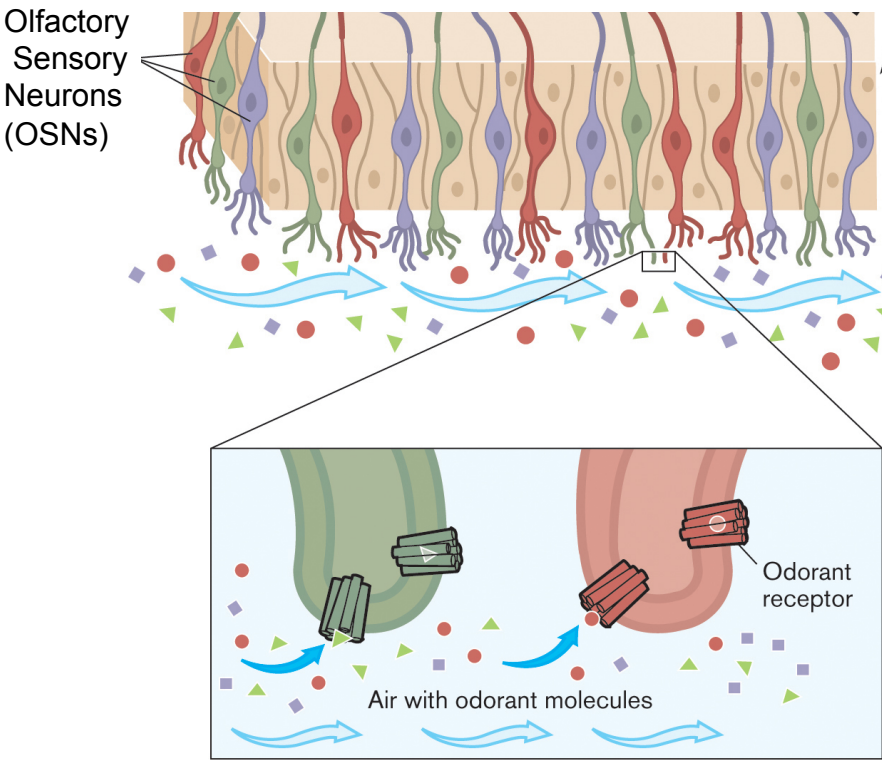

Figure 1: Olfaction and ORs. ORs are individually expressed in on olfactory sensory neurons (OSNs) that line the olfactory epithelium. Volatile chemicals in the air activate certain combination of ORs to initiate the perception of smell. From http://www.nobelprize.org

Mammalian genomes typically encode large numbers of ORs (Table 1), with approximately 400 intact ORs in human, more than 1000 in mouse and close to 2000 in African elephant (Godfrey et al., 2004; Malnic et al., 2004; Niimura et al., 2014). An exception is found in cetaceans (whales and dolphins), which only have small numbers of intact ORs (McGowen et al., 2008). The ORs are undergoing rapid evolutionary changes including extensive gene gains and losses, as seen by a high percentage of OR pseudogenes in mammalian genomes (Gilad et al., 2003). ORs also exhibit high genetic variability within the human population (Menashe et al., 2003).

Table 1: Numbers of OR genes in the genomes of several mammal species as reported in (Niimura et al., 2014). The truncated OR was defined as a partial, intact sequence located at a contig end, and may be either a functional OR or a pseudogene.

Species	Intact ORs	OR pseudogenes	Truncated ORs
Human	396	425	0
Chimpanzee	380	414	19
Mouse	1130	236	0
Rat	1207	508	52
Dog	811	278	11
Elephant	1948	2230	89

1.2 'De-orphanization' of ORs

Most of the mammalian ORs are 'orphan', meaning their cognate ligands are still unknown. Deciphering the coding of olfactory information requires the comprehensive identification of ORs that respond to a given odor (Peterlin et al., 2014). Various *in vivo, ex vivo,* and *in vitro* methods, such as virus-mediated OR overexpression (Touhara et al., 1999; Zhao et al., 1998), calcium imaging of dissociated OSNs combined with single-cell RT-PCR (Nara et al., 2011; Touhara et al., 1999), use of fluorescent-labeled transgenic mice (Bozza et al., 2002), and *in vitro* expression (Saito et al., 2004) have been used to match ORs and their cognate ligands. Notably, molecular receptive range analyses of a few ORs have revealed diverse odor tuning properties among the tested ORs (Araneda et al., 2000; Grosmaitre et al., 2009; Li et al., 2012; Reisert and Restrepo, 2009).

Currently, large-scale identification of active ligands for mammalian ORs relies on *in vitro* heterologous cell systems (Mainland et al., 2014; Saito et al., 2009; Saito et al.,

2004). Though many studies have shown that *in vitro* responses predict OSN activation in mice and behavior in humans (Dong et al., 2014; Keller et al., 2007; Mainland et al., 2014; Menashe et al., 2007; Shirasu et al., 2014), the lack of high-throughput *in vivo* mapping methods makes it difficult to estimate the correspondence between *in vitro* and *in vivo* results across a large number of ORs activated by a given odor.

Recently I developed a high-throughput method to map odorant responding ORs *in vivo,* using phosphorylated ribosome immunoprecipitation and RNA-Seq. The S6 ribosomal subunit is phosphorylated following neuronal activation (Cao et al., 2008; Knight et al., 2012; Valjent et al., 2011; Villanueva et al., 2009; Zeng et al., 2009). This phosphorylation is comparable to induction of immediate early gene expression, such as c-Fos and Egr-1, which is widely used to mark active neurons (Haga et al., 2010; Isogai et al., 2011; Shirasu et al., 2014). However, unlike immediate early gene expression, phospho-S6 has the advantage of being physically associated with mRNA species expressed in the activated neurons. Thus, phosphorylated S6 immunoprecipitation (pS6-IP) followed by mRNA profiling with RNA-Seq has been developed as a method to identify mRNAs expressed in activated neurons (Knight et al., 2012). This approach has been successfully applied in brain regions such as the hypothalamus in identifying markers of neurons that respond to feeding, starvation and high salt (Knight et al., 2012).

In my graduate work, I established a method using pS6-IP and RNA-Seq in the olfactory system to map odor-activated ORs. In the OE, each OSN chooses to express

one OR allele out of over 1000 possible OR genes. Thus ORs associated with activated OSNs are likely responding to the odor.

1.3 Functional evolution of ORs

Central to understanding the functional evolution of ORs is functional comparison within orthologous ORs (ORs evolved from a common ancestral OR by speciation) as well as within paralogous ORs (related ORs by duplication within a genome). It is usually assumed that orthologs perform equivalent functions, while paralogs may be more diverged in function (Koonin, 2005), but the extent to which this assertion holds for ORs needs to be tested. For instance, the human OR OR1A1 and its mouse ortholog Olfr43 show similar responses to (S)-(−)-citronellol (Schmiedeberg et al., 2007), supporting the notion that orthologous ORs are equivalent in function. However, the mouse and rat I7 orthologs show different preferences for heptanal and octanal, as a result of a single amino acid change (Krautwurst et al., 1998). In addition, in a study involving orthologs of Olfr154 (OR912-93) across multiple species, the ligand selectivity of orthologs was found to be similar but not the same: the Olfr154 orthologs of mouse, pig and four primates all responded to both 2- and 3-heptanone, the squirrel-monkey ortholog only responded to 3-heptanone, and the human and orangutan orthologs were not functional (Gaillard et al., 2004). When quantitative characteristics of OR responses were taken into account, primate orthologs of the human OR OR7D4 exhibited extremely diverse potency and efficacy to two steroid ligands, androstenone and

androstadienone (Zhuang et al., 2009). Finally, a large scale analysis of 18 OR orthologs in the human, chimpanzee and/or rhesus macaque and of 17 in mouse and rat against a panel of 42 ORs revealed that a striking ~90% of orthologous ORs showed functional variances affecting receptor potency and/or efficacy (Figure 2 A, B). When only considering binary responses, orthologous ORs respond to a common ligand 83% of the time, while paralogs respond to a common ligand only 33% of the time (Adipietro et al., 2012). These studies suggest that the ligand selectivity of OR orthologs are more conserved than paralogs, while the response ranges among orthologs can dramatically differ. In the future, it will be important to evaluate how the functional differences in OR orthologs and paralogs contribute to the adaptation of different species.

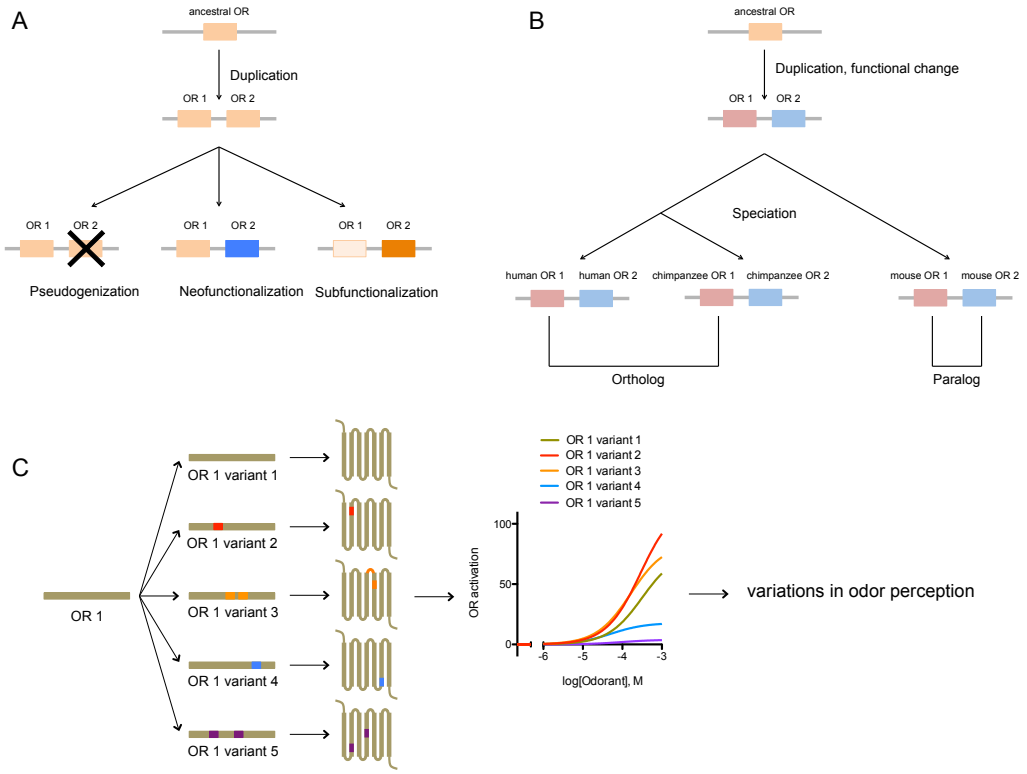

Figure 2: Functional evolution and variation of ORs. (A) Expansion of the OR repertoire. During evolution, an ancestral OR gene is duplicated. The gained copy of this OR can become a pseudogene (pseudogenization), or obtain new functions such as responsiveness to a different odor (neofunctionalization). In addition, the two copies of the ancestral OR can each maintain a subset of the original function, each responding to subsets of the ligands of the ancestral OR (subfunctionalization). (B) OR orthologs and paralogs. Orthologs are defined as genes related via a speciation event (for example, human OR 1 and chimpanzee OR 1). Paralogs are defined as genes related via a gene duplication event (for example, mouse OR 1 and mouse OR 2). (C) ORs exhibit high degrees of variation within human populations. A given OR can have several variants with altered protein sequences, and the functions of these variants are sometimes also altered.

