BIOINFORMATIC AND MOLECULAR ANALYSIS OF OLFACTORY RECEPTOR GENE REGULATION IN ZEBRAFISH

by

Ahmet Burak Kaya

B.S., Molecular Biology and Genetics, Bilkent University, 2012

Submitted to the Institute of Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree
Master of Science

Graduate Program in Molecular Biology and Genetics
Boğaziçi University

2015

ACKNOWLEDGEMENTS

I would like to thank to my thesis supervisor Assoc. Prof. Stefan Fuss for his supervision and guidance during my graduate studies.

I am very grateful to my committee members, Prof. Uğur Sezerman and Assoc. Prof Necla Birgül-İyison for accepting to read and review my thesis.

I was supported by the Scientific and Technological Research Council of Turkey (TÜBITAK) with the scholarship in conjunction with "2210-E Doğrudan Yurt İçi Yüksek Lisans Burs Programı (2012-2)".

Work carried out in this project was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) Grant number: 112T168 – "Zebra balığında koku reseptör gen düzenleyici bölgelerin işlevsel karakterizasyonu".

I would like to express my sincere thanks to my lab mates Metin, Mehmet Can, Tuba, Büşra, Serdar, Burak, Kerem, Yusuf, Özge and Xalid for the great times in the lab. I would like to especially thank my lovely friend Gizem for being more than a lab partner.

I would like to thank my dearest friends Cüneyt, Emrecan, Efe, Ece, Ayşe, Kaan, Çağlayan, Burcu, Nehir and Fatih for everything.

I would like to express my gratitude to Elif Begüm Gökerküçük for providing me food, shelter and infinite support.

Last but not least, I would like to express my deepest thanks to my family for their love and support.

ABSTRACT

BIOINFORMATIC AND MOLECULAR ANALYSIS OF OLFACTORY RECEPTOR GENE REGULATION IN ZEBRAFISH

In vertebrates, an individual olfactory sensory neuron (OSN) expresses only one allele of one olfactory receptor (OR) gene from a huge genomic repertoire. The molecular mechanism governing "one neuron-one receptor rule" are not fully understood and may be regulated by a combination of a variety of cellular mechanisms. Two OR locus-related regulatory mechanisms previously identified are the interaction of transcription factors with specific DNA-binding motifs as a short-range control and the activity of Locus Control Regions as long-range control. Curiously, several regulatory motifs such as Olf1/Ebf1, homeodomain (Lhx2/Emx2) and BPTF motifs were identified in promoter and LCR regions of mice. Here, RNA-Sequencing and bioinformatic analysis were performed in order to detect conserved regulatory motifs within zebrafish OR promoters. Transcript structures and expression levels of OR repertoire were obtained by analysis of transcriptome data. Also, TSS of 161 OR genes were resolved and promoter regions directly upstream of TSSs were identified. De novo motif search with motif-finding bioinformatic tools resulted in a zebrafish-specific motif which is highly similar to Ebf1. Investigation of known regulatory motifs Ebf1, TBP, Lhx2, Emx2 and BPTF indicated presence and distribution of these motifs in OR promoters. Next, a member of highly expressed OR gene family, OR132-5, was investigated with *in situ* hybridization and a correlation between FPKM levels and number of OSNs that express a given OR was observed. Furthermore, a candidate regulatory motif was identified in this highly expressed OR gene family.

ÖZET

ZEBRABALIĞINDA KOKU DUYUSUNA AİT RESEPTÖR GEN REGÜLASYONUNUN BİYOİNFORMATİK VE MOLEKÜLER ANALİZİ

Omurgalılarda her bir koku algılayıcı sinir hücresi büyük bir genomik repertuar içerisinden yalnızca bir koku reseptör geninin tek bir alelini ifade etmektedir. "Bir nöron-bir reseptör kuralı" nı kontrol eden moleküler mekanizma tam olarak anlaşılabilmiş değildir ve çeşitli hücresel mekanizmaların kombinasyonlarıyla düzenleniyor olabileceği düşünülmektedir. Daha önce belirlenen iki adet koku reseptörü lokus-bazlı düzenleyici mekanizmalar kısa menzilli kontrol olarak transkripsiyon faktörlerinin DNA'ya bağlanan özel motiflerle etkileşimi ve uzun menzilli kontrol olarak Lokus Kontrol Bölgeleri'nin aktivitesidir. Ilginç bir biçimde, Olf1/Ebf1, homeodomain (Lhx2/Emx2) ve BPTF gibi birkaç düzenleyici motif faredeki promotör ve LCR bölgelerinde belirlenmiştir. Zebrabalığının koku reseptör genlerinin promotör kısımlarındaki korunmuş düzenleyici motifleri saptamak için bu projede RNA-sekanslama ve biyoinformatik analizi yapıldı. Koku reseptör repertuarının transkript yapıları ve ifade edilme seviyeleri transkriptom data analizi ile belirlenmiştir. Aynı zamanda, 161 koku reseptörü geninin transkripsiyon başlangıç bölgeleri çözümlenmiş ve transkripsiyon başlangıç bölgelerinin direkt olarak yukarı yönündeki promotör bölgeleri belirlenmiştir. Motif bulucu biyoinformatik araçlarıyla de novo motif araştırması Ebf1'e oldukça benzer zebrabalığına özel bir motif tespit etmiştir. Ebf1, TBP, Lhx2, Emx2 ve BPTF gibi bilinen düzenleyici motiflerin araştırılması bu motiflerin varlığını ve koku duyusu reseptör gen promotör bölgelerindeki dağılımını göstermiştir. Daha sonra, in situ hibridizasyon ile koku reseptör gen ailesinin yüksek derecede ifade edilen bir elemanı olan OR132-5 araştırılmış ve FPKM seviyelerinin verilen koku reseptörünü ifade eden OSN'lerin sayısıyla bağlantısı gözlemlenmiştir. Buna ek olarak, yüksek derecede ifade edilen bu koku reseptör gen ailesinden aday bir düzenleyici motif belirlenmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS / ABBREVIATIONS

BAC	Bacterial Artificial Chromosome
bp	Base Pair
cDNA	Complementary Deoxyribonucleic Acid
DNA	Deoxyribonucleic Acid
GPCR	G-protein-coupled Receptor
kb	Kilobase Pair
mRNA	Messenger Ribonucleic Acid
OB	Olfactory Bulb
OE	Olfactory Epithelium
OMP	Olfactory Marker Protein
OR	Odorant Receptor
OSN	Olfactory Sensory Neuron
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PWM	Positional Weight Matrix
RNA	Ribonucleic Acid
TSS	Transcription Start Site
UTR	Untranslated Region

1. INTRODUCTION

1.1. The Olfactory System

Chemosensation is one of the essential and evolutionary ancient sensory abilities that is present in almost all organisms, from unicellular organisms to the most complex life forms. Even in a primitive process like chemotaxis in bacteria, main characteristics of chemosensation, such as ligand induced chemoreceptor activation and behavioral response to the external stimuli, are present (Adler, 1966). In higher organisms, chemosensory systems evolved to detect chemical cues from their environment and turn these cues into signals that can lead to complex behavior such as locating food, finding mates, detecting preys and avoiding predators (Prasad and Reed, 1999). The vertebrate olfactory system is specialized to detect common odorant and pheromones. In this system, olfactory sensation is mediated by special olfactory sensory neurons (OSNs) located in the olfactory epithelium (OE) of the nasal cavity. Odorant chemicals bind to olfactory receptors (ORs) expressed by OSNs and induce electrical signals that are transmitted to the brain. Expression of ORs in vertebrates is a tightly regulated process. Interestingly, an individual OSN only express a single OR gene from an enormous repertoire (Malnic *et al*., 1999). However, the exact mechanisms that regulate the expression of chemoreceptor genes are not properly understood.

1.1.1. Anatomy of the Olfactory System

Detection of odorants takes place in the OE of the peripheral olfactory system. In vertebrates, the OE resides inside the nasal cavity and its function is linked to the respiratory system (Song *et al*., 2013). Active inhalation through the nose delivers chemicals from the environment to OSNs that express ORs, which interact with odorants at the molecular level. Curiously, each OSN expresses only a single OR from a large and diverse repertoire of OR

genes (Malnic *et al*., 1999). Thus, OSNs are specialized in the detection of odorants by virtue of the OR that they express and the ligands that can bind to the receptor.

OSNs send axonal projections to glomeruli of a forebrain structure called the olfactory bulb ((OB), Vassar *et al*.,1994; Ressler *et al*., 1994; Mombaerts *et al*., 1996). Importantly, OSNs which express the same OR converge onto the same glomerulus within the OB (Mombaerts *et al*., 1996; Vassar *et al*., 1994; Ressler *et al*., 1994). Because of the one to one relationship between OR expression and glomerular projection, glomeruli form a representational map of odor identity in the brain. Thus, monogenic OR expression in OSNs contributes to both, the architecture of the glomerular map and molecular basis of smell perception.

In rodents, a second peripheral sensory organ called the vomeronasal organ (VNO) functions is the detection of pheromones and kairomones (Firestein *et al*., 2001). Neurons in the VNO express a related family of chemosensory receptors, the vomeronasal receptors, and project their axons to the accessory olfactory bulb (AOB) located at the posterior end of the OB (Belluscio *et al*., 1999).

In contrast to the mammalian system, the zebrafish olfactory system only has a single OE, which harbors OSNs and VNO-like neurons that are able to detect both, odorants and pheromones. The zebrafish OE has a rosette like structure and is formed through the folding of the tissue into a large number of staggered lamellae (Hansen and Zeiske, 1998).

Four distinct types of OSNs have been identified so far in the zebrafish OE: cilliated, microvillous, crypt and kappe neurons (Hansen and Zeiske, 1998; Hamdani and Døving, 2007; Ahuja *et al*., 2014). Cilliated neurons are located in more basal layers of OE and either express ORs or trace amine-associated receptors (TAARs; Liberles and Buck 1996). Compared to ORs, the odorant spectrum of TAARs is narrower and restricted to volatile amines (Ihara *et al*., 2013).

Cilliated neurons can be distinguished at the molecular level by their expression of olfactory marker protein (OMP, Çelik *et al*., 2002). Microvillous OSNs are located in the VNO in rodents, whereas in zebrafish they are also located in the same OE as ciliated cells, but in more apical layers. They express V1R and V2R type receptors and can be identified by their expression of the transient receptor potential channel C2 (TRPC2; Dulac and Axel, 1995; Ben-Shaul *et al*., 2010). Another zebrafish-specific subtype of chemoreceptor cells are the crypt neurons, which express a single V1R-related gene, ORA4, and which are located in the most apical layer of the OE (Oka *et al*., 2012). It was shown that crypt cells are immunoreactive to TrkA and S-100 calcium binding protein but most likely do not actively express these proteins (Catania *et al*., 2003; Germana *et al*., 2004). Recently, a new type of OSNs, the kappe neurons were discovered in zebrafish (Ahuja *et al*., 2014). Similar to cyrpt cells, kappe neurons are located in the apical OE and can be labeled by virtue of Go-like immunoreactivity.

At the level of the OB, the organization of glomerular map in zebrafish is similar to its vertebrate counterpart. Using DiI tracing from the OE to the OB a stereotyped glomerular map of 80 glomeruli with left right symmetry could be identified (Baier and Korsching, 1994). More recently, by using specific antibodies which recognizes general neuronal markers, it was possible to clearly and invariantly identify about 140 individual glomeruli along with their spatial distribution in in various areas of OB (Braubach *et al*., 2012).