The OR repertoire is highly variable across the human population, encompassing large numbers of single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) (Gilad and Lancet, 2003; Keydar et al., 2013; Olender et al., 2012; Waszak et al., 2010). A human OR has a median of five different alleles with an allele frequency greater

than 1% in the 1000 Genomes Project data (Mainland et al., 2014). Recently, a functional analysis of variants of 27 human ORs in the heterologous expression system revealed that 63% of these ORs have variations that alter *in vitro* functions, and two individuals have an average of over 30% functional differences in their OR alleles (Mainland et al., 2014). Among the variants that alter OR protein sequences, 7.9% lead to nonsense mutations that would disrupt ORs, 6.8% are hypofunctional (lower sensitivity and dynamic range) and 11% are hyperfunctional (higher sensitivity and dynamic range) as compared to reference alleles (Figure 2 C) (Mainland et al., 2014). In this study, only SNPs in the open reading frames of ORs were considered. Additional functional differences could result from variations in the regulatory regions that control OR expression, as well as CNVs that cause gains and losses of certain ORs. Recent transcriptome analysis is starting to reveal differences in OR expression levels among individuals (Verbeurgt et al., 2014). CNVs are also often observed (Hasin et al., 2008), with reported CNVs for ~33% of the human ORs, and an average of 8 CNVs of OR loci per individual (Waszak et al., 2010).

1.4 Links between OR sequence and ligand selectivity

To date, very little is know how OR sequences determine their ligand selectivity, except that OR orthologs share more common ligands than paralogs. Based on this idea, Man et al. proposed 22 amino acid residues important for ligand selectivity (Man et al., 2004), but these residues remain largely untested experimentally. A study performed on

Drosophila odorant receptors showed that sequence similarity do not correlate strongly with ligand specificity (Boyle et al., 2013). For mammalian ORs, it remains largely unclear if the OR sequences contain any information that can be learned to predict their function, partly due to the limited availability of OR response profiles. A better understanding on the relationship between OR sequence and their ligand selectivity will guide the de-orphanization of ORs, and may allow us to design novel ORs as custom chemical sensors. Here I attempt to model OR responses based on their sequences.

2. Data description

I systematically analyzed the response profile of mouse ORs to an odorant, acetophenone, using both *in vivo* and *in vitro* methods. This leads to the largest dataset of OR response profiles to a given chemical to date. 52 ORs responding to acetophenone both *in vivo* and *in vitro* are considered as acetophenone ORs. 367 ORs do not respond to acetophenone under either condition, thus are considered not acetophenone receptors.

2.1 In vivo identification of acetophenone receptors

Transcriptomes of olfactory neurons responding to acetophenone were profiled by phosphorylated ribosome IP and RNA-Seq. Samples from unstimulated animals were used as controls. Experiments were performed in triplicates, and within each replication the experimental and control mice were littermates of the same sex.

2.1.1 Mapping of RNA-Seq reads and handling multiple alignment

Short reads were aligned to the mouse reference genome mm10 using Bowtie (Langmead et al., 2009). The reads mapped to annotated genes were then counted using BEDTools (Quinlan and Hall, 2010). Because UTRs of many ORs are not well defined, we only used the coordinates of coding exons for all the genes to count the reads mapped to them.

The similarity between ORs can result in multiple alignments. The traditional stringent method is to suppress multiple alignments, discarding all the reads that map to more than one genomic locus. This approach apparently does not make full use of the

data. The traditional unstringent method allows for multiple alignments, and randomly assigns the read to one of the loci it maps to. This approach introduces some uncertainty in the result. To make the best use of the sequence data and obtain exact estimates of expression levels, I estimated the distribution of reads mapped to multiple locations based on unique mapping results. For example, if a read was mapped to two different genes, and the read counts of these two genes were 3 and 7 respectively based on uniquely mapped read counts, then 0.3 and 0.7 were added to the previous read counts and the new counts became 3.3 and 7.7. Based on this idea, I implemented a stepwise mapping scheme that allows reads aligned to multiple regions to be distributed proportionally to these regions based on the previous read counts from these regions (Wang et al., 2008). Briefly, the reads were aligned to mm10 by Bowtie with -m 20 -a (report all reads that map to at most 20 different regions on the reference genome), and the resulted mapping file was separated into 20 sub-files that includes reads only mapped to 1, 2, … , 20 regions respectively. BEDTools (Quinlan and Hall, 2010) was used to first count reads mapped to coding exons of genes using only uniquely mapped reads. Reads mapped to two regions were distributed based on read counts from uniquely mapped reads. The iteration was allowed for 20 cycles and the resulted read counts were then rounded. Iteration number 20 was chosen because pilot parameter tuning indicated that the majority of read counts stopped increasing after 20 cycles (Figure 3).

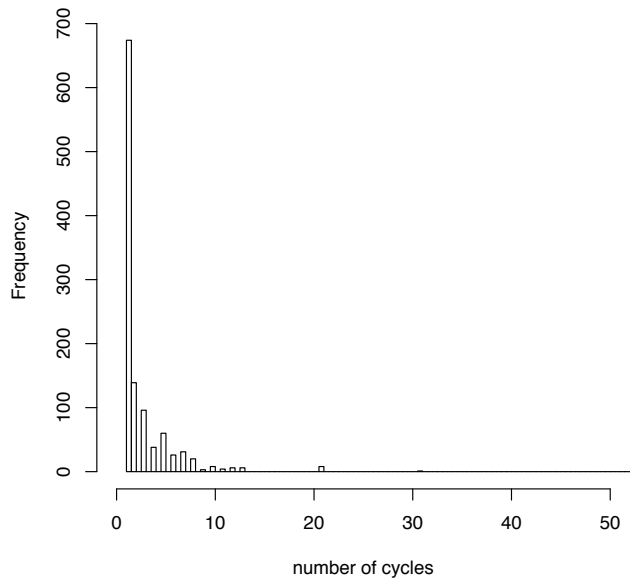

Figure 3: Distribution of cycle numbers when maximum read counts are achieved. Note that for most of the genes, maximum read counts are reached within the first few cycles.

2.1.2 Differential expression analysis

The read counts table for all the genes were then imported into R and analyzed using the negative-binomial model based tool EdgeR under default parameter settings (Robinson et al., 2010). A similar result was obtained using a separate tool DESeq2 employing similar model assumptions (Anders and Huber, 2010). Littermate information was included in the design matrix to enhance the power of detection (McCarthy et al., 2012). The generalized linear model for EdgeR is

$$log\mu_{gi} = x_i^T \beta_g + logN_i$$

where μ_{gi} is the mean of mapped read numbers in library i of gene g, and N_i is the total number of mapped reads in library i. β_g is the vector of coefficients and x_i denote the vector of covariates that specify the treatment of sample i.

Multiple comparison was adjusted within the detected OR group by controlling false discovery rate (FDR) (Benjamini and Hochberg, 1995). DESeq was used to calculate the size factors of individual libraries (Anders and Huber, 2010). Experiments were performed with both high (100%) and low (1%) concentrations of acetophenone (Figure 4).

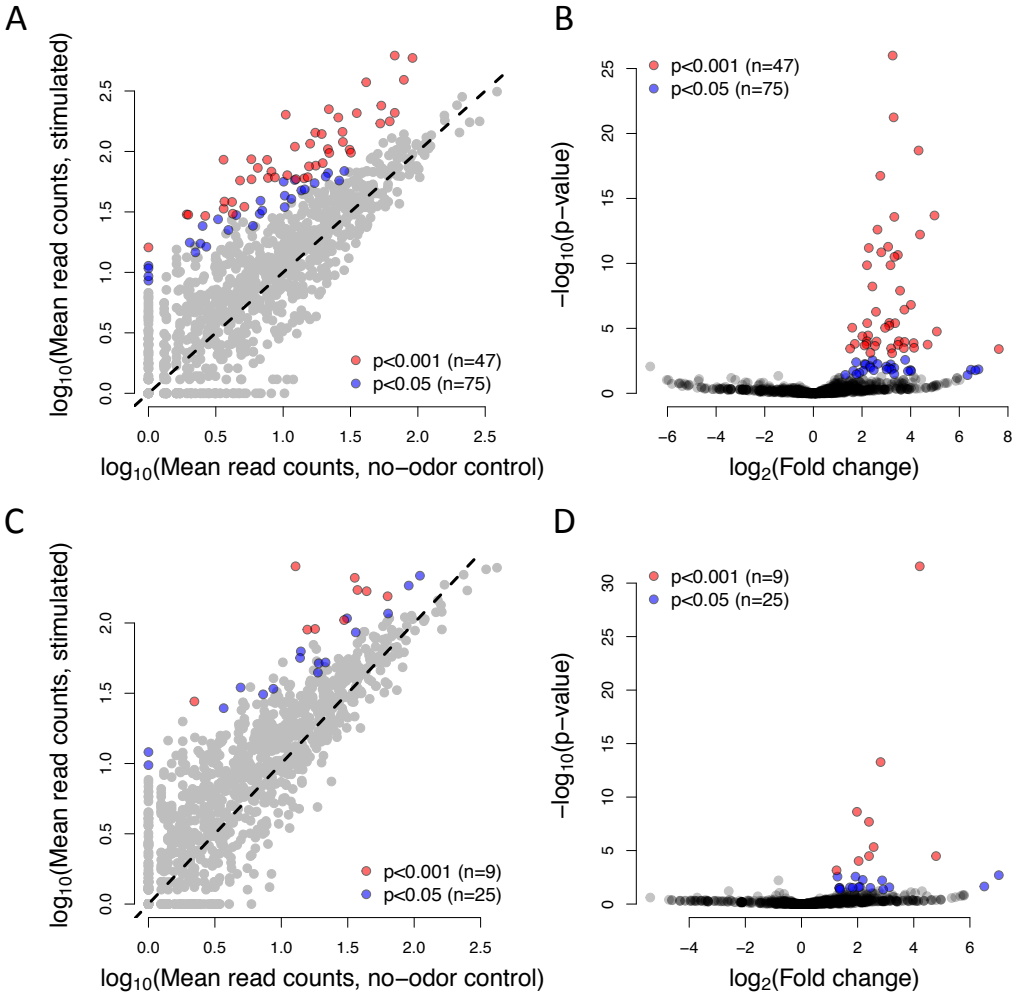

Figure 4: Enrichment of OR transcripts following odor stimulation. Red and blue dots represent ORs enriched in acetophenone-stimulated group (n=3 pairs). A,B: 100% acetophenone; C,D: 1% acetophenone.

In agreement with previous notions, a high concentration of odor stimulation activated additional ORs (Figure 5).

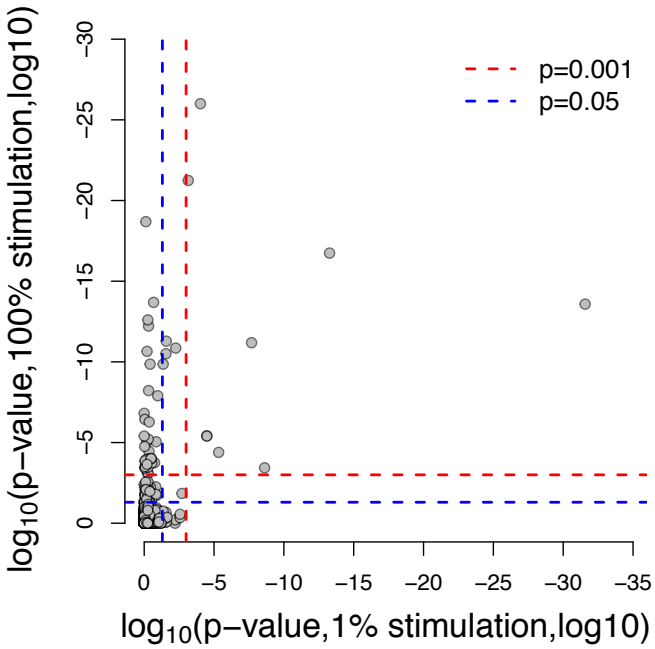

Figure 5: High concentraion of odor stimulation recruits additional odorant receptors (top left corner). Note the lack of points in the bottom right corner.

2.2 In vitro validation for ORs

To validate that the *in vivo* receptor mapping identified acetophenone-activated ORs, we leveraged the previously established cAMP-mediated luciferase reporter gene assay in heterologous cells (Li and Matsunami, 2011; Zhuang and Matsunami, 2008) to measure the response of these ORs to acetophenone. To systematically test whether the enriched ORs were more likely to respond to acetophenone, we tested 71 ORs enriched by 100% acetophenone (out of 75, $p<0.05$ with FDR correction) and 449 control ORs that were not enriched (Figure 6).

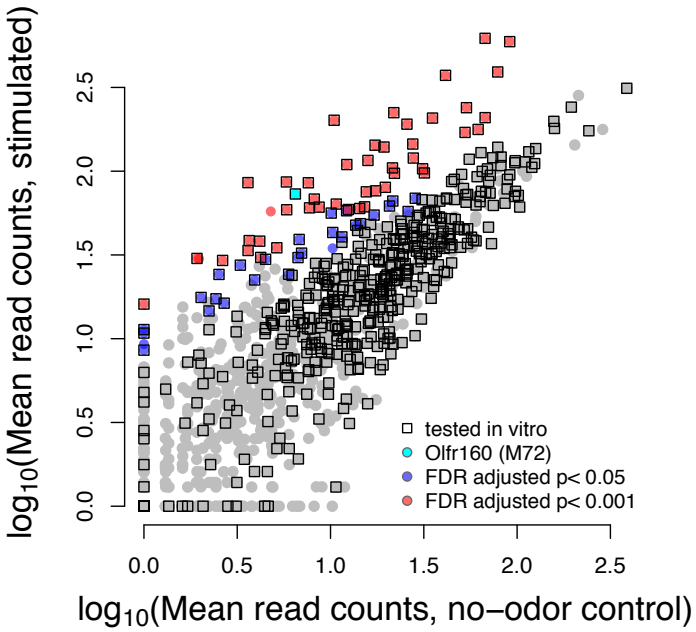

Figure 6: ORs that are selected for *in vitro* validation. This pool covers almost all the enriched ORs and a large fraction of the ORs not enriched. A known mouse OR that responds to acetophenone, Olfr160 (M72), is also enriched in our experiment.

We independently expressed these ORs in Hana 3A cells, which are a HEK293T-derived cell line that supports the robust expression of various transiently expressed ORs (Saito et al., 2004; Zhuang and Matsunami, 2008), and stimulated the cells with 3μM, 30μM, or 300μM of acetophenone as well as no-odor controls. We quantified the degree of OR activation by determining the relative fold-increase of luciferase activity as compared to control stimulation. As expected, acetophenone-induced *in vitro* activation of enriched ORs was higher than that of control ORs ($p=0.005$ at 30μM, $p=7\times10^{-21}$ at 300μM, Wilcoxon rank-sum test, Figure 7).

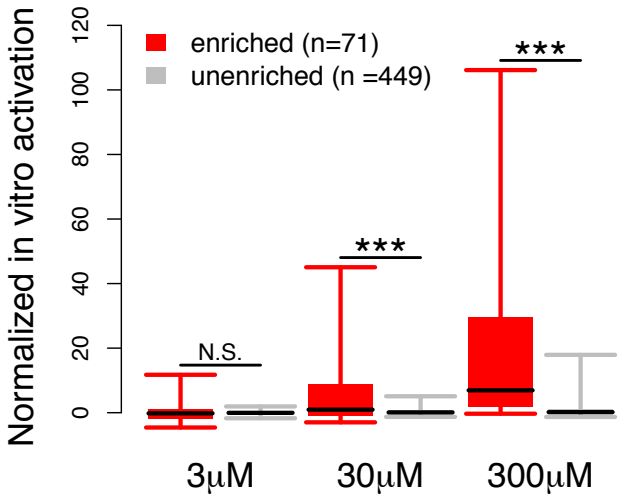

Figure 7: *In vitro* activation of 71 enriched and 449 un-enriched ORs. Y-axis: normalized fold of increase in luciferase signals. 100% is determined by the fold of increase of Olfr1126 stimulated with 300μM acetophenone. 0% is determined by the fold of increase of empty vector stimulated with 3μM acetophenone. Red: ORs enriched at $p<0.05$. Gray: ORs not called enriched at $p=0.05$. Black bar: median; Box: 25 and 75 percentile; Whisker: 5 and 95 percentile. Wilcoxon rank sum test for difference of *in vitro* responses between enriched and not enriched: 3μM $p=0.1$, 30μM $p=0.005$ 300μM $p=7\times10^{-21}$.

Similarly, the ORs that were activated by acetophenone *in vitro* were enriched in the pS6-IP analysis as a group ($p=1\times10^{-14}$, Wilcoxon rank-sum test, Figure 8).

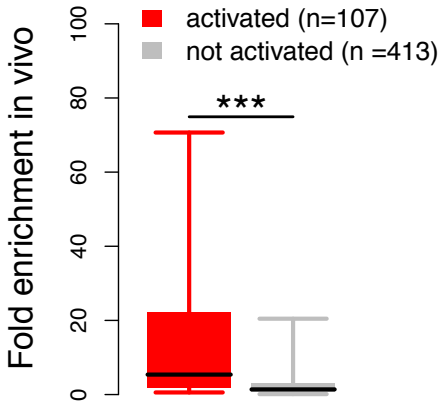

Figure 8: *In vivo* enrichment of 107 activated and 413 not activated ORs. Y-axis: fold of enrichment of transcripts by pS6 IP. Red: ORs activated *in vitro*. Gray: ORs not activated *in vitro*. Black bar: median; Box: 25 and 75 percentile; Whisker: 5 and 95 percentile. Wilcoxon rank sum test for difference of *in vivo* enrichment between activated and not activated: $p=1\times10^{-14}$.

Enrichment in pS6-IP RNA-Seq predicted whether the OR responded to acetophenone *in vitro* (area under the receiver operating characteristic (ROC) curve (AUC) = 0.754, $p=6\times10^{-16}$, Wilcoxon rank-sum test). *In vitro* response also predicted the p-value of enrichment in pS6-IP RNA-Seq (AUC = 0.845, $p=7\times10^{-21}$, Wilcoxon rank-sum test, Figure 9).

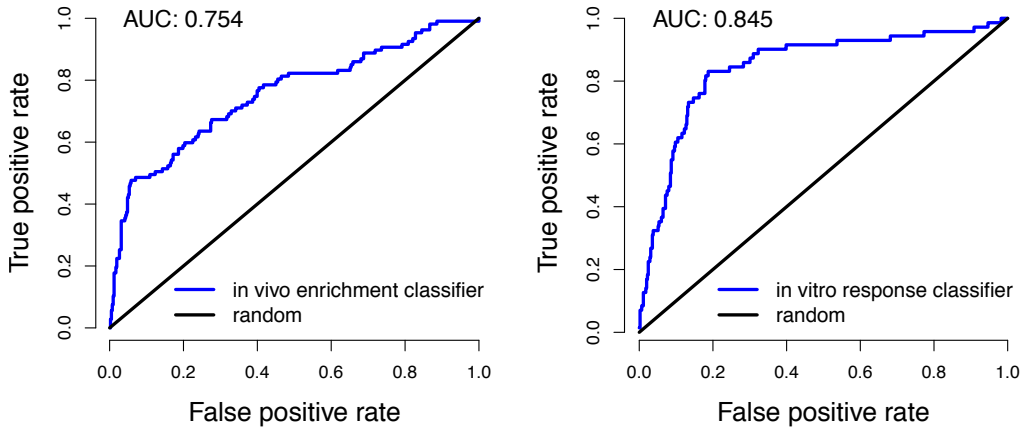

Figure 9: *In vivo* and *in vitro* responses predict one another. Left: ROC curves illustrating performance of classifiers using *in vivo* enrichment p-values to predict whether the OR responds to acetophenone *in vitro*. Area Under Curve: 0.754, $p=6\times10^{-16}$, Wilcoxon rank-sum test. Right: ROC curves illustrating performance of classifiers using *in vitro* responses to predict whether the OR is enriched at p<=0.05. Area Under Curve: 0.845, $p=7\times10^{-21}$, Wilcoxon rank-sum test.