1.1.2. Odorant Receptor (OR) Repertoire

ORs were discovered in 1991 by seminal research of Linda Buck and Richard Axel (Buck and Axel, 1991). ORs are rhodopsin-like type A GPCR proteins with seven transmembrane (TM) domains and short N- and C- termini (Buck and Axel, 1991). Further analysis of the entire OR repertoire revealed conserved amino acid motifs: GN in TMI domain, PMYF/LFL in TMII domain, MAYDRYVAIC in TMIII domain, KAFSTCA/GSHLSVV in TMVI, PMLNPFIYSLRN in TMVII (Buck and Axel, 1991). The binding pockets for ligand

binding are proposed to be formed by TM domains: one pocket is suggested to be located in TMIII, V and VII whereas the other pocket is suggested to be from TM3 to TM7 (Emes *et al*., 2004; Liu *et al*., 2003).

Binding of odor ligands to ORs activates an OSN-specific isoform of the heterotrimeric G-protein complex (Jones and Reed 1989). Dissociation of Gaolf triggers the activation of olfactory-specific adenylate cyclase type III (Adcy3, Bakalyar and Reed 1990), which causes an increase in intracellular cAMP levels. Elevated cAMP in turn opens a cyclic nucleotide-gated cation channel, OCNC1 (Wang and Reed 1993). Inward flux of calcium through OCNC1 then triggers an outward-directed chloride current through the calcium activated chloride channel ANO2 (Stephan *et al*., 2009).

Recent development of sequencing technologies and bioinformatics methods facilitated the discovery of OR genes in a large number of diverse species (Niimura and Nei, 2007; Gilad *et al*., 2005; Glusman *et al*., 2001; Go and Niimura, 2008; Quignon *et al*., 2005) and let to the discovery of basic principles of the molecular and genomic organization of OR genes. OR genes are principally found in clusters in the genome and do not mingle with non-OR genes (Sullivan *et al*., 1996; Rouqier *et al*., 1998). The gene structure of ORs is rather simple: the coding sequence of about 1 kb length is located on a single exon while several non-coding 5'-exons may exist (Hoppe *et al*., 2003; Mombaerts, 1999, Sosinsky *et al.*, 2000; Volz *et al*., 2003; Young *et al*., 2003).

In mammals, it was found that the OR gene repertoire varies between 260 and 1.300 genes and in most species, OR genes are the largest gene family (Grus et al., 2005, Niimura and Nei 2007, Zhang et al., 2007). However, not all OR genes are and of the fraction of pseudogenes varies between 20% in the mouse to 72% in humans (Zhang et al., 2009). The size of the OR repertoire in fish is much smaller, ranging only between 40 and 140 genes (Alioto and Ngai, 2005; Niimura and Nei, 2005). In zebrafish, 136 functional OR genes have been described so

far (Alioto and Ngai, 2005) and the genomic organization is similar to mammals (Alioto and Ngai, 2005). Generally, OR subfamily members have the same transcriptional orientation, suggesting tandem duplications of OR genes during evolution (Korsching, 2009).

1.1.3. Class Distinction in Odorant Receptors

ORs can be divided into two phylogenetic class based on their amino acid sequence: class I and class II ORs. Class I ORs were first identified in fish (Ngai *et al*., 1993) and frog (Freitag *et al.*, 1996) and it was suggested that these receptors play a role in detecting water-soluble odorants. About 10% of the mammalian ORs belong to class I ORs and are located in a single large cluster (Zhang and Firestein, 2002), which suggested that class I genes are evolutionarily more ancient than Class II ORs (Zhang and Firestein, 2002). It was also shown that class I genes are responsive to the water-soluble odorants such as aldehydes, aliphatic acids and alcohol (Malnic *et al*., 1999; Kobayakawa *et al*., 2007). The OR gene repertoire of the zebrafish, however, is almost entirely composed of class I or phylogenetically related genes (Alioto and Ngai, 2005; Freitag et al., 1998). In contrast, most of the OR gene repertoire in mammals is comprised of Class II genes (90%) and these ORs are primarily detect volatile odorants (Zhang and Firestein, 2002). In zebrafish, only a single class II-related OR has been identified, the evolutionary relationship of which to mammalian class II receptors remains to be elucidated (Alioto and Ngai, 2005; Niimura and Nei, 2005).

1.2. Expression of Odorant Receptor Genes

1.2.1. Monogenic Expression

A hallmark of the olfactory sensory system is the fact that each OSN expresses only a single OR gene from the large genomic repertoire (Malnic *et al*., 1999). In their seminal paper,

Buck and Axel initially proposed the ‘one neuron-one receptor hypothesis’ based on the frequencies of OR genes represented in cDNA libraries from OE tissue (Buck and Axel, 1991). Even though hard to prove unequivocally, additional evidence for the one neuron – one receptor rule has emerged over the past years (Mombaerts, 2004).

In situ hybridization against OR transcripts provides the abilities to analyze both, the number of cells which express any given OR and the topographic arrangement of OSNs expressing the OR in the OE. Based on OMP expression, it was reported that the OE at three weeks of age contains 15 million OSNs in rats whereas in the adult 22 million OSNs are present (Meisami, 1989; Youngentob *et al*., 1997). Therefore, several thousand OSNs must express a given OR gene in the rat epithelium. These numbers were confirmed for several OR genes in mouse and rat, although large differences exist for different OR genes (Ressler *et al*., 1993; Strotmann *et al*., 1994; Kubick *et al*., 1997; Royal and Key, 1999; Iwema *et al*., 2003).

The best direct evidence for the one neuron – one receptor rule comes from reverse transcription polymerase chain reaction (RT-PCR) studies (Malnic et al., 1999). Using single OSNs as templates, only one OR gene could be amplified in half of the OSNs, whereas no OR could be amplified from the other half. Additional support comes from functional studies that showed that odorant responsiveness of OSNs depends to their respectively amplified OR gene (Touhara *et al*., 2000; Kajiya *et al*., 2001). Theoretically, if an OSN is capable of expressing only a single gene, altering expressed OR gene might alter the odorant responsiveness of the cell. Changing expressed odorant receptor from M71 TO I7 resulted in a change of odorant responsiveness from acetophenone to octanal as expected (Bozza *et al*., 2002).

Despite the evidence summarized above, violations to the rule have also been reported. The first cases of co-expression was reported for the rat i9 and HGL-SL2 OR genes by in situ hybridization (Rawson *et al*., 2000). Also, 2% (P0) and 0.2% (1 month) of coexpression rate was observed in double *in situ* experiments in mouse septal organ (Tian and Ma, 2008). In

zebrafish, in situ hybridization experiments showed coexpression of the OR103-1 and OR103-5/2 genes (Sato *et al*., 2007).

1.2.2. Monoallelic Expression

In addition to monogenic expression, OR genes are expressed in monoallelic fashion (Chess *et al*., 1994). A given OSN expresses either the maternal or the paternal allele but not both. This was initially shown by using primers which recognize polymorphic sequences in the I7 OR by RT-PCR, where only one of the alleles could be amplified from individual OSNs (Chess *et al*., 1994).

A more direct evidence is provided by DNA/RNA in situ hybridization experiments in OSN nuclei showing nascent RNA only from one of the two genomic loci coding for the expressed OR (Ishii *et al*., 2001). In addition, transgenic mice expressing different fluorescent reporters from the two alleles show no overlap in signals (Li *et al*.,2004). Cell counts for tagged ORs in homozygous and heterozygous mice also show a predicted 2:1 ratio, indirectly supporting monoallelic expression (Mombaerts *et al*., 1996).

1.2.3. Zonal Expression

Another important aspect of OR gene expression is zonal expression, where each OR gene is expressed in a restricted area (zone) of the OE. In situ hybridization experiments initially showed that four distinct zones were present in MOE and a given area may only express a subset of OR genes which are specific for that zone (Ressler *et al*., 1993; Vassar *et al*., 1994). Subsequent experiments disproved the four zone model and showed a much higher number of zones (Miyamichi *et al*., 2005). These results suggested that expression zones are continuous

and overlapping (Norlin *et al*., 2001; Iwema *et al*., 2003; Miyamichi *et al*.,2005). Similarly, in zebrafish four overlapping expression rings are observed by using four different OR probes, suggesting a zebrafish analogue of zonal restriction (Weth *et al*., 1996).

1.3. Transcriptional Regulation of OR Gene Expression

In order to explain monogenic and monoallelic expression of OR genes, various mechanisms have been offered involving long range elements, proximal promoter elements, negative feedback signals and most recently, epigenetic mechanisms.

1.3.1. Long Range Elements

In 2002**,** a study investigating homologies between human and mouse OR clusters identified a remarkably long region of homology between the two species (Serizawa et al., 2003). This 2 kb H region is located 75kb upstream of Mouse MOR28 gene and has important regulatory function in MOR28 expression (Serizawa *et al*., 2003). A yeast artificial chomosome (YAC) transgenic line expressed the transgenic MOR28 only in the presence of H (Serizawa et al., 2003). Furthermore, relocating H region closer to the MOR28 cluster, caused an increase in the number of OSNs that express MOR28 (Serizawa, 2003). These result suggested that the H-element acts as a *cis*-acting locus control region (LCR) which ensures the selection and expression of ORs from a nearby gene cluster. The cis-regulatory function of the H element was later confirmed in gene targeted mice (Fuss *et al*., 2007, Nishizumi *et al*., 2007).

In 2009, a second candidate cis-acting LCR, was identified in the P2 cluster by its homology with the promoter of its adjacent OR gene P3 (Bozza *et al*, 2009). Similar to the H

region, deletion of P resulted in loss of expression of proximal OR genes (Khan et al., 2011). Thus, it was suggested that P and H elements regulate the probability of OR gene choice in a critical manner (Khan et al., 2011).

Recently, 35 OR-linked intergenic sequences were proposed as possible LCRs based on their similarity to the epigenetic properties of H element (Markenscoff-Papadimitriou *et al*., 2014). Reporter screens in zebrafish indicated that, 12 of the 32 tested elements were functional in OR gene regulation in OSNs. Furthermore, three of these elements (Lipsi, Sfaktiria, Kefallonia) showed activity as enhancers in mouse (Markenscoff-Papadimitriou *et al*., 2014).

Similar elements may exist in zebrafish (Nishizumi et al., 2007). Two regions on zebrafish chromosome 15 were shown to be critical for expression of adjacent OR genes (Nishizumi *et al*., 2007).

1.3.2. Promoter Architecture of The OR genes

The main obstacle to the investigation of promoter regions and to the detection of proximal promoter elements is the absence of proper characterization of OR transcript structures. Since OR gene repertoire is very large in many species, new OR genes are mostly identified by bioinformatics methods based on sequence similarity. A protein coding gene was considered as OR gene if its coding region was around 1 kb and if it contained the OR motifs in expected positions (Zhang and Firestein, 2002). However, these approaches resulted annotations of OR genes without their untranslated regions (UTRs), therefore ended up with the lack of information regarding transcription start sites (TSSs) and the OR transcript structure.

Until now, several studies investigated the transcript structure and promoter regions of OR genes based on 5'-RACE and RLM RACE experiments (Hoppe *et al*., 2000; Hoppe *et al*., 2003; Bulger *et al*., 2000; Michaloski *et al*,. 2006; Michaloski *et al*., 2011; Clowney *et al*., 2011). 5'-RACE and bioinformatics based approaches were performed to investigate OR37 (OR262) subfamily promoter regions an intron-exon structure of 5 members of this genome family were predicted. Most of the transcripts contained 5' introns, 5' noncoding exons and an intronless 3' UTR. Also, alternative transcipts for two OR genes are predicted. TSSs were found in a range varying between 2000-4000 bp upstream of the translational start site. Furthermore, several conserved regions (two A-T rich blocks, two G-A rich blocks and two blocks with a conserved "TCCCA" motif) observed in promoters of these gene as a result of local sequence alignments (Hoppe et al., 2000). Then, similar analysis were performed to identify transcript structures of the remaining six members of OR262 family. All genes contained several exon and four of them were alternatively spliced. TSSs were located between approximately 750 bp and 7200 bp upstream of translation start. Comparison of 1 kb upstream of TSSs resulted in identification of conserved sequence regions in the OR promoters. In order to test the function of these motifs, electromobilty shift assays were performed with proteins extracted from OE against a motif block sequence and an interaction was observed. Then, yeast one-hybrid experiments identified interaction between the conserved sequences and a set of transcription factors including Ptx1, BEN, O/E-2, Alx-3, Lhx2 and AP-2. However, only two of the factors O/E-2 (Olfactory neuron specific factor 1/ early B-cell-factor-like-2) and Lhx2 (Limhomeobox-2 factor) could be visualized by in situ hybridization on the OE (Hoppe *et al*., 2003). Interestingly, these two TF motifs are implicated in promoters of OR genes in later studies (Plessy *et al*., 2012; Michaloski *et al*., 2006; Hirota and Mombaerts, 2004).

A comprehensive analysis which investigated the promoter regions of 198 OR genes resulted in identification and spatial preference of O/E–like motifs in most of the upstream regions (Michaloski et al., 2006). Transcripts structure and TSSs of 198 OR genes were determined by RLM-RACE, and investigation of common promoter elements were carried in their corresponding promoters. Motifs resembling the O/E-like sites were mostly identified

within 200 bp upstream of TSSs. Furthermore, it was shown that these motifs interact with nucleus proteins obtained from the OE (Michaloski et al., 2006).

In 2011, a high-throuhput approach based on RLM-RACE and DNA microarray was performed to map the TSSs of 1085 (of 1400) Mouse OR genes and to reveal conserved motifs among these promoters such as homeodomain and homeoboxes (Lhx and Dhx family members) and previously unobserved motifs which belongs to BRN, ARID, and NKX6 families (Clowney *et al*., 2011).

In the most comprehensive study to date, OR promoters of mice were investigated with nanoCAGE technology to identify cis-regulatory elements and transcription factors that might regulate the OR gene expression (Plessy *et al*., 2012). nanoCAGE is a recently developed method which has the advantage of having minimum requirement of total RNA amount compared to other methods (Plessy *et al*., 2010). Map and architecture of the promoters were elucidated for 87.5% of the mouse OR genes by using nanoCAGE technology. It has been observed that OR genes mostly had a non-coding first exon and median distance from translation start site to TSS were 3125 bp. In 21% of OR promoters, a well-defined TATA-box motif was detected at expected location -33 to -29. EBF1(O/E-like) motif was shown to have preference for positions between 50 and 150 bp upstream. Also, homeobox binding motifs were observed to accumulate at 100–150 bp upstream of the TSSs. Furthermore, binding of TBP, EBF1 and MEF2A transciption factors to OR promoters were identified by chromatin immunoprecipitation (Plessy *et al*., 2012).

In contrast to the dependence on long-range regulatory elements certain OR genes have been shown to be controlled largely by short range promoter elements. In those cases, short transgenic constructs were able to drive OR expression similar to the endogenous gene (Qasba and Reed, 1998; Vassali *et al*., 2002; Rothman *et al*., 2005; Vassali *et al*., 2011). A comprehensive analyis of the 161 bp M71 minimal proximal promoter identified homeodomain

and O/E-like binding sites (Rothman *et al*., 2005; Vassalli et al., 2011). Mutagenesis experiments showed that these binding sites are required for OR gene expression, suggesting that homeodomain and *olf-1* (O/E-like) transcription factors have a critical role in OR gene expression. A comprehensive bioinformatic investigation of OR gene promoters showed the presence of these homeodomain and O/E-like binding sites in most OR promoters (Hoppe *et al*., 2006; Michaloski *et al*., 2006; Michaloski et al, 2011; Vassali *et al*, 2011; Plessy *et al*, 2012). Curiously, O/E-like binding sites are also present in many olfactory-specific genes, for instance Golf, Adenylyl cyclase III (ACIII), olfactory cyclic nucleotide-gated channel, and OMP (Wang et al, 1993). Interestingly, it has also been shown that the H and P regions contain homeodomain and O/E-like binding sites (Hirota et al., 2007; Nishizumi *et al*., 2007; Vassali *et al*., 2011). Furthermore, mutations in these sites caused abolishment of OR gene expression in transgenic animals (Nishizumi *et al*., 2007).

Up until now, two homeodomain transcription factors have been shown to play a role in OR gene expression, Lhx2 and Emx2 (Hirota and Mombaerts, 2004; McIntyre *et al*., 2008). Lhx2, a LIM-homeodomain protein, was identified by its ability to bind to the M71 promoter (Hirota and Mombaerts, 2004). Lhx2 knock-out mice showed defects in OSN development indicating a potential role of Lhx2 in this process (Hirota and Mombaerts, 2004; Kolterud 2004). Subsequent studies indicated that Lhx2 is required for expression of Class II ORs while Class I ORs remain mostly unaffected in knock-out mice (Hirota *et al*., 2007). Therefore, it can be argued that Lhx2 has a critical role in Class I OR gene expression but not in Class II ORs.

Another protein which was implicated in OR gene regulation is Emx2 homeobox transcription factor (McIntyre *et al*., 2008). Emx2 also binds to the M71 promoter region (Hirota and Mombaerts, 2004) and is expressed in the OE (Nedelec *et al*., 2004). No OE development defect was observed in Emx2 knock-out mice, however reduction in number of mature OSNs was detected (McIntyre et al., 2008). Furthermore, reduction in the expression of many OR genes which was greater than reduction in number of OSNs (42 %) suggested that Emx2 have a role in transcriptional regulation of ORs (McIntyre et al., 2008).

1.3.3. Negative Feedback Mechanism

A negative feedback mechanism which stabilizes and maintains monogenic OR expression was proposed by Serizawa and colleagues, stating that a functional OR protein might be triggering a negative feedback signal that can prevents expression of any other OR gene and by this way maintains the singular expression of ORs (Serizawa *et al*., 2003). In order to test this hypothesis, the endogenous MOR28 gene was replaced with a sequence coding for a fluorescent protein and it was observed that OSNs which express the fluorescent reporter gene also express other ORs. A frameshift mutation in MOR28 resulted in the same phenotype, suggesting that the protein not the transcript is required to initiate the negative feedback signal (Serizawa *et al*., 2003). Similar observations were made for the M4 (Lewcock and Reed, 2004), P2 (Shykind *et al*., 2001) and SR1 genes (Fuss *et al*., 2012).

In a recent study, activating $G_{\beta\gamma}$ signaling through the OR protein resulted in decrease of expression in the given OR genes. Furthermore, $G_{\beta\gamma}$ signaling inhibition caused increase in the number of OSNs which express a given or and also triggered the multigenic expression of ORs (Ferreira *et al*., 2014). Finally, RNA-Seq experiments showed that $G_{\beta\gamma}$ signaling influences the expression of histone modifying enzymes which are responsible for repressive histone methylation marks (Ferreira *et al*., 2014). This study suggest a possible link between OR protein function and epigenetic mechanisms.

In zebrafish, a bacterial artificial chromosome (BAC) transgenic line which includes a cluster of OR genes has been used to investigate the negative feedback mechanism (Sato *et al*., 2007). The BAC comprised 16 OR genes from three different subfamilies and two OR genes, OR 103-1 and OR111-7, were substituted with fluorescent proteins. In situ experiments showed that OSNs that express fluorescent proteins also express another OR gene indicating that second choice mechanism exists in zebrafish. Interestingly, the expressed ORs were inform the same

subfamily on the same cluster. This restriction is different from mouse where second choice mostly occurs among ORs from other chromosomes (Sato *et al*., 2007).

1.3.4. Epigenetic Mechanisms

Recent studies have shown epigenetic regulation of OR expression. In immature OSNs, OR loci are silenced in regions of constitutive heterochromatin (Magklara *et al*., 2011). Presence of heterochromatin markers such as H3K9 methylation in the OR loci led to investigation of the enzymes that might regulate methylation processes (Magklara *et al*., 2011). It was shown that LSD1, a lysine-specific demethylase, has a dual function in OR gene expression as both coactivator and corepressor (Lyons *et al*., 2013). Also, a second group of histone methyltransferases G9a (KMT1C) and GLP (KMT1D) were shown to be functional in OR expression by pharmacological inhibition of these enzymes in zebrafish (Ferreira *et al*., 2014). Furthermore, deletion of G9a/GLP in mice resulted in disruption of "one receptor one neuron" hypothesis (Lyons *et al*., 2014).

Dalton and colleagues proposed that activation of UPR in OSNs trigger a cascade, which eventually prevents expression of more than one OR per OSN. This negative feedback model is based on the observation that upon OR expression activation of Perk and phosphorylation of eif2a occurs in OSNs as a result of UPR. Then, translation of activating transcription factor 5 (ATF5) and transcription of Adcy3 downregulate Lsd-1 activity and maintain the initial OR choice by preventing further OR expression (Dalton *et al*., 2013). The epigenetic mechanisms which are shown to be involved in singular gene expression in OSNs made the OR gene expression process even more complicated and suggested that proper understanding of this process also relies on studies in cellular level.

1.4. RNA Sequencing for Differential Gene and Transcription Expression Analysis

1.4.1. TopHat and Cufflinks

High throughput mRNA sequencing (RNA-Seq) is an experimental technique which allows researchers to discover new genes and transcripts and to quantify levels of expression (Mortazavi *et al*., 2008; Cloonan *et al*., 2008; Nagalakshmi *et al*., 2008). In order to analyze enormous and complex data yielded from RNA-Seq experiments, several bioinformatics software has been developed (Garber *et al*., 2011). These software need to be robust, efficient and statistically principled and should be able to perform three main tasks read alignment, transcript assembly or genome annotation, transcript and gene quantification (Trapnell *et al*., 2012). Two bioinformatics tools are able to perform all three tasks when they are performed in concert: TopHat and Cufflinks. TopHat (http://tophat.cbcb.umd.edu/) is capable of alignment of reads to the genome and identification of transcript splice sites (Trapnell *et al*.2009). Output of TopHat is considered as input for Cufflinks (http://cufflinks.cbcb.umd.edu/) which uses these alignments as a map for assembling the reads into the transcripts (Trapnell *et al*., 2010). There are several considerations prior to use of TopHat and Cufflinks. First, a sequenced (reference) genome is required. Second, RNA-Seq should be performed with either Illumina or SOLiD Sequencing machines. Lastly, UNIX shell is necessary for running both software (Trapnell *et al*., 2012).

The most critical step of RNA-Seq is the alignment of sequencing reads to a reference genome. Therefore, an efficient alignment tool strengthens the quality of whole RNA-Seq experiment and one of the most efficient software is called as Bowtie (Langmead *et al*., 2009). The only limitation of Bowtie is its inability to align a read which contains a large gap into the genome, thus making it incapable of aligning reads that contain introns. However, TopHat, which uses Bowtie as an alignment “engine”, overcomes this problem by breaking alignments into smaller parts called “segments”. If some of the segments align far apart from each other,

TopHat recognizes the presence of a splice junction and estimates splice sites position of the read (Trapnell *et al*., 2012).