After establishing the concordance between in vivo and in vitro responses, I set out to further investigate the acetophenone-activated ORs. I selected ORs that were enriched by S6 phosphorylation analysis with acetophenone, expressed these ORs in Hana3A cells (Saito et al., 2004) and challenged the cells with 10^{-9} - 10^{-3} M acetophenone. The dose-dependent luciferase response profiles confirmed that 52 ORs were indeed activated by acetophenone. I reason that these receptors, supported by both in vivo and in vitro evidence, are likely to be bona fide acetophenone receptors (Figure 10).

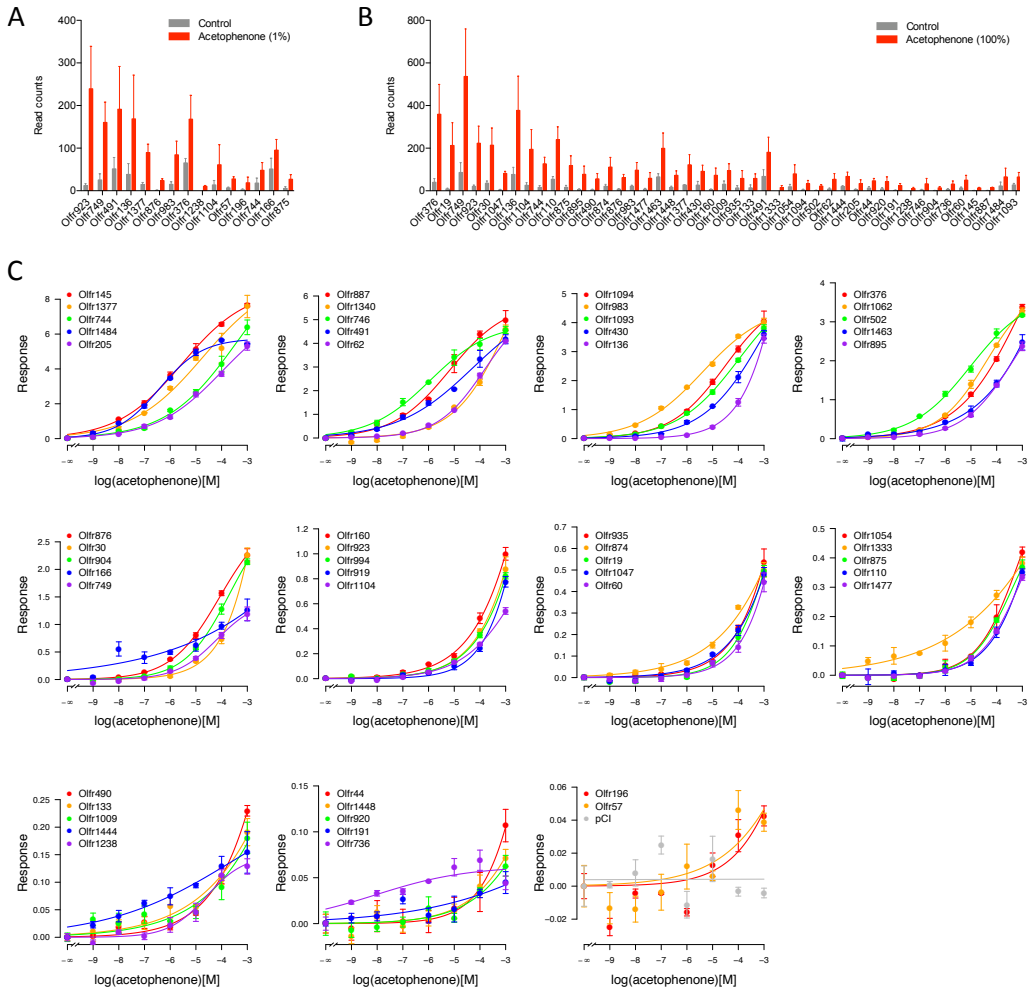

Figure 10: Identification of acetophenone receptors. (A) Read counts of acetophenone receptors in control and 1% acetophenone stimulated samples. (B) Read counts of acetophenone receptors in control and 100% acetophenone stimulated samples. (C) *In vitro* responses of the acetophenone receptors. Responses are scaled to Olfr160 (M72). The maximum response of Olfr160 is defined as 1.

3. Data exploration

The collection of OR response profiles to acetophenone provides the first opportunity to systematically study what determines the responsiveness of ORs to a chemical. Previous studies have found that the overall sequence similarity is a poor predictor for receptor response within species. OR paralogs only share the common ligand 33% of the time. On the other hand, OR orthologs share the common ligand 83% of the time. Indeed, the assertion that orthologs tend to share the same function is the basis of the use of model organisms in biomedical research. So far, it is also the only way to infer the response of an OR from one species based on the OR from another species. Given this big collection of OR response profiles, it might be possible to construct better models for OR response prediction that works for both paralogs and orthologs, and outperform the overall sequence similarity based approach.

ORs share very similar protein sequences and predicted transmembrane structures. They are seven-transmembrane G protein-couple receptors. The transmembrane domain regions are well conserved. The domains implicated in down-stream signaling are also well conserved. This provides the opportunity to align these receptors together, and hypothesize that receptor functions may be determined by specific amino acid residues at certain positions (Figure 11).

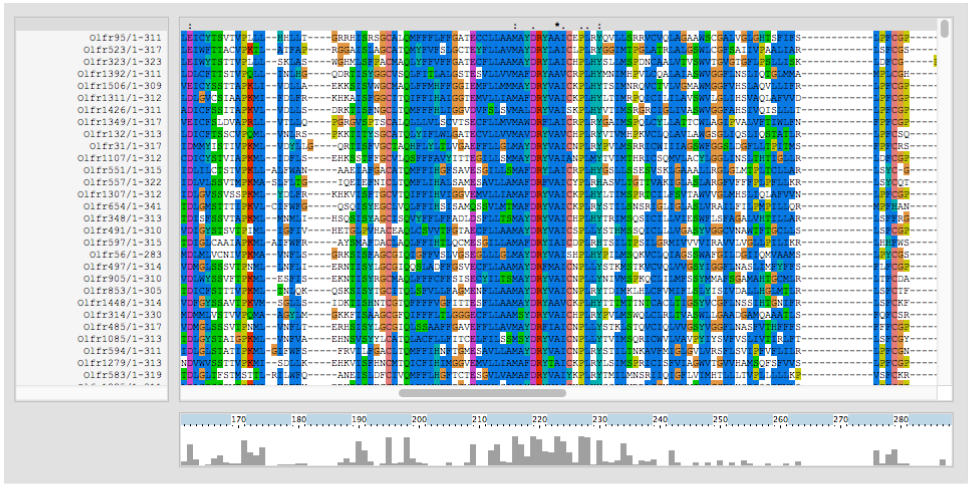

Figure 11: A fraction of aligned ORs. Different colors indicate different amino acids. Gray bars on the bottom panel indicate conservation at each position.

3.1 Sequence diversity of acetophenone ORs

I first asked if the acetophenone ORs have greater overall homology with each other. When plotted on the OR phylogenetic tree, the acetophenone ORs were not clustered in one or a few subfamilies, though there were a few notable examples of closely related ORs. Rather, their sequences were strikingly diverse: at least one member of 27 out of the 286 receptor families defined by Zhang and Firestein (Zhang and Firestein, 2002) were represented (Figure 12).

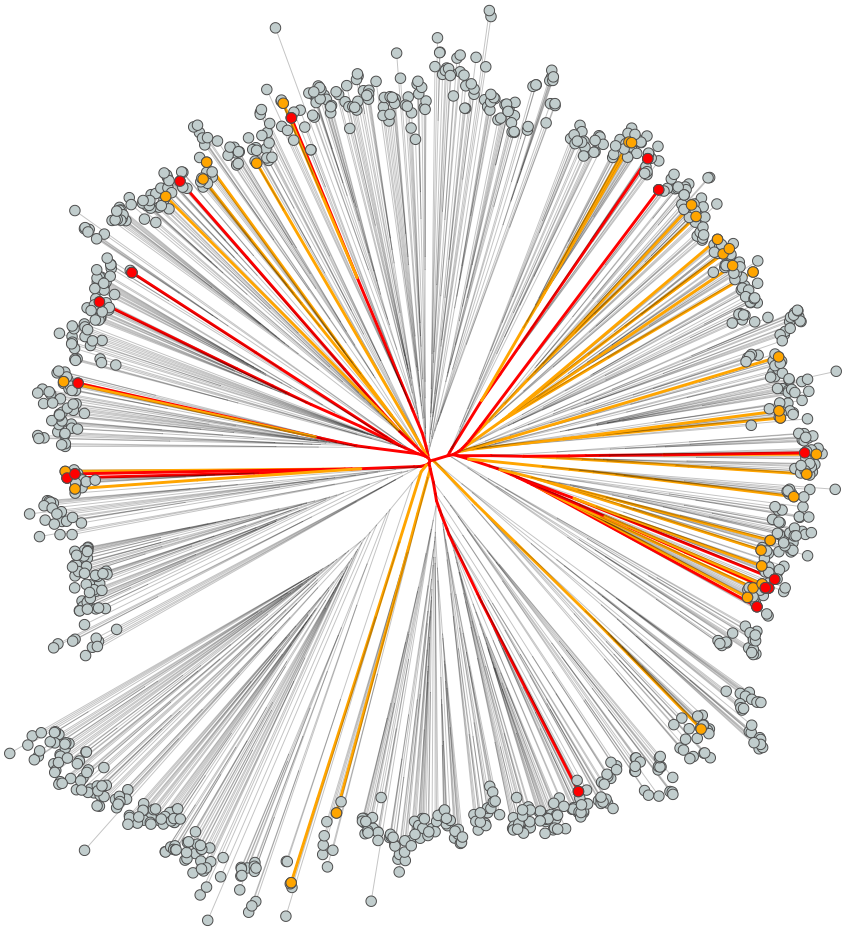

Figure 12: Distance tree of OR protein sequences. Red, ORs enriched by 1% acetophenone and confirmed *in vitro*. Orange, ORs additionally enriched by 100% acetophenone and confirmed *in vitro*.

3.2 Clustering ORs based on amino acid properties

I then asked if ORs could be clustered based on their sequences to reflect their functions. Grantham (Grantham, 1974) defined three quantitative features for amino acid residues, including polarity, composition and volume. Composition is defined as

the atomic weight ratio of hetero (noncarbon) elements in end groups or rings to carbons in the side chain, reflecting composition differences between amino acids. Polarity reflects the hydrophobicity of the residue. Volume reflects size of the residue. Using the multiple alignment of mouse OR proteins, I calculated these three properties as defined by Grantham for amino acid residues within the alignment. 291 amino acid residues common to at least 10% of the ORs are included in this analysis, resulting in $3 \times 291 = 873$ predictors. The missing values in this matrix (0.17%) resulting from occasional gaps in alignment were imputed from column means. To construct binary response variables, I used ORs that were supported by both *in vivo* and *in vitro* evidence, with the 52 ORs that were enriched in our pS6 RNA-Seq and responded to acetophenone *in vitro* labeled as 1's, and 367 ORs that were not enriched in pS6 RNA-Seq ($p>0.05$ in both 1% and 100% acetophenone stimulation experiments) and did not respond to acetophenone *in vitro* labeled as 0's.

I first performed a principle component analysis (PCA) on the amino acid property data to see if it can cluster ORs into groups that reflect their responses to acetophenone. The first three PC's (percentage of variance explained 6.8%, 3.7%, 2.6% respectively) cluster the ORs into four groups, and the acetophenone responding ORs are mainly present in two of them (Figure 13). This suggests that a combination of amino acid properties might contain information related to OR function.

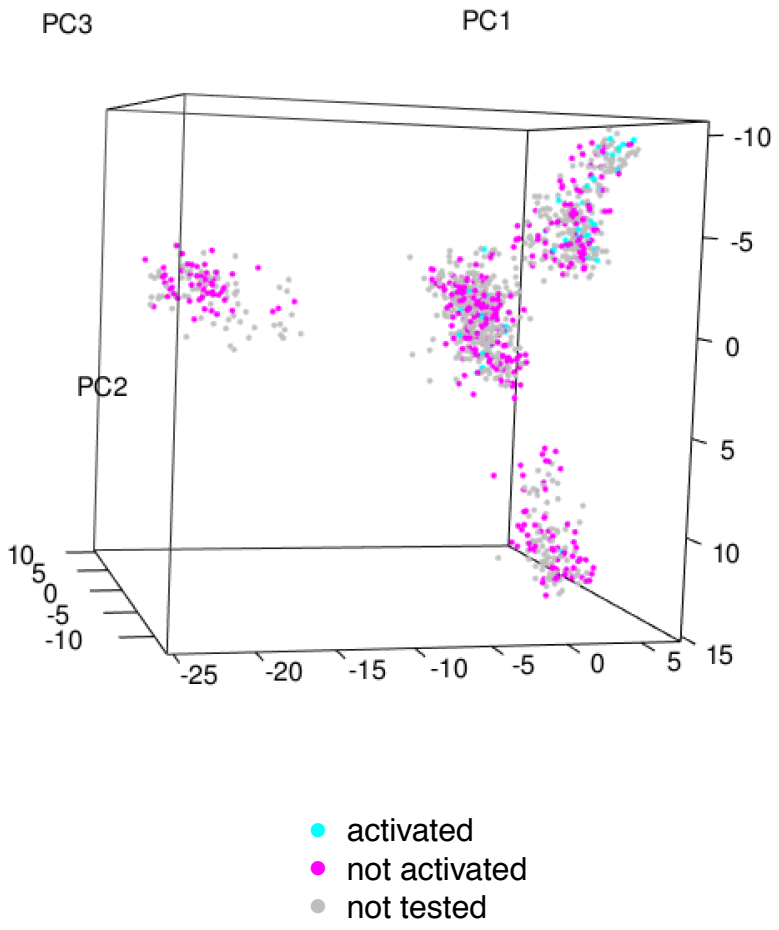

Figure 13: Principle component analysis of ORs. The first three PC's are used to separate ORs into four clusters. The acetophenone ORs are mainly present in two of the clusters (cyan).

3.3 Some amino acid residues are more conserved within acetophenone ORs

I therefore asked whether particular sites were conserved among the acetophenone ORs as a group. To identify residues that were more conserved among the 43 acetophenone ORs than by random chance, I simulated the distributions of mean Grantham distances within random OR sets for each residue in the alignment. To

simulate these distributions, 52 random ORs were randomly sampled from the mouse OR repertoire. At each amino acid residue, the pairwise Grantham distances across these 52 ORs were calculated and their mean was computed. Random sampling was iterated 20000 times. The per-residue mean Grantham distances were also computed for the group of 52 acetophenone ORs, and used to find p-values under the null-hypothesis that they were samples from the distributions of random OR sets. Locations with gaps for more than 10% of the ORs were excluded from the study. Multiple comparison was adjusted by controlling false discovery rate (FDR) (Benjamini and Hochberg, 1995). This comparison of mean Grantham distances of amino acid properties at individual sites among the acetophenone ORs with that of random OR sets identified 50 sites with higher amino acid similarity among acetophenone ORs than random ($n=50$, $p<0.05$; $n=23$, $p<0.01$; FDR adjusted, Figure 14). These sites are not highly conserved among ORs as a whole and the majority of them are located within transmembrane helices implicated in odorant binding (Abaffy et al., 2007; Baud et al., 2011; Doszczak et al., 2007; Singer, 2000).

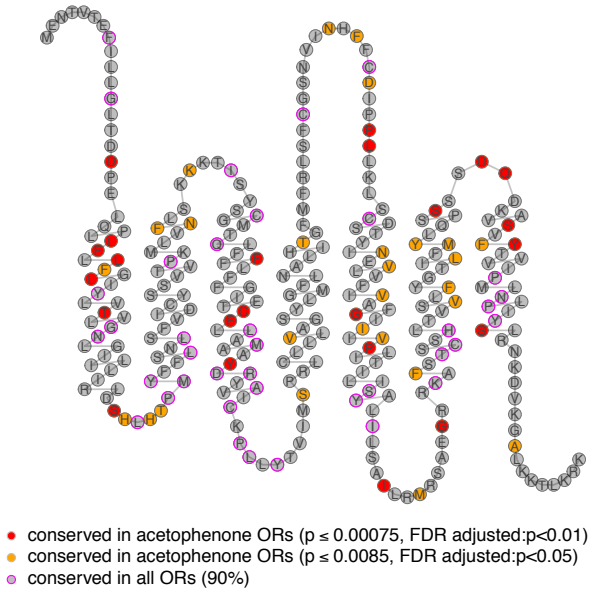

Figure 14: Amino acid residues that are more conserved in the acetophenone receptors than random OR sets of the same size. Red, p<0.01. Orange, p<0.05.

4. Modeling OR responses

The PCA and conserved residues in acetophenone ORs suggest certain combination of ORs might be able to predict OR responses. To test this hypothesis, I use machine learning techniques and generalized linear models to model the responses of ORs to acetophenone.

4.1 Overall sequence similarity based approach

As a first step, I implemented a classifier based on overall sequence similarity to compare all the other approaches with. This classifier searches for the OR in the training data that is most similar with the test OR, and label the test OR accordingly with a weight reflecting the sequence similarity between the two ORs. To test its performance, a 10-fold cross-validation scheme was performed for 100 iterations. At each iteration, we reserved 10% of the data as test subset at each iteration, trained the model with the remaining 90% and asked the model to predict the response of the test subset. The testing results from 100 iterations were merged to evaluate the model. The same scheme is used in all the following cross-validations. Indeed, overall sequence similarity within species (paralogs) predicts OR response poorly, with an AUC close to random (AUC=0.528, p=0.01, Figure 15).

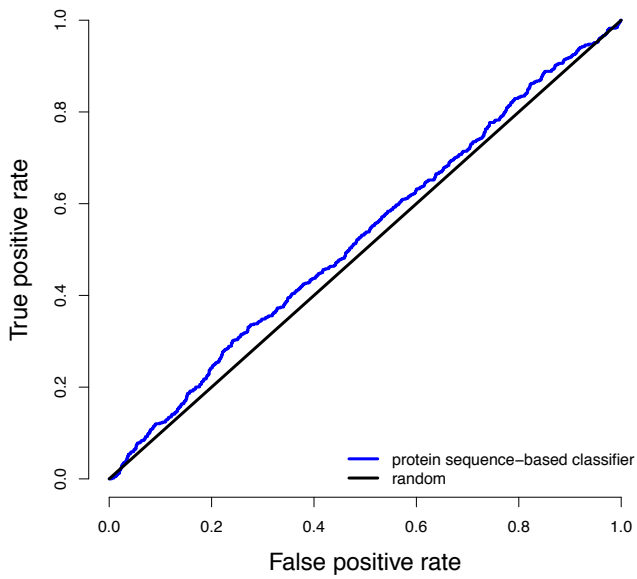

Figure 15: ROC of overall similarity based classifier, 10-fold cross-validation. AUC=0.528, p=0.01

4.2 Support Vector Machine approach

One candidate approach is to train the model using support vector machines (SVM). SVM searches for the optimal separating hyper-plane between the two classes through maximizing the margin between the closest points of the two classes.

4.2.1 Linear SVM

A C-classification SVM minimizes the following error function

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

with constraint

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

This turns out to involve only the dot product $x^T x$, the linear kernel. A gradient search for the parameter C was performed to tune for best cross-validation (100 iterations) performance. A model with $C = 0.01$ gives an AUC of 0.798 ($p=3\times10^{-112}$, Figure 16).