In order to quantify gene and transcript expression, alignments of RNA-seq can be used since there is a correlation between the number of reads belong to a transcript and its expression level. The most accurate quantification relies on identification of the proper isoform of a given gene. Cufflinks assembles individual transcripts for each isoform, then these assemblies are merged by using the Cuffmerge utility. The merged assembly is required for providing a uniform background for calculating gene and transcript expression (Trapnell *et al*., 2012). Even though the number of RNA-Seq reads is directly proportional to relative abundance of given transcript in the sample, further normalization steps needed for proper calculation of expression level. The main reason for this requirement is the fact that library construction process perform size selection for cDNA fragments, resulting more sequencing fragments for longer transcripts (Trapnell *et al*., 2012). A statistical model called fragments per kilobase of transcript per million mapped fragmens (FPKM) ensures normalization of data for unbiased comparison of expression levels (Trapnell *et al*., 2010). Bigger FPKM value indicates higher expression of a given transcript. The main disadvantage of this model is the requirement of transcript structure prediction for calculating FPKM levels. The genes that have lower expression levels might not have enough amount of sequencing reads for prediction of their transcript structure. Thus, FPKM value for these genes might not be calculated. In order to overcome this problem increasing the depth of sequencing is required.

1.4.2. Olfactory Transcriptome

In recent years, several RNA-Seq based studies were performed to investigate the olfactory transcriptome of mice and zebrafish. In a study which employed deep RNA sequencing in mouse, expression microarray and quantitative RT-PCR, transcriptome profile of

both olfactory mucosa and vomeronasal organ were determined (Ibarra-Sorria *et al*., 2014). Evidence of expression was found in for whole VR repertoire and nearly all OR genes that were annotated as functional in database. Also, new and multi-exonic annotations were generated for over 1100 receptor genes. OR and VR genes both had reproducible distribution of expression and neither of them expressed equally or randomly (Ibarra-Sorria *et al*., 2014).

Another study which also investigated the mouse olfactory transcriptome employed deep RNA sequencing and identified nearly all OR and TAAR genes previously annotated as functional (Kanageswaran *et al*., 2015). In the group of the most highly expressed 200 genes in OE, the genes that participate in olfactory signaling pathways were found. Expression profile of OSNs were compared with different mouse tissues in order to identify OE specific genes (Kanageswaran *et al*., 2015).

In a very recent paper, olfactory transcriptome of zebrafish was produced and analyzed by RNA sequencing (Saraiva *et al*., 2015). Since zebrafish lacks VNO, comparison of mouse and zebrafish transcriptome was performed to investigate the relationship between these two physiologically different olfactory systems in biochemical and evolutionary level. Enormous molecular similarity between mouse and zebrafish chemosensory receptor classes (OR,Taar,V1r,V2r and Gucy) and OSN specific markers were observed since expression of orthologs of these mouse genes could be detected in zebrafish OE. Furthermore, it was demonstrated that chemosensory receptor expression levels correlated with number of OSNs expressing given chemosensory receptor (Saraiva *et al*., 2015). These studies indicated the adequacy of utilizing RNA sequencing for comprehensive investigation of chemosensory gene families compared to other expression analysis techniques such as in situ hybridization, quantitative RT-PCR and microarray.

1.5. Regulatory Sequence Analysis Tool for Motif Discovery

Non-coding DNA sequences have a critical function in biological systems by regulating the gene transcription by means of spatial and temporal control. The *cis*-regulatory elements are the site where interactions between transcription factor (TF) proteins and their target genes occur. Therefore, detecting novel cis-regulatory elements and their interacting partners is an important aspect of understanding transcriptional regulation.

Regulatory Sequence Analysis Tools (RSAT) (http://rsat.ulb.ac.be/rsat) is a suite which is composed of collection of software tools for the the prediction of *cis*-regulatory sites in non-coding DNA sequences. Several programs are included in this suite with various functions: Sequence retrieval, pattern discovery, phylogenetic footprint detection, pattern matching, genome scanning and feature map drawing. In order to provide statistical rigidity, RSAT allows random controls which can be performed by random gene selections or by generating random sequences by using statistical background models such as Bernoulli or Markov. In addition to word-based pattern discovery tools, RSAT includes statistically powerful matrix-based scanning tools.

2. PURPOSE

The aim of this study is to detect regulatory DNA motifs within OR gene promoters of zebrafish which might regulate OR gene expression. It has previously been shown that OR promoters of mice contain several binding motifs such as Ebf1, Lhx2 and Emx2. In order to perform unsupervised motif search and investigate the presence of previously proposed TF binding motifs, combination of transcriptome analysis and bioinformatic tools was employed. To pinpoint the TSS of OR genes, RNA-Sequencing performed against zebrafish olfactory epithelium tissue. Bioinformatic tools were utilized to identify presence of *de novo* and previously known motifs in OR promoters and to detect their positional preference in relation to TSS. Then, promoters of highly expressed gene family were investigated to identify an enriched candidate regulatory motif.

3.MATERIALS AND METHODS

3.1. Materials

3.1.1. Fish

Zebrafish (*Danio rerio)* of the AB/AB and AB/Tü strains obtained from the Zebrafish International Resource Center (ZIRC) at the University of Oregon, USA were used in this study. For in situ hybridization, animals maintained at the zebrafish facility at the Boğaziçi University Life Sciences Center (Vivarium) were used; for RNA-seq animals of the AB/AB background were obtained and processed at the University of Münster, Germany.

3.1.2. Equipment and Supplies

The list of chemicals, equipments, and consumables can be found in Appendix A and Appendix B.

3.1.3. Buffers and Solutions

The buffers and solutions for common molecular biology procedures were either obtained from manufacturers or prepared according to the the instructions in Sambrook and Russell (2001). Zebrafish specific buffers and solutions were prepared according to Westerfield (1997).

3.2. Methods

3.2.1. Maintenance and Breeding of Fish

Zebrafish strains (AB/AB, AB/Tü) were kept at 28ºC under a 14 hours light / 10 hours dark light cycle at appropriate housing densities in a Stand Alone Zebrafish Housing System (Aquatic Habitats) at the Vivarium of Bogazici University Life Sciences Center. Individual tanks were connected to the housing system with aeration, UV sterilization and five stage filtration capacities. System water was prepared by mixing 2 g of sea salt, 7.5 g sodium bicarbonate, and 0.84 g of calcium sulphate in 100 liters of reverse osmosis water. Adult zebrafish were fed three times a day, twice with live brine shrimp (*Artemia sp*.) and once with flake food (TetraMin, Sera Vipan).

3.2.2. Polymerase Chain Reaction (PCR)

Polymerase chain reactions were performed using the GoTaq Flexi DNA Polymerase (Promega), OneTaq (NEB), or Advantage Taq (Clonetech) PCR kits according to the manufacturer's instructions. For standard PCR reactions, 100 ng of plasmid DNA, 0,5 μM of forward and reverse primer, 1x reaction buffer, 1.5 mM $MgCl_2$ (if not supplied in the reaction buffer), 0.2 mM dNTP mix and 1-2 units of Taq polymerase were used. The PCR conditions were set to a 3 min initial denaturation step at 95˚C, 24-30 cycles of 45 s denaturation at 95ºC, 30 s annealing at the appropriate annealing temperature (4 degree lower than the melting temperature of the lowest melting oligonucleotide) and 1 min / 1 kb target amplicon at 72°C, followed by a final elongation step of 10 min at 72˚C. For colony PCR reactions, which were performed for identification of positive clones after ligation of DNA fragments to vector backbones, the cycle number was increased to 36 cycles and the initial denaturing step was extended to 10 minutes to allow for sufficient lysis of bacteria. Home-made Taq polymerase

was used. A typical reaction mixture for colony PCR contained 0,5 μM of forward and reverse primer, 1x reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTP mix and 1 μL of homemade Taq polymerase.

3.2.3. Restriction Enzyme Digestion of DNA

Restriction endonuclease enzymes from Promega, Invitrogen or New England Biolabs (NEB) were used for restriction digestions. In a typical reaction mixture, 3-6 units of restriction enzyme was used per microgram of DNA in a reaction mixture containing appropriate buffer (1x concentration) and 1x BSA if if required. Digestion reactions were incubated at 37ºC for 1 to 8 h.

3.2.4. Agarose Gel Electrophoresis

DNA fragments were analyzed by agarose gel electrophoresis using 1% agarose gels (in 1x TAE) supplemented with 0.5 μg/ml Ethidium bromide. 1x TAE was used as running buffer at 60 – 100 V. Gels were visualized under UV light in a GelDoc XR (Bio-Rad Labs, USA) system and stored electronically as TIF images. The 1kb and 100bp DNA ladders (NEB, USA) were used as molecular weight markers.

3.2.5. PCR and Gel Purification

The High Pure PCR Purification Kit (Roche, USA) was used to purify DNA fragments from agarose gels. For gel extraction, the desired DNA fragment was cut out from the gel using a scalpel and weighed. Per 0.1 g of agarose gel, 300 μl of binding buffer was added and the mixture was incubated at 56ºC for 10 minutes to melt the agarose. Then, 150 μl isopropanol

were added per 0.1g of agarose gel and the mixture was loaded to spin columns provided by the kit, washed, and eluted using wash and elution buffers from the kit. Quantification of eluted DNA performed by using the NanoDrop Spectrophotometer and DNA fragment visualized on agarose gels.

3.2.6. Ligation of DNA Fragments to Vectors

To ligate DNA fragments to vector backbones typically a 3:1 or 1:1 molar ratio of insert to vector was used. The total combined amount of DNA (vector and insert) was kept below 10ng per µl. A typical ligation reaction consisted of 1 µl of 10x Ligase buffer in a final reaction volume of 10 µl. The reaction mixture was incubated at 25ºC for 1 hour and transformed into competent cells.

The pGEM-T Easy (Promega) vector system was used for direct ligation of PCR products into vector plasmids. For a typical reaction, up to 3.5 µl purified PCR product, 0.5 µl pGEM-T Easy vector, 5 µl of 2x ligase buffer, 1 µl of T4 DNA ligase (NEB) and dH_2O to 10µl total reaction volume were combined. Incubation was performed at 25ºC for 1 hour and the reaction was transformed into competent *E. coli*.

3.2.7. Transformation of Plasmid DNA into Competent Cells

For each transformation reaction 50 µL competent cells were thawed on ice for 5 min, mixed with up to 10 µl of plasmid DNA or ligation product and incubated on ice for 30 min. Typically, 10-50 ng of plasmid DNA for retransformations and 10 µl of ligation reaction were used for a transformation. Following incubation on ice, the tubes were transferred to a heat block adjusted to 42ºC and incubated for 90 sec, followed by incubation on ice for 5 minutes to

recover. The transformed bacteria were resuspended in 500 µl of LB medium and incubated for 1 hour at 37ºC on a shaking platform. 100 µl of the transformation reaction were spread on LB agar plates with appropriate antibiotics. The remaining reaction mixture was centrifuged for 1 minute at maximum speed, the supernatant was removed by decanting and the cells were resuspended with the remaining supernatant before being spread on selection plates. Bacteria plates were incubated at 37ºC overnight.

3.2.8. Plasmid Isolation

For small scale preparations of up to 20 µg the Plasmid MiniGeneJet Isolation kit (Thermo Scientific) and for larger samples of up to 100 µg, the Qiagen midi kit were used. Plasmid DNA was isolated from bacteria according to the protocol provided by manufacturer.

3.2.9. In Situ Hybridization

For expression analysis by RNA in situ hybridization, olfactory epithelia were dissected and embedded in OCT mounting medium and frozen at -20C. Samples were cut at 12µm thickness on a LEICA CM3050S crystat, dried at 60C and stored at -20C until use.

The section were fixed with 4% PFA (in 1x PBS) for 10 min, washed in PBS for 5 min, treated with 0.2 M HCl for 10 minutes, digested with 1 ug/ml Proteinase K in 0.1 M Tris-Cl for 10 min, and TEA for 10 min. Slides were washed in 1x PBS for at least 5 min between individual processing steps. For hybridization, section were covered with Hybridization Mixture (HM, 50% Formamide, 5X SSC, 0.1% Heparin, 5% yeast RNA 0.05% Tween 20, 1% Citric Acid and the RNA Probe), covered with cover slip and hybridized at 65C in a moist chamber (50% Formamide, 5X SSC) over night. Following hybridization, consecutive series of 10 min washes

at 65 ºC performed to gradually change from HM to 2X SSC (75% HM . 2x SSC, 50% HM / 2x SSC, 25% HM / 2x SSC, 2x SSC). Then two washes with 0.2 X SSC were performed for 30 min at 65 ºC and the 0.2 SSC solution was gradually replaced with PBST by a series of 10 min steps at room temperature (75% 0.2 x SSC/ PBST, 50% 0.2 x SSC/ PBST, 25% 0.2 x SSC/ PBST, PBST). Aftrer an additional 5 min wash with PBST, slides were incubated with 0.5% Blocking Reagent in PBST for 1-3 hours and with anti-DIG-AP in 0.5% Blocking Reagent (1:750) at 4 ºC overnight. Following antibody incubation, sections were washed with PBST three times for 15 minutes and Detection Buffer (100Mm NaCl, 100mM Tris-Cl pH=8,10 mM Magnesium Chloride) for 3 min at room temperature. Then, sections were covered with 250 µl of (10 µl 25 mg/ml Fast Red and 10 µl HNPP in 1ml Detection Buffer) for 60 min and visualized under confocal microscope.

In order to perform in situ hybridization against highly expressed OR132gene family, primers were designed against 3'-UTR regions of these genes. Then, probe sequences were obtained by PCR for four OR genes, - OR132-2, OR132-3, OR132-4, OR132-5. These sequences were cloned into pGEM-T Easy vector and their orientations were determined. Then, in vitro transcription was performed according to the instruction provided within the manual.

3.2.10. TO-PRO Staining

TO-PRO®-3 was applied to visualize nuclei after ISH. 1:500 volume of TO-PRO in PBS was applied for 30 minutes and washed with PBS.

3.2.11. RNA-Seq and Bioinformatic Analysis

RNA-seq of olfactory epithelium and brain samples was performed at BGI, Hong Kong and preprocessed. The Bowtie Suite (Trapnell *et al*., 2012) was used for local realignment of

the sequencing reads to the Zv9 Zebrafish Genome. The TopHat (Trapnell *et al*., 2012) algorithm was used for remapping of the reads. The Cufflinks (Trapnell *et al*., 2012) program was used for transcript prediction and determining rpkm values of the transcripts. We used default values for Cufflinks except multihit value 100 since our target genes are highly similar. RPKM value were calculated as follows (Trapnell *et al*., 2012).

RPKM = reads per kilobase per million

= [# of mapped reads]/[length of transcript in kilo base]/[million mapped reads]

= [# of mapped reads]/([length of transcript]/1000)/([total reads]/10^6)

In order to extract upstream sequences of OR genes, the TSS was determined by visual inspection of transcript structure and mapped reads to the zebrafish genome in IGV and the first 2000 and 500 bp upstream of the TSS were selected.RSAT Suite (Turatsinze *et al*.,2008) was chosen for investigation of enriched motifs on OR promoter regions. Oligo Analysis (van Helden *et al*., 1998) was used in order to detect enriched oligomers in OR promoters. Matrix Scan (Turatsinze *et al*.,2008) was used to identify presence and distribution of known motifs on Or promoters. *Danio Rerio* was chosen as the background model in all of the analysis. Statistical thresholds were always considered. PWMs are obtained from Jaspar (Mathelier *et al*., 2014) and UniProbe (Newburger and Bulyk, 2009) database. TOMTOM tool (Tanaka *et al*., 2011) from MEME Suite (Bailey *et al*., 2009) were used to detect similarity of identified motifs to the PWMs in database. Statistical threshold is determined as 4 for each RSAT- Matrix Scan Analysis. For de novo Motif finding analysis Oligo Analysis (van Helden *et al*, 1998) performed with setings with oligomer length as 6,7 and 8. Obtained PWM from this analysis used as input for TOMTOM tool in MEME Suite.

3.2.12. Linear Regression Analysis

Linear Regression Analysis performed by using Graph Pad to investigate the strength of the association between fpkm values and cell counts.

4. RESULTS

4.1. RNA Sequencing Outline

Major aim of the studies presented in this thesis was to identify regulatory motifs and their corresponding transcription factors in zebrafish OR gene promoter regions. Similar to previous studies in mice (Michaloski 2006, Michaloski 2011, Plessy *et al*., 2012), it was intended to identify conserved DNA sequence motifs located near or around the TSS of OR genes by bioinformatic analysis. In addition, using *ab initio* motif finding algorithms, putative OR promoter regions were scanned for the presence of known and suspected binding motifs. Because OR gene expression is strictly regulated and complex process that requires molecular coordination among hundreds to thousands of different genomic loci, it was expected that this phenomenon may, at least in part, be reflected at the level of OR promoters and that common promoter elements may be present for all or specific OR subsets.

Therefore, the primary objective of this study was to identify candidate OR promoter sequences and to find conserved sequence motifs within these sequences using unsupervized bioinformatic tools. To guide the bioinformatic analysis, several assumptions regarding the presence and characteristics of presumptive regulatory motifs were made. It was expect that candidate regulatory motifs locate in proximity (upstream or downstream) of the TSS of OR transcripts. It was also expected that motifs guiding olfactory tissue-specific expression should be conserved in all or most OR gene promoters. On the other hand, factors that govern spatial or temporal profiles of OR subsets would be present in the promoters of coordinately expressed genes. The identified candidate sequences would then provide the basis to guide further biochemical and transgenic studies to elucidate the role of these factors in OR gene expression.

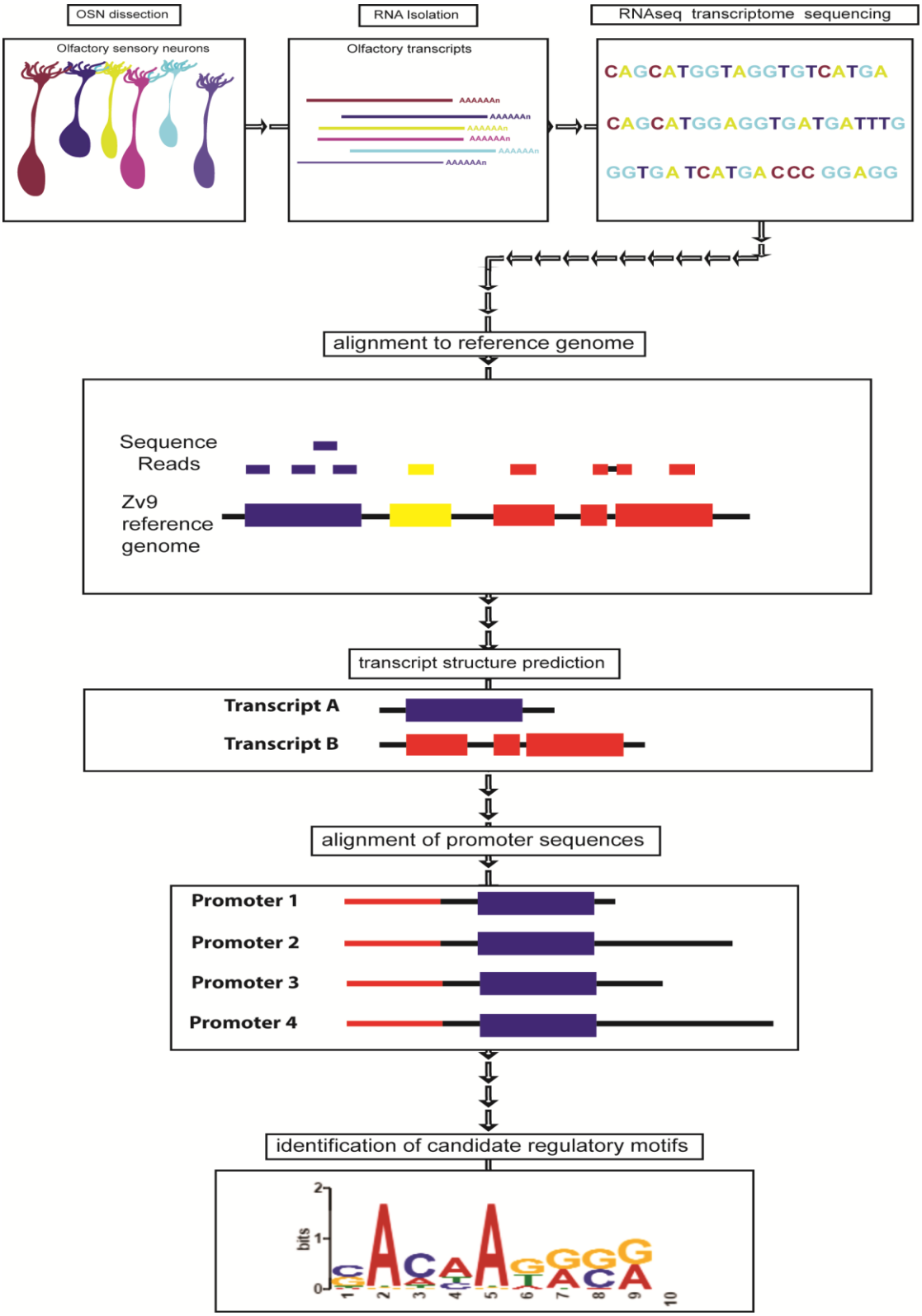

Figure 4. 1. Experimental workflow for the identification of candidate regulatory motifs within OR gene promoters by transcriptome analysis and bioinformatic search.

The general strategy for this approach included the isolation of total RNA from OSNs using pools of OE from a large number of animals, followed by RNA sequencing (RNA-seq). Subsequent mapping of RNA-seq reads to the Zv9 zebrafish genome allows for the unambiguous experimental determination of TSSs, including the TSSs of all OR genes, and extraction of candidate promoter sequences. These sequences were then subjected to a bioinformatic search for conserved sequences using a variety of different tools.

4.1.1. Evaluation of Mapping Efficiency

RNA-seq was performed in two runs, an initial run with 4 GB depth to estimate the required depth of sequencing to obtain reliable structural data on OR transcripts suing OE and brain tissue, followed by a second run with an additional 8 GB depth on OE tissue to increase the quality of OR transcript structures.

The initial run of RNA-Seq provided by BGI resulted in 52 million paired reads (4.7 GB, average read length of 90 bases) for the OE sample and 49 million paired reads (4.4 GB, average read length of 90 bases) for the brain sample. Sequencing reads of both samples were then aligned to the Zv9/DanRer7 reference genome and resulted in the mapping of 76% of all reads (40 million reads) to the genome for the OE sample. Of those, 35 million reads (or 87.5 %) could be mapped to annotated genes. Alignment of brain sample to the Genome Researchulted in mapping of 80% of all reads (39 million reads), 30 million (or 77 %) of which could be mapped to annotated zebrafish genes.

After the second RNA-seq run, the combined sequencing reads were remapped to the Zv9/DanRer7 reference genome using local computing power on the Bogazici University Genome Server (http://trgenom.bio.boun.edu.tr) and Tophat algorithms (http://tophat.cbcb.umd.edu/; Trapnell *et al*., 2012). Further analysis, such as transcript structure

assembly and determination of expression levels, was performed using the Cufflinks suite (http://cufflinks.cbcb.umd.edu/; Trapnell *et al*., 2012) and results were visualized using the Integrated Genome Viewer (IGV, http://www.broadinstitute.org/igv/; Thorvaldsdóttir *et al*., 2013).