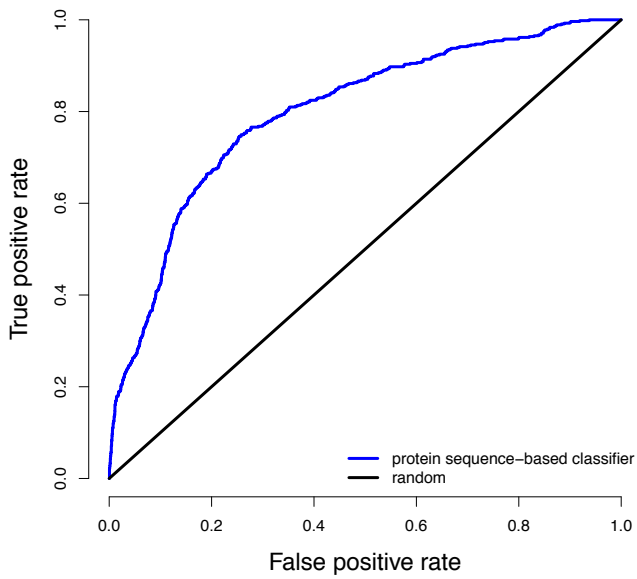

Figure 16: ROC of the linear SVM classifier, 10-fold cross-validation. AUC=0.798, $p=3\times10^{-112}$, Wilcoxon rank-sum test.

4.2.2 Kernel method

A commonly used kernel is the Radial basis function (RBF) kernel. The RBF kernel function is

$$K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right).$$

Let $\gamma = -\frac{1}{2\sigma^2}$, this becomes

$$K(x, x') = \exp(\gamma \parallel x - x' \parallel^2),$$

with an additional parameter γ to optimize.

Using a grid search for parameters that maximize cross-validation AUC, the parameters are chosen as $\gamma = 10, C = 0.001$. The AUC is 0.800 ($p=7\times10^{-114}$).

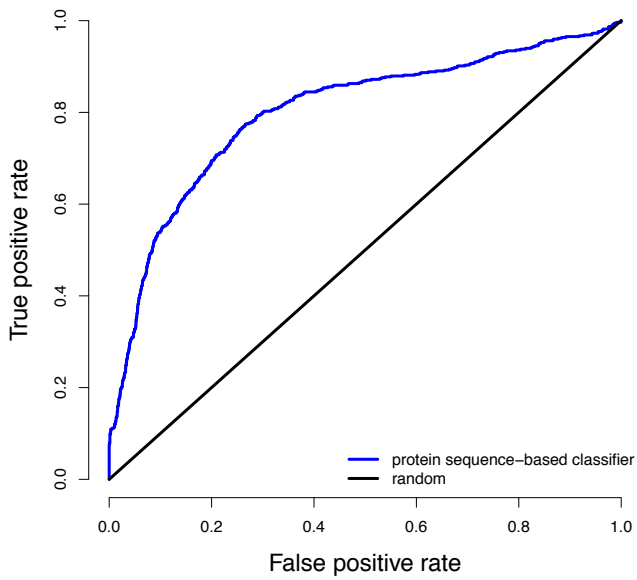

Figure 17: ROC of the RBF kernel SVM classifier, 10-fold cross-validation. AUC=0.800, $p=7\times10^{-114}$, Wilcoxon rank-sum test.

4.3 Semi-Supervised SVM (S3VM)

We do not have labels for about half of the ORs. The unlabeled data might imply underlying patterns of the data, and might guide model construction. To this end, I used a S3VM algorithm with a linear kernel. A S3VM algorithm minimizes the error fuction

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, (y_j)} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i + C^* \sum_{j=1}^{n} \xi_j$$

with the constraint

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0;\ y_j\left(w^T x_j + b\right) \geq 1 - \xi_j, \xi_j \geq 0$$

However, no dramatic enhancement in performance is seen (AUC = 0.736, $p=2\times10^{-181}$).

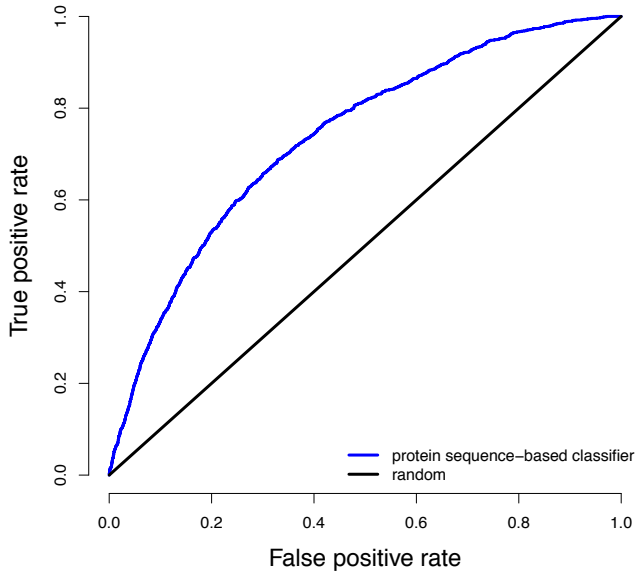

Figure 18: ROC of the S3VM classifier, 10-fold cross-validation. AUC=0.736, $p=2\times10^{-181}$, Wilcoxon rank-sum test.

4.4 Logistic regression and variable selection

Although SVM methods can be used to construct predictors, it is difficult to interpret the model coefficients. For better interpretability, I built a model using logistic regression that takes amino acid properties of ORs as input to predict their responses to acetophenone. To reduce over-fitting, I performed variable selection with elastic-net penalty (Friedman et al., 2010). For a logistic model

$$\pi(x) = p(y = 1|x) = \frac{\exp(x\beta)}{1+\exp(x\beta)},$$

find the β that minimizes the penalized log-likelihood

$$\frac{1}{N}\sum_{i=1}^{N}\left[1_{\{y_i=1\}}\log\pi(x_i)+1_{\{y_i=0\}}\log(1-\pi(x_i))\right]-\lambda P_\alpha(\beta),$$

with the penalty term being a combination of L1 and L2 norms of β weighted by α:

$$P_\alpha(\beta)=(1-\alpha)\frac{\|\beta\|_2^2}{2}+\alpha|\beta|.$$

A grid search finds the optimized $\alpha = 0.11$. The selected model has 47 predictors with AUC=0.834 ($p=2\times10^{-135}$, Figure 19).

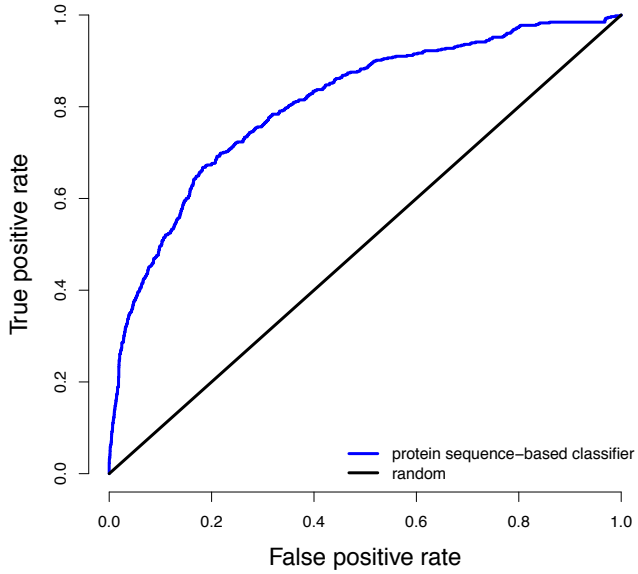

Figure 19: ROC of logistic regression with 47 predictors selected by elastic-net penalty, 10-fold cross-validation. AUC=0.834, $p=2\times10^{-135}$, Wilcoxon rank-sum test.

The selected predictors may reflect residues that contribute to the responsiveness of acetophenone. When plotted on a hypothetic OR, the selected predictors correspond

to residues that show little overlap with residues that are known to be conserved within the entire OR repertoire (Figure 20).

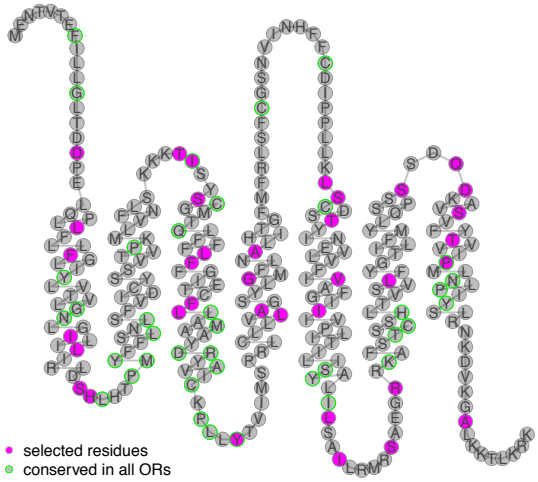

Figure 20: Distribution of selected residues on a hypothetic OR (magenta), and compared with residues that are conserved across all ORs, regardless of the response profile (green circles).

5. External Validation of the model

So far, the examination of model performance is based on cross-validation within the mouse ORs. However, it is desired that the models be tested on a separate dataset. In addition, the model derived from mouse ORs would be much more useful if it can generalize to human ORs. There are two main difficulties working with human ORs. First, our ability to measure their response *in vivo* is extremely limited. Second, only a small number of human ORs have been de-orphanized using *in vitro* systems, which may implicate the low efficiency of *in vitro* expression. Thus, a negative response *in vitro* may be false negative. To this end, I measured the *in vitro* responses of 27 human ORs that have been successfully expressed *in vitro* to acetophenone (Mainland et al., 2014). When stimulated with acetophenone, 9 of these human ORs showed response *in vitro.* In this chapter, I use these human ORs for external validation of the previously described models (Figure 21).