Changing the parameters of alignment during mapping of RNA-seq reads to the genome (e.g allowing multi hits, changing accepted mismatch values) increased the percentage of mapped reads. Yet, because this manipulations might also increase the number of false positive mapping to the genome, we used common parameters except multihits for the subsequent remapping since some of the OR gene sequences share high sequence similarity. Thus, for final remapping of the dataset, the mismatch number was set to default (2), while the maximum number of multi hits was increased to 100 instead of the default value of 20.

An important factor skewing the apparent mapping efficiency to zebrafish genes is the fact that the current zebrafish genome assembly lacks annotations for a large number of genes, including OR genes. Thus, the real overall number of reads mapping to identifiable transcripts is much higher. A total of 57,892 and 54,704 transcripts could be identified by de-novo transcript assembly using cufflinks algorithms from OE and brain samples, respectively. The zebrafish genome contains about 36.000 genes (Howe *et al*. 2013) suggesting that a high number of alternatively spliced transcripts and noncoding transcripts were uncovered.

A curious observation from the visualization of mapping data was the distribution of mapping density along transcripts where the 5'-end of annotated transcripts showed a higher coverage than the respective 3'-ends. A critical procedure that affects the read distribution along a transcript is the cDNA library preparation method. It has been shown that oligo(dT) priming causes a bias towards 3'-ends (especially in long transcripts) whereas random hexamer priming tends to generate a bias towards the 5'-ends (Hansen *et al*., 2010). In the analysis presented here, both oligo(dT) (during RNA extraction) and random hexamer priming (during library

generation) was applied in successive steps. An additional RNA fragmentation step was applied during cDNA library preparation to prevent bias towards either end (Hansen *et al*., 2010). Our visualization of mapping data and randomness assessment provided by BGI showed that our sequencing reads covers the transcripts more or less evenly except for a slight decrease in 3'-ends. Proper coverage of 5'-end of transcripts, however, is a prerequisite for the intended identification of exact OR TSSs positions.

4.1.2.Visualization of RNA Sequencing Data

In order to determine the quality of the generated transcriptome dataset, transcripts, which are specifically expressed in olfactory tissue, such as members of the olfactory signal transduction cascade and genes which have an olfaction-related function, were investigated. These genes include molecular markers for different types of OSNs, such as the olfactory marker protein (omp) gene for cillated OSNs (Celik *et al*., 2002; Sato *et al*., 2005) and the transient receptor potential family C type 2 (trpc2) gene for microvillous OSNs (Sato *et al*., 2005). In addition, other olfaction-related genes, the guanine nucleotide binding protein alpha (gnao1b), adenylate cyclase 3b (adcy3b), cnga4 (cyclic nucleotide gated channel alpha 4), ebf-1 (Olf1/Ebf1 transcription factor), and beta 2 microglobulin (b2ml) genes were analyzed. Gnao1b, adcy3b, and cnga4 were chosen because these genes are important components of olfactory signal transduction (reviewed in Restrepo and Schild, 1998), ebf1 was chosen because ebf-1 binding sites (O/E-like motifs) were shown to be enriched OR promoters and promoters of genes that are expressed in OSNs (Vassali *et al*.,2002; Hoppe *et al*., 2003; Michaloski *et al*., 2003). B2ml was chosen because of its known role in pheromone detection by forming a multimolecular complex with V2R-type vomeronasal receptors and the MHC class I gene M10 (Loconto *et al*., 2003). As expected, all of these genes were highly enriched in OE samples when compared to their expression level in the brain.

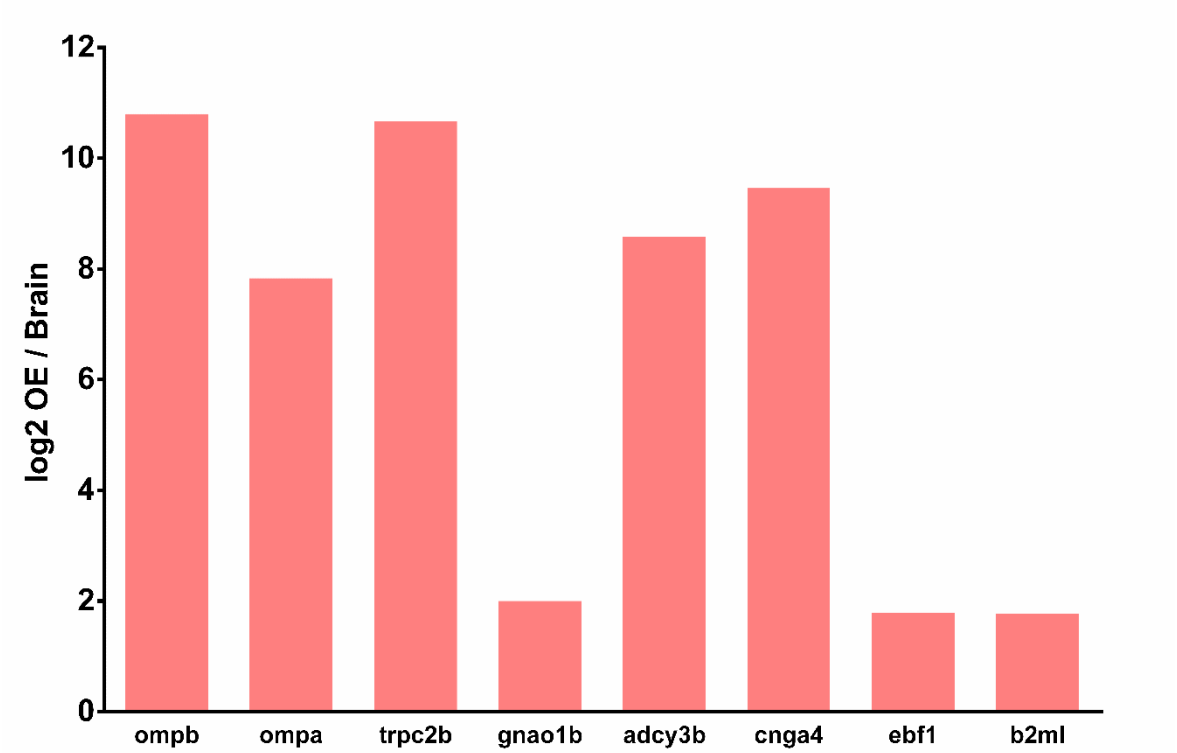

Figure 4. 2. Comparison of known olfactory genes between brain and olfactory epithelium. Olfactory marker protein a (ompa) olfactory marker protein b (ompb), transient receptor potential channel (trpc2b) guanine nucleotide binding protein alpha (gnao1b), adenylate cyclase 3b (adcy3b), cyclic nucleotide gated channel alpha 4 (cnga4), Olf1/Ebf1 transcription factor (ebf-1), and beta 2 microglobulin (b2ml).

Thus, the obtained sequencing data, in principle, allows for the identification of major constituents of the olfactory signal transduction cascade and molecular markers of OSNs. Both, the transcript structure and the relative expression levels of these genes within and between OE and brain tissue could be revealed with high accuracy.

Visual inspection of highly expressed and structurally characterized genes in the brain and OE tissue (Figure 4.3.) reveals that the transcript structure can be identified with high accuracy from the RNa-seq data. As shown for the tbp gene, which is expressed in both brain

and OE at high levels, visualization in IGV displays accurate agreement between read alignment and transcript structure from the Ensembl database. Thus, with sufficient sequencing depth and a sufficiently large number of mapped reads, it is possible to predict the exact transcript structure with high precision, including the position of the TSS.

Figure 4. 3. Visualization of the mapping of RNA-Seq reads for tbp gene. Sequencing reads and their distribution along genome can be observed. Reads are color coded according their direction. Grey region indicates the coverage distribution. Transcript structure previously annotated in Ensembl is depicted in blue.

Because the ultimate goal is to extract structural data from OR transcripts, representation of the OR repertoire was analyzed. Figure 4.4 depicts the expression levels of each OR in ascending order for each chromosome. Within chromosomes and across the repertoire, a wide dynamic range of expression can be observed. FPKM values range from < 0.1 for the lowest expressed genes to 325 for the highest expressed genes. Expression levels also vary considerably among members of the same OR gene family (e.g or111 family). The OR genes with the highest

overall expression levels collectively belong to the or132 family, which is located on chromosome 21. For these genes up to 10-fold higher expression levels were observed than for the average OR.

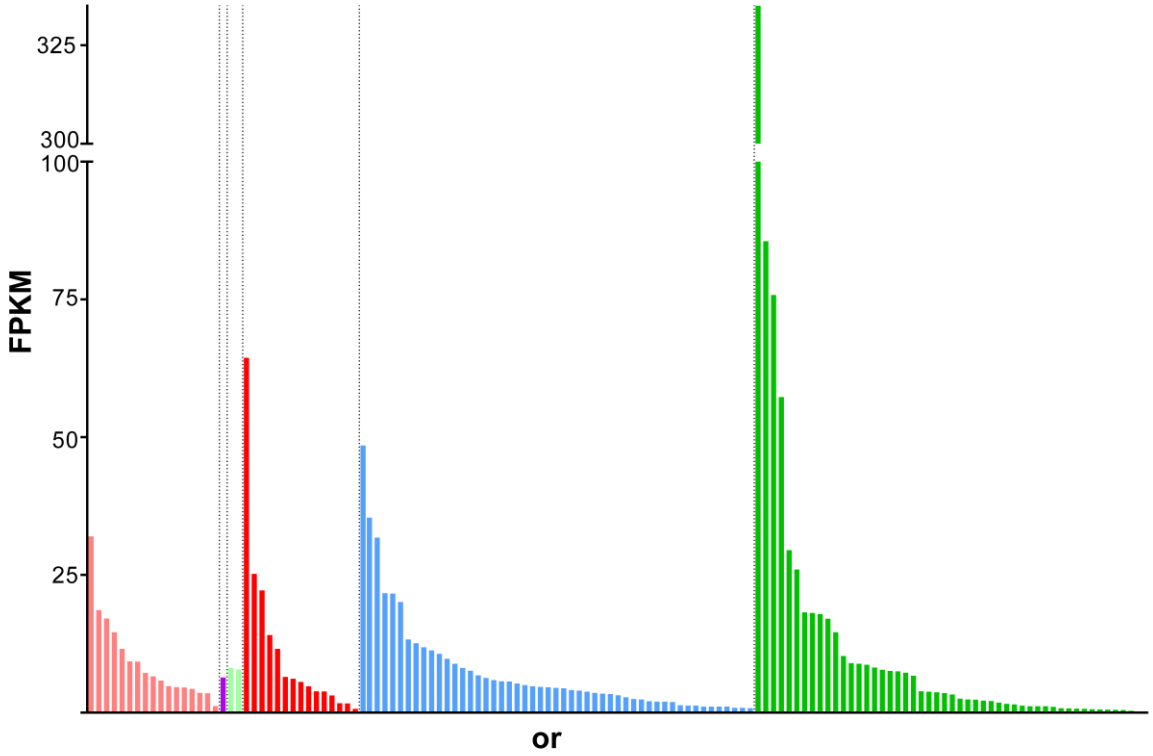

Figure 4.4. Genome wide overview of expression levels of OR repertoire. Pink is chromosome 6, purple is chromosome 7, light green is chromosome 8, red is chromosome 10, blue is chromosome 15, green is chromosome 21.

When the possibility to accurately map OR transcript structures was investigated, generally accurate OR transcript structures could be obtained for some of the known OR transcripts. Figure 4.6 exemplifies mapping of RNA-seq reads to an OR locus for the or111-2 gene. Reads are successfully mapped to the both coding region and UTR regions of the OR gene

and provided a proper transcript structure. Also, intron structure is properly determined by Cufflinks.

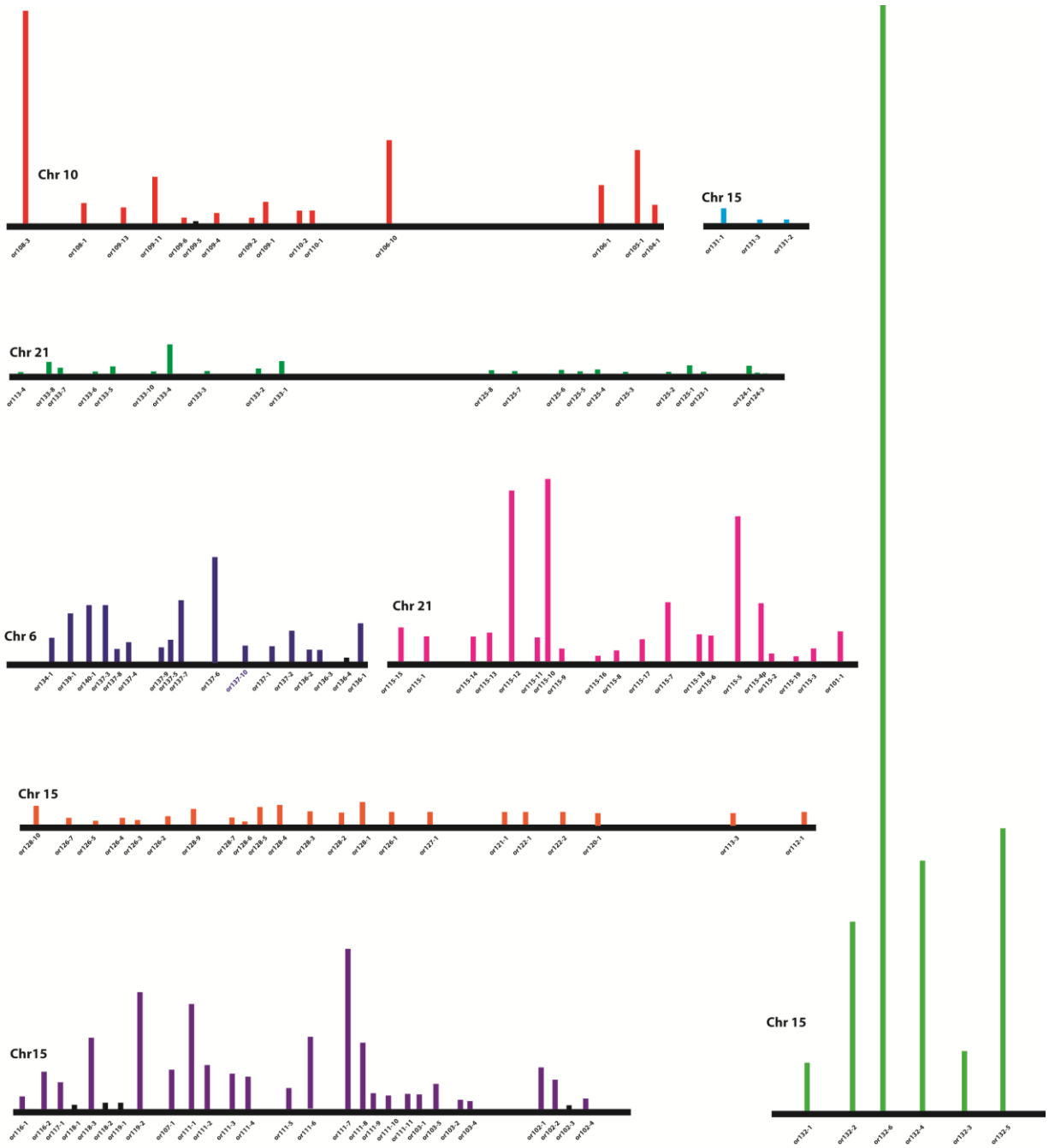

Figure 4.5. Expression levels of OR genes across the chromosomes.

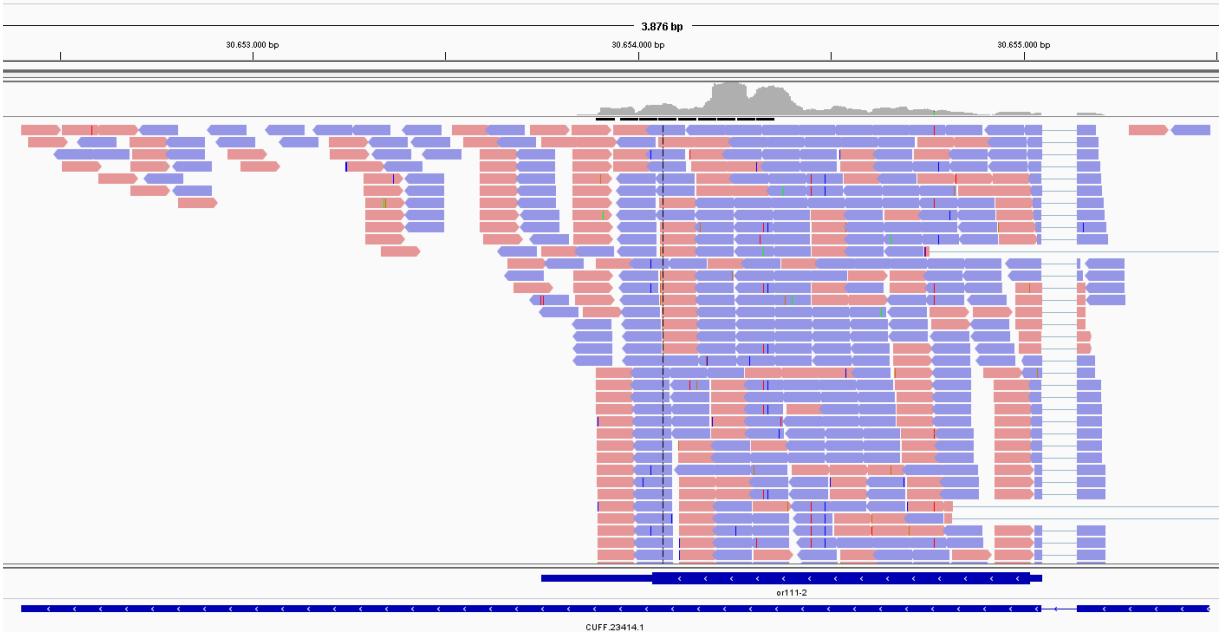

Figure 4.6. Visualization of the mapping of RNA-Seq reads for *or111-2* gene. Sequencing reads and their distribution along genome can be observed. Reads are color coded according their direction. Grey region indicates the coverage distribution. Transcript structure previously annotated in Ensembl is depicted in blue (middle). Cufflinks transcript prediction is also depicted (bottom).

4.1.3. Repertoire-wide Identification of Transcript Structure and Transcription Start Sites of OR genes

OR genes are expressed in a monogenic and monoallelic fashion (Malnic *et al.*, 1999), thus, any given OR gene would only be expressed in a small number of OSNs unlike the other olfactory-related genes mentioned above, which are expressed by the entire OSN population or by large OSN subsets. Thus, individual OR transcripts are less abundantly represented in the transcriptome data, mandating increased sequencing depth to obtain meaningful structural data and to obtain reliable TSSs of OR genes.

As a first analysis, the genomic locations of all known zebrafish OR genes were analyzed. In two previous bioinformatic data mining studies, 131 and 137 zebrafish OR genes were identified from the Zv5 reference genome by iterative searches for sequence homology among OR genes (Niimura and Nei, 2005; Alioto and Ngai, 2005). A in-depth comparison of the results from both studies by BLAST analysis resulted in a list of 166 unique zebrafish OR genes. Using a combination of one-by-one inspection of transcripts within OR gene clusters and subsequent BLAST and / or sequence alignment resulted in the identification of additional 13 OR gene loci, which were not previously reported. Thus, in total 179 candidate OR genes and their respective genomic positions in Zv9 based on their homology to OR genes could be identified (Table 4.1.).

The initial RNA-seq dataset contained 4.68 GB of clean data (52 million sequence reads), however this information was only sufficient to resolve transcript structure and TSS of 129 of these 179 OR genes. The accuracy of TSSs and transcript structures varied with regard to the expression level of individual genes. Thus, increasing the depth of sequencing would increase the quality of TSS mapping and transcript prediction for OR genes, especially for those ORs, which were expressed at low level. Therefore, a second run of RNA-seq was performed on the OE sample to increase the depth of coverage to a combined total of 12 GB of clean data.

The increased combined data set was remapped to the Zv9 zebrafish genome using Tophat (Trapnell *et al*., 2012) and Cufflinks (Trapnell *et al*., 2012) algorithms on a local server. Then, the previously identified 179 zebrafish OR gene loci were re-investigated individually by inspection of RNA sequence reads using IGV software.

As expected, the extended data set allowed for the unambiguous identification of transcript structures and TSS positions of 161 of the 179 (90%) identified OR genes with sufficient resolution (Table 4.1). Of the remaining 18 OR genes, 10 loci were covered by RNA-seq reads, suggesting that these genes are expressed in the OE, albeit at very low levels, which

did not sufficiently cover the 5'-UTRs and therefore did not provide accurate information on TSS position. The remaining 8 ORs genes probably are not expressed as these genes were not supported by RNA-seq and may be transcriptional pseudogenes. When the coding sequences of these 8 genes were analyzed, only one gene (or102-6p) contained a stop codon within the open reading frame whereas remaining genes had intact reading frames, suggesting that expression of these genes has been lost due to mutations within their respective promoter regions.

When the 161 OR genes, which could be mapped with high confidence, were investigated, 68 OR genes (39.7%) contained 5'-intronic sequences, whereas no indication of 3'-intronic sequences could observed for any of the OR genes. Similar observations were made for 198 mouse OR genes for which the transcript structure has been mapped (Michaloski, 2006). In one case, a transcript containing two 5'-introns, or115-10, was observed. The 5'-intron length ranged from 78 bp to 2356 bp. Transcription of the remaining 93 OR genes initiated from the first coding exon, which also includes 5'-untranslated sequences.

Thus, although transcription of the majority of OR genes is initiated in close proximity to the OR coding sequence, transcription can initiate as far as 5000 bp upstream of the coding sequence. Without mapping of proper TSSs for these genes, relevant regulatory sequences located around the TSS would have been missed if the beginning of the coding sequence would have been taken as a reference point.

All but one of the transcripts, or139-1, contained an uninterrupted coding sequence located within a single exon. For the or139-1 gene, however, an unusual transcript in which the coding sequence was distributed over two adjacent exons and interrupted by a 2055 bp intronic sequence was observed. Curiously, evidence for alternatively spliced introns and intron retention could be observed for 39 of the 68 OR genes (57.4%), which contained 5'-introns. In such cases, two different kind of sequence read distribution could be observed for a single

transcript: split reads that mapped to two distant genomic locations and uninterrupted reads that filled the intronic region.

Thus, increasing the depth of RNA sequencing significantly improved the quality of TSS mapping and transcript structure prediction by 50%. An accurate and precise map of the TSSs and transcript structure for 94% of the expressed zebrafish OR gene repertoire could be generated. The TSSs were located upstream of the OR coding sequences and in a significant proportion of genes the two sites were separated by intronic sequence. The reason for this disparity in transcript structure are not understood, but it has been shown generally that splicing can increase transcription and / or translation of genes (Le Hir *et al*., 2003).

Table 4.1. Mapping of OR gene TSSs and intron exon structure for 161 OR genes. TSS denotes the genomic position of TSSs, OR genes that Show alternative splicing by intron retention are indicated in the last column. Strand shows the orientation of transcription.