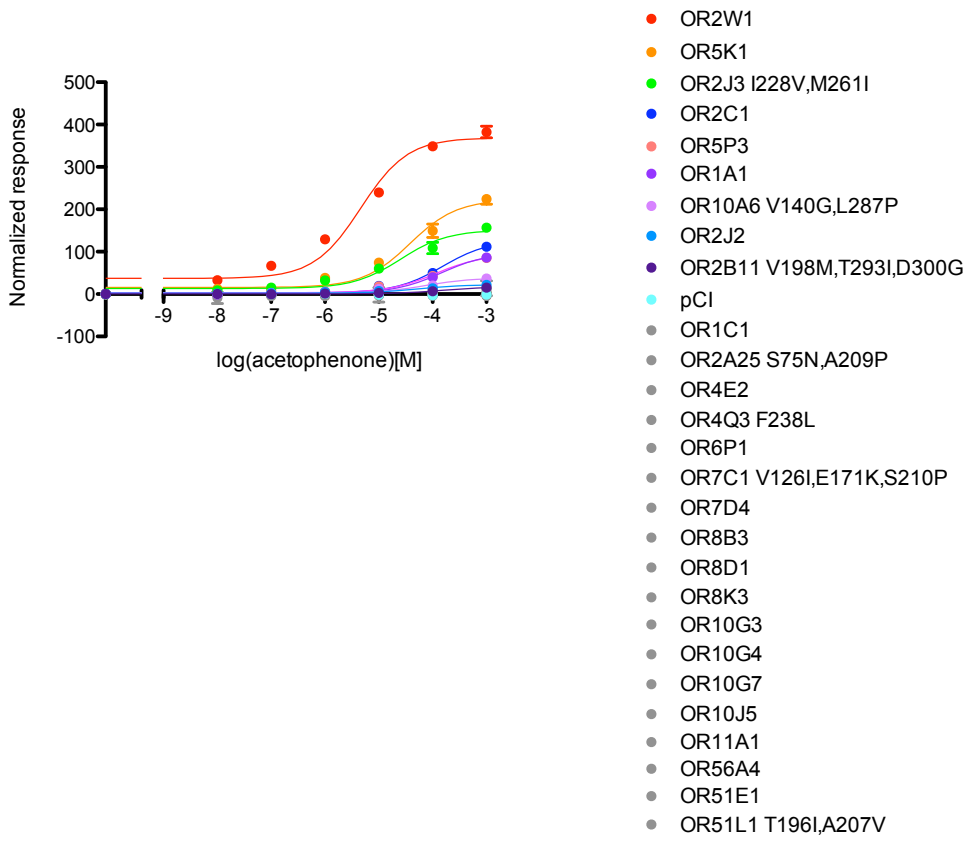

Figure 21: Dose response curves of 27 human ORs to acetophenone, used for external validation of the model.

5.1 Overall sequence similarity based approach

Again I start with the overall sequence similarity based approach. This approach is expected to show better performance when used to predict ORs from another species (orthologs) than within the same species (paralogs), as OR orthologs share a common ligand much more often than paralogs. Indeed, the AUC here is higher than cross-validation (AUC=0.66, p=0.1).

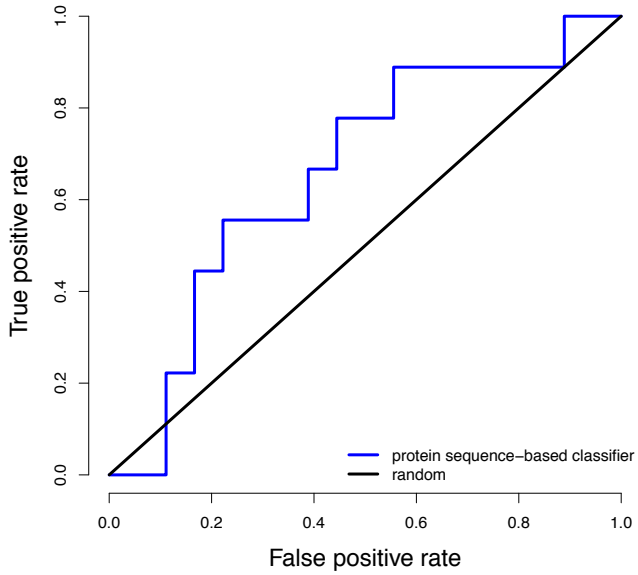

Figure 22: ROC of overall sequence similarity based classifier, external validation on 27 human ORs. AUC=0.660, p=0.1, Wilcoxon rank-sum test.

5.2 SVM and Logistic regression

The linear SVM model ($C = 0.01$) gives an AUC of 0.696 (p=0.05). The SVM model with RBF kernel ($\gamma = 10, C = 0.001$) gives an AUC of 0.713 (p=0.04). The semi-supervised linear SVM model gives an AUC of 0.684 (p=0.06). A logistic regression model with the selected 47 predictors gives an AUC of 0.728 (p=0.03).

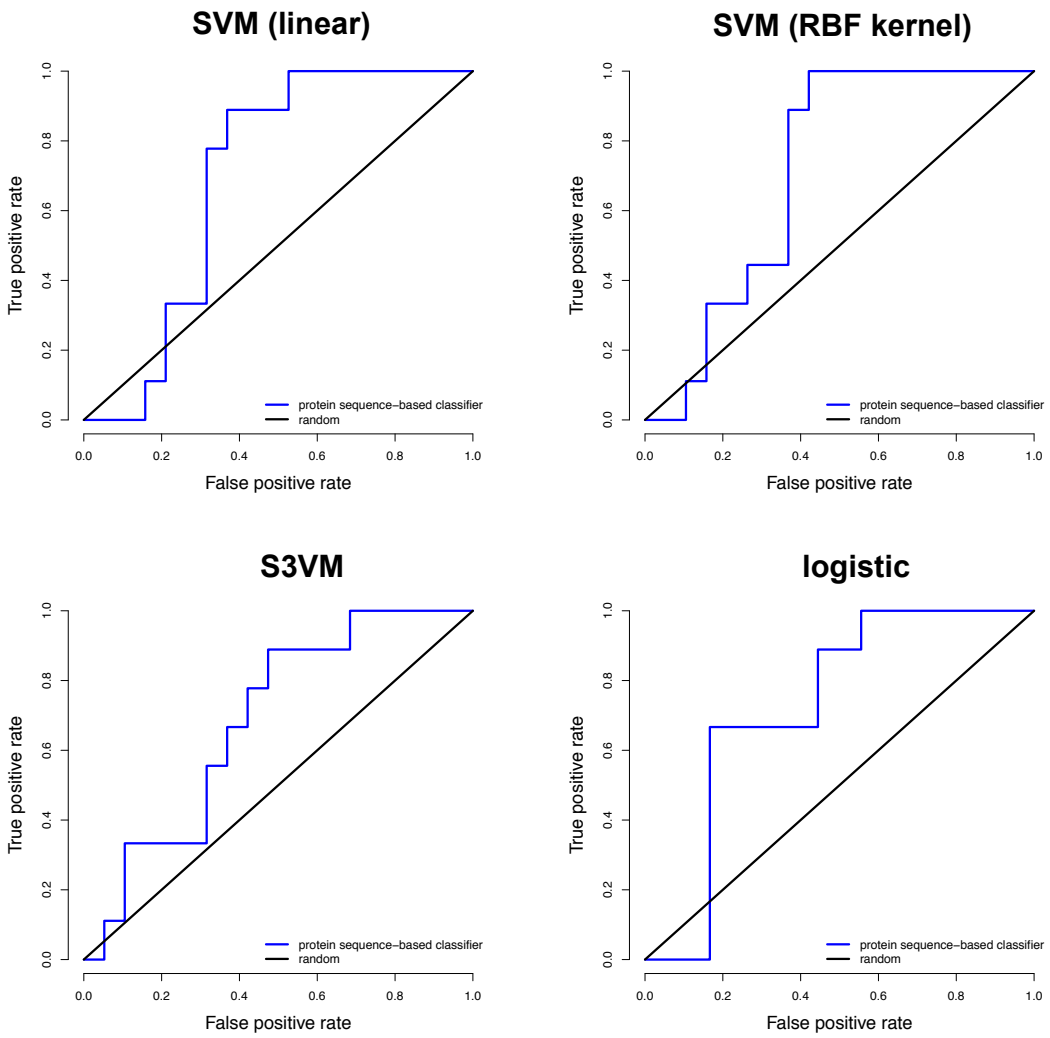

Figure 23: ROC of linear SVM (AUC=0.696, p=0.05), RBF kernel SVM (AUC=0.713, p=0.04), S3VM (AUC=0.684, p=0.06) and logistic regression with 47 predictors selected by elastic-net penalty (AUC=0.728,p=0.03). External validation on 27 human ORs. P-values are from Wilcoxon rank-sum test.

5.3 Conclusion

Finally, I summarize the model performances in cross- and external validations (Table 2). The large-scale identification of ORs for a given odorant, acetophenone, has allowed the construction of models that is superior to the current overall similarity-

based approach in several aspects. First, the new models are of better predictive values when used to predict novel OR responses. Second, the new models work generally for ORs both within the same species and from another species. Third, the variable selection approach provides hint for amino acid residues that may contribute to the responsiveness to specific odorants, thus generating hypotheses for future experimental investigation.

Table 2: Summary of model performance.

	Cross-validation		External validation	
	AUC	p-value	AUC	p-value
Similarity	0.528	0.01	0.660	0.10
SVM (linear)	0.798	3×10^{-112}	0.696	0.05
SVM (RBF)	0.800	7×10^{-114}	0.713	0.04
S3VM	0.736	2×10^{-181}	0.684	0.06
Logistic	0.834	2×10^{-135}	0.728	0.03

References

Abaffy, T., Malhotra, A., and Luetje, C.W. (2007). The molecular basis for ligand specificity in a mouse olfactory receptor: a network of functionally important residues. The Journal of biological chemistry *282*, 1216-1224.

Adipietro, K.A., Mainland, J.D., and Matsunami, H. (2012). Functional evolution of mammalian odorant receptors. PLoS genetics *8*, e1002821.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Araneda, R.C., Kini, A.D., and Firestein, S. (2000). The molecular receptive range of an odorant receptor. Nature neuroscience *3*, 1248-1255.

Baud, O., Etter, S., Spreafico, M., Bordoli, L., Schwede, T., Vogel, H., and Pick, H. (2011). The mouse eugenol odorant receptor: structural and functional plasticity of a broadly tuned odorant binding pocket. Biochemistry *50*, 843-853.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Boyle, S.M., McInally, S., and Ray, A. (2013). Expanding the olfactory code by in silico decoding of odor-receptor chemical space. eLife *2*, e01120.

Bozza, T., Feinstein, P., Zheng, C., and Mombaerts, P. (2002). Odorant receptor expression defines functional units in the mouse olfactory system. The Journal of neuroscience : the official journal of the Society for Neuroscience *22*, 3033-3043.

Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell *65*, 175-187.

Buck, L.B. (2004). Olfactory receptors and odor coding in mammals. Nutrition reviews *62*, S184-188; discussion S224-141.

Bushdid, C., Magnasco, M.O., Vosshall, L.B., and Keller, A. (2014). Humans can discriminate more than 1 trillion olfactory stimuli. Science *343*, 1370-1372.

Cao, R., Lee, B., Cho, H.Y., Saklayen, S., and Obrietan, K. (2008). Photic regulation of the mTOR signaling pathway in the suprachiasmatic circadian clock. Molecular and cellular neurosciences *38*, 312-324.

Dong, P.Y., Gong, N.N., and Matsunami, H. (2014). Cell-Based System for Identification of Olfactory Receptors. In Bioelectronic Nose, T.H. Park, ed. (Springer Netherlands), pp. 83-95.

Doszczak, L., Kraft, P., Weber, H.P., Bertermann, R., Triller, A., Hatt, H., and Tacke, R. (2007). Prediction of perception: probing the hOR17-4 olfactory receptor model with silicon analogues of bourgeonal and lilial. Angew Chem Int Ed Engl *46*, 3367-3371.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software *33*, 1-22.

Gaillard, I., Rouquier, S., Chavanieu, A., Mollard, P., and Giorgi, D. (2004). Amino-acid changes acquired during evolution by olfactory receptor 912-93 modify the specificity of odorant recognition. Human molecular genetics *13*, 771-780.

Gilad, Y., and Lancet, D. (2003). Population differences in the human functional olfactory repertoire. Molecular biology and evolution *20*, 307-314.

Gilad, Y., Man, O., Paabo, S., and Lancet, D. (2003). Human specific loss of olfactory receptor genes. Proceedings of the National Academy of Sciences of the United States of America *100*, 3324-3327.

Godfrey, P.A., Malnic, B., and Buck, L.B. (2004). The mouse olfactory receptor gene family. Proceedings of the National Academy of Sciences of the United States of America *101*, 2156-2161.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862-864.

Grosmaitre, X., Fuss, S.H., Lee, A.C., Adipietro, K.A., Matsunami, H., Mombaerts, P., and Ma, M. (2009). SR1, a mouse odorant receptor with an unusually broad response profile. The Journal of neuroscience : the official journal of the Society for Neuroscience *29*, 14545-14552.

Haga, S., Hattori, T., Sato, T., Sato, K., Matsuda, S., Kobayakawa, R., Sakano, H., Yoshihara, Y., Kikusui, T., and Touhara, K. (2010). The male mouse pheromone ESP1 enhances female sexual receptive behaviour through a specific vomeronasal receptor. Nature *466*, 118-122.

Hasin, Y., Olender, T., Khen, M., Gonzaga-Jauregui, C., Kim, P.M., Urban, A.E., Snyder, M., Gerstein, M.B., Lancet, D., and Korbel, J.O. (2008). High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. PLoS genetics *4*, e1000249.

Healy, M.D., Smith, J.E., Singer, M.S., Nadkarni, P.M., Skoufos, E., Miller, P.L., and Shepherd, G.M. (1997). Olfactory receptor database (ORDB): a resource for sharing and analyzing published and unpublished data. Chemical senses *22*, 321-326.

Isogai, Y., Si, S., Pont-Lezica, L., Tan, T., Kapoor, V., Murthy, V.N., and Dulac, C. (2011). Molecular organization of vomeronasal chemoreception. Nature *478*, 241-245.

Keller, A., Zhuang, H., Chi, Q., Vosshall, L.B., and Matsunami, H. (2007). Genetic variation in a human odorant receptor alters odour perception. Nature *449*, 468-472.

Keydar, I., Ben-Asher, E., Feldmesser, E., Nativ, N., Oshimoto, A., Restrepo, D., Matsunami, H., Chien, M.S., Pinto, J.M., Gilad, Y., *et al.* (2013). General olfactory sensitivity database (GOSdb): candidate genes and their genomic variations. Human mutation *34*, 32-41.

Knight, Z.A., Tan, K., Birsoy, K., Schmidt, S., Garrison, J.L., Wysocki, R.W., Emiliano, A., Ekstrand, M.I., and Friedman, J.M. (2012). Molecular profiling of activated neurons by phosphorylated ribosome capture. Cell *151*, 1126-1137.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annual review of genetics *39*, 309-338.

Krautwurst, D., Yau, K.W., and Reed, R.R. (1998). Identification of ligands for olfactory receptors by functional expression of a receptor library. Cell *95*, 917-926.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Li, J., Haddad, R., Chen, S., Santos, V., and Luetje, C.W. (2012). A broadly tuned mouse odorant receptor that detects nitrotoluenes. Journal of neurochemistry *121*, 881-890.

Li, Y.R., and Matsunami, H. (2011). Activation state of the M3 muscarinic acetylcholine receptor modulates mammalian odorant receptor signaling. Science signaling *4*, ra1.

Mainland, J.D., Keller, A., Li, Y.R., Zhou, T., Trimmer, C., Snyder, L.L., Moberly, A.H., Adipietro, K.A., Liu, W.L., Zhuang, H., *et al.* (2014). The missense of smell: functional variability in the human odorant receptor repertoire. Nature neuroscience *17*, 114-120.

Malnic, B., Godfrey, P.A., and Buck, L.B. (2004). The human olfactory receptor gene family. Proceedings of the National Academy of Sciences of the United States of America *101*, 2584-2589.

Malnic, B., Hirono, J., Sato, T., and Buck, L.B. (1999). Combinatorial receptor codes for odors. Cell *96*, 713-723.

Man, O., Gilad, Y., and Lancet, D. (2004). Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. Protein science : a publication of the Protein Society *13*, 240-254.

McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research *40*, 4288-4297.

McGowen, M.R., Clark, C., and Gatesy, J. (2008). The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. Systematic biology *57*, 574-590.

Menashe, I., Abaffy, T., Hasin, Y., Goshen, S., Yahalom, V., Luetje, C.W., and Lancet, D. (2007). Genetic elucidation of human hyperosmia to isovaleric acid. PLoS biology *5*, e284.

Menashe, I., Man, O., Lancet, D., and Gilad, Y. (2003). Different noses for different people. Nature genetics *34*, 143-144.

Nara, K., Saraiva, L.R., Ye, X., and Buck, L.B. (2011). A large-scale analysis of odor coding in the olfactory epithelium. The Journal of neuroscience : the official journal of the Society for Neuroscience *31*, 9179-9191.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome research *24*, 1485-1496.

Niimura, Y., and Nei, M. (2005). Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. Gene *346*, 23-28.

Olender, T., Waszak, S.M., Viavant, M., Khen, M., Ben-Asher, E., Reyes, A., Nativ, N., Wysocki, C.J., Ge, D., and Lancet, D. (2012). Personal receptor repertoires: olfaction as a model. BMC genomics *13*, 414.

Peterlin, Z., Firestein, S., and Rogers, M.E. (2014). The state of the art of odorant receptor deorphanization: A report from the orphanage. The Journal of general physiology *143*, 527-542.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Reisert, J., and Restrepo, D. (2009). Molecular tuning of odorant receptors and its implication for odor signal processing. Chemical senses *34*, 535-545.

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.

Saito, H., Chi, Q., Zhuang, H., Matsunami, H., and Mainland, J.D. (2009). Odor coding by a Mammalian receptor repertoire. Science signaling *2*, ra9.

Saito, H., Kubota, M., Roberts, R.W., Chi, Q., and Matsunami, H. (2004). RTP family members induce functional expression of mammalian odorant receptors. Cell *119*, 679-691.

Schmiedeberg, K., Shirokova, E., Weber, H.P., Schilling, B., Meyerhof, W., and Krautwurst, D. (2007). Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. Journal of structural biology *159*, 400-412.

Shirasu, M., Yoshikawa, K., Takai, Y., Nakashima, A., Takeuchi, H., Sakano, H., and Touhara, K. (2014). Olfactory receptor and neural pathway responsible for highly selective sensing of musk odors. Neuron *81*, 165-178.

Singer, M.S. (2000). Analysis of the molecular basis for octanal interactions in the expressed rat 17 olfactory receptor. Chemical senses *25*, 155-165.

Touhara, K., Sengoku, S., Inaki, K., Tsuboi, A., Hirono, J., Sato, T., Sakano, H., and Haga, T. (1999). Functional identification and reconstitution of an odorant receptor in single olfactory neurons. Proceedings of the National Academy of Sciences of the United States of America *96*, 4040-4045.

Valjent, E., Bertran-Gonzalez, J., Bowling, H., Lopez, S., Santini, E., Matamales, M., Bonito-Oliva, A., Herve, D., Hoeffer, C., Klann, E., *et al.* (2011). Haloperidol regulates the state of phosphorylation of ribosomal protein S6 via activation of PKA and phosphorylation of DARPP-32. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology *36*, 2561-2570.

Verbeurgt, C., Wilkin, F., Tarabichi, M., Gregoire, F., Dumont, J.E., and Chatelain, P. (2014). Profiling of olfactory receptor gene expression in whole human olfactory mucosa. PloS one *9*, e96333.

Villanueva, E.C., Munzberg, H., Cota, D., Leshan, R.L., Kopp, K., Ishida-Takahashi, R., Jones, J.C., Fingar, D.C., Seeley, R.J., and Myers, M.G., Jr. (2009). Complex regulation of mammalian target of rapamycin complex 1 in the basomedial hypothalamus by leptin and nutritional status. Endocrinology *150*, 4541-4551.

Wang, S., Mortazavi, L., and White, K.A. (2008). Higher-order RNA structural requirements and small-molecule induction of tombusvirus subgenomic mRNA transcription. Journal of virology *82*, 3864-3871.

Waszak, S.M., Hasin, Y., Zichner, T., Olender, T., Keydar, I., Khen, M., Stutz, A.M., Schlattl, A., Lancet, D., and Korbel, J.O. (2010). Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. PLoS computational biology *6*, e1000988.

Zeng, L.H., Rensing, N.R., and Wong, M. (2009). The mammalian target of rapamycin signaling pathway mediates epileptogenesis in a model of temporal lobe epilepsy. The Journal of neuroscience : the official journal of the Society for Neuroscience *29*, 6964-6972.

Zhang, X., and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. Nature neuroscience *5*, 124-133.

Zhao, H., Ivic, L., Otaki, J.M., Hashimoto, M., Mikoshiba, K., and Firestein, S. (1998). Functional expression of a mammalian odorant receptor. Science *279*, 237-242.

Zhuang, H., Chien, M.S., and Matsunami, H. (2009). Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates. Proceedings of the National Academy of Sciences of the United States of America *106*, 21247-21251.

Zhuang, H., and Matsunami, H. (2008). Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. Nature protocols *3*, 1402-1413.