#	OR name	Ensembl ID	TSS	Exon 1	Intron 1	Exon 2	CDS	Strand	Intron Retention
					chromosome 6				
1	or138-1	no Ensembl ID	6:52806039	6:52806039-52807546	none	none	6:52806070-52807080	1	
2	or134-1	ENSDARG00000070341	6:52809741	6:52809741-52809796	6:52809797-52812152	6:52812153-52814828	6:52812177-52813124	1	+
3	or139-1	ENSDARG00000087254	6:52816777	6:52816777-52817390	6:52817391-52819445	6:52819446-52820369	6:52817133-52817390 6:52819446-52820165	1	
4	or140-1	ENSDARG00000094030	6:52822550	6:52822550-52824174	none	none	6:52822577-52823509	1	
5	or137-3	ENSDARG00000070344	6:52826537	6:52826537-52826550	6:52826551-52826667	6:52826668-52828661	6:52826825-52827757	1	+
6	or137-8	ENSDARG00000055495	6:52829826	6:52829826-52831451	none	none	6:52830072-52831004	1	
7	or137-4	ENSDARG00000070345	6:52833012	6:52833012-52834274	none	none	6:52833236-52834168	1	
8	or137-9	ENSDARG00000070347	6:52841755	6:52841755-52841935	6:52841936-52842021	6:52842022-52843762	6:52842149-52843081	1	
9	or137-5	ENSDARG00000055485	6:52844352	6:52844352-52846243	none	none	6:52844807-52845739	1	
10	or137-7	ENSDARG00000091877	6:52847613	6:52847613-52847657	6:52847658-52847749	6:52847750-52850194	6:52847916-52848845	1	+
11	or137-6	ENSDARG00000093273	6:52855841	6:52855841-52855862	6:52855863-52855991	6:52855992-52859577	6:52856167-52857090	1	
12	**or137-10**	ENSDARG00000055467	6:52863868	6:52863868-52863878	6:52863879-52864009	6:52864010-52866825	6:52864185-52865111	1	
13	or137-1	ENSDARG00000055465	6:52871196	6:52871196-52872341	none	none	6:52871413-52872345	1	
14	or137-2	ENSDARG00000044704	6:52876687	6:52876687-52878383	none	none	6:52876934-52877878	1	
15	or136-2	ENSDARG00000092236	6:52880387	6:52880387-52880422	6:52880423-52880500	6:52880501-52882221	6:52880526-52881449	1	+
16	**or136-3**	ENSDARG00000095628	6:52884490	6:52884490-52888318	none	none	6:52884642-52885571	1	
17	**or136-4**	ENSDARG00000093400	6:52889185	6:52889185-52889223	6:52889224-52891204	6:52891205-52893104	6:52891207-52892130	1	
18	or136-1	ENSDARG00000055458	6:52895533	6:52895533-52895568	6:52895569-52895686	6:52895687-52896627	6:52895689-52896606	1	+
					chromosome 7				
19	**or114-1**	ENSDARG00000077414	7:21496519	7:21493398-21496519	none	none	7:21494699-21495613	-1	
					chromosome 8				
20	or135-1	ENSDARG00000053648	8:38639746	8:38639746-38641298	none	none	8:38639923-38640843	1	
21	or130-1	ENSDARG00000057354	8:5869916	8:5869889-5869916	8:5869634-5869888	8:5866936-58696221	8:5868675-5869622	-1	
22	**or130-2**	no Ensembl ID	8:5893739	8:5892064-5893739	none	none	8:5892616-5893608	-1	
					chromosome 10				
23	or108-3	ENSDARG00000043142	10:37358732	10:37358732-37358871	10:37358872-37359026	10:37359027-37363192	10: 37359097-37360074	1	+
24	or108-2	no Ensembl ID	10:37364917	10:37364917-37366319	none	none	10:37365104-37366084	1	
25	or108-1	ENSDARG00000068704	10:37374626	10:37374626-37374701	10:37374702-37374825	10:37374826-37376152	10:37374870-37375835	1	+
26	or109-13	ENSDARG00000091064	10:37387834	10:37387642-37387834	10:37386482-37387641	10:37385383-37386482	10: 37385487-37386326	-1	+
27	or109-11	ENSDARG00000087002	10:37394852	10:37392397-37394852	none	none	10:37393624-37394622	-1	
28	or109-7	no Ensembl ID	10:37401345	10:37397870-37401345	none	none	10:37398904-37399881	-1	
29	or109-6	ENSDARG00000090279	10:37403116	10:37401887-37403116	none	none	10:37401933-37402910	-1	
30	or109-5	ENSDARG00000090513	10:37405925	10:37404836-37405925	none	none	10:37404799-37405797	-1	
31	or109-4	ENSDARG00000089560	10:37411973	10:37410819-37411956	none	none	10:37410796-37411812	-1	
32	or109-3	no Ensembl ID	10:37416504	10:37415185-37416504	none	none	10:37415302-37416306	-1	
33	or109-2	ENSDARG00000087707	10:37421162	10:37418018-37421162	none	none	10:37419845-37420855	-1	
34	or109-1	ENSDARG00000091854	10:37424959	10:37422600-37424959	none	none	10:37423839-37424828	-1	
35	or110-2	ENSDARG00000091172	10:37434527	10:37434447-37434527	10:37434086-37434446	10:37432926-37434085	10:37433099-37434055	-1	
36	or110-1	ENSDARG00000091203	10:37440730	10:37440450-37440730	10:37440140-37440449	10:37438674-37440139	10:37439208-37440104	-1	
37	**or106-11**	no Ensembl ID	10:37445256	10:37445256-37447535	none	none	10:37445380-37446591	1	
38	**or106-12**	no Ensembl ID	10:37452281	10:37452281-37455087	none	none	10:37452601-37453548	1	
39	or106-10	ENSDARG00000091143	10:37457613	10:37457613-37459707	none	none	10:37457939-37458880	-1	
40	or106-9	no Ensembl ID	10:37465659	10:37465659-37467491	none	none	10:37465990-37466931	1	
41	or106-8	no Ensembl ID	10:37474678	10:37474678-37476742	none	none	10:37475074-37476090	1	
42	or106-7	no Ensembl ID	10:37479779	10:37479779-37481061	none	none	10:37479974-37480951	1	
43	or106-6	no Ensembl ID	10:37483039	10:37483039-37484808	none	none	10:37483376-37484380	1	
44	or106-5	no Ensembl ID	10:37489175	10:37489175-37491099	none	none	10:37489276-37490412	1	
45	or106-4	no Ensembl ID	10:37494127	10:37494127-37496836	none	none	10:37494726-37495676	1	
46	or106-3	no Ensembl ID	10:37503036	10:37503036-37505244	none	none	10:37503365-37504327	1	
47	or106-2	no Ensembl ID	10:37508794	10:37508794-37509102	10:37509103-37509213	10:37509214-37511038	10:37509145-37510107	1	+
48	or106-1	ENSDARG00000068661	10:37514713	10:37514713-37516202	none	none	10:37514958-37515917	1	
49	or105-1	ENSDARG00000068660	10:37523673	10:37523673-37523694	10:37523695-37525298	10:37525299-37526434	10:37525308-37526285	1	+
50	or104-2	ENSDARG00000068659	10:37529323	10:37529323-37529391	10:37529392-37529546	10:37529547-37530836	10:37529548-37530514	1	
51	or104-1	no Ensembl ID	10:37534953	10:37534953-37535033	10:37535034-37535280	10:37535281-37538818	10:37535282-37536265	1	

Table 4.1. Mapping of OR gene TSSs and intron exon structure for 161 OR genes (cont.).

chromosome 15									
52	or131-1	ENSDARG00000005865	15:17167394	15:17165454-17167394	none	none	15:17166355-17167320	-1	
53	**or131-3**	ENSDARG00000077898	15:17176340	15:17175185-17176340	none	none	15:17175326-17176306	-1	
54	or131-2	ENSDARG00000073812	15:17183778	15:17182130-17183778	none	none	15:17182749-17183726	-1	
55	or116-1	ENSDARG00000054734	15:30594072	15:30594072-30594093	15:30594094-30594169	15:30594170-30595766	15:30594176-30595111	1	
56	or116-2	ENSDARG00000089994	15:30600911	15:30600911-30600954	15:30600955-30601028	15:30601029-30602942	15:30601035-30601973	1	+
57	or117-1	ENSDARG00000090892	15:30605376	15:30605376-30605428	15:30605429-30605539	15:30605540-30606709	15:30605547-30606527	1	+
58	or118-1	ENSDARG00000054720	15:30611484	15:30609151-30611484	none	none	15:30610064-30611002	-1	
59	or118-3	ENSDARG00000054719	15:30616844	15:30616801-30616844	15:30616608-30616800	15:30614344-30616607	15:30615532-30616473	-1	+
60	or118-2	ENSDARG00000054718	15:30621182	15:30620156-30621182	none	none	15:30620222-30621160	-1	
61	**or119-1**	ENSDARG00000041034	15:30625018	15:30625018-30625133	15:30625134-30625222	15:30625223-30626476	15:30625232-30626176	1	
62	or119-2	ENSDARG00000041033	15:30629890	15:30629890-30630226	15:30630227-30630315	15:30630316-30632153	15:30630325-30631269	1	+
63	or107-1	ENSDARG00000041032	15:30642037	15:30642037-30642060	15:30642061-30642308	15:30642309-30644080	15:30642326-30643321	1	
64	or111-1	ENSDARG00000041030	15:30649682	15:30649544-30649682	15:30649461-30649543	15:30647715-30649460	15:30648450-30649431	-1	+
65	or111-2	ENSDARG00000025658	15:30655483	15:30655140-30655483	15:30655048-30655139	15:30652402-30655047	15:30654037-30655017	-1	
66	or111-3	ENSDARG00000086633	15:30664297	15:30664123-30664297	15:30664026-30664122	15:30662431-30664025	15:30663015-30663995	-1	+
67	or111-4	ENSDARG00000088846	15:30670377	15:30669923-30670377	15:30669796-30669922	15:30665938-30669795	15:30668786-30669766	-1	+
68	or111-5	ENSDARG00000091791	15:30681040	15:30679147-30681040	none	none	15:30679948-30680928	-1	
69	or111-6	ENSDARG00000041024	15:30688623	15:30688421-30688623	15:30688315-30688420	15:30685809-30688314	15:30687306-30688283	-1	
70	or111-7	ENSDARG00000088453	15:30699782	15:30694500-30699782	none	none	15:30697711-30698688	-1	
71	or111-8	ENSDARG00000086383	15:30705108	15:30703714-30705108	15:30703607-30703713	15:30702485-30703606	15:30702611-30703585	-1	+
72	or111-9	ENSDARG00000092238	15:30706852	15:30705144-30706582	none	none	15:30705392-30706336	-1	
73	or111-10	ENSDARG00000091216	15:30711671	15:30711475-30711671	15:30711353-30711474	15:30710108-30711352	15:30710342-30711322	-1	+
74	or111-11	ENSDARG00000041019	15:30717347	15:30717211-30717347	15:30717099-30717210	15:30714374-30717098	15:30716091-30717068	-1	+
75	or103-1	ENSDARG00000094080	15:30724068	15:30724068-30724091	15:30724092-30724319	15:30724320-30725654	15:30724333-30725319	1	
76	or103-5	ENSDARG00000075128	15:30725695	15:30725695-30726591	15:30726592-30726709	15:30726710-30729904	15:30726748-30727691	1	+
77	or103-2	ENSDARG00000094125	15:30733216	15:30733216-30733393	15:30733394-30733511	15:30733512-30735994	15:30733549-30734493	1	+
78	or103-4	ENSDARG00000003090	15:30742179	15:30742078-30744072	none	none	15:30742206-30743162	1	
79	or102-1	ENSDARG00000041005	15:30760223	15:30760223-30760332	15:30760333-30760488	15:30760489-30763758	15:30760404-30761450	1	+
80	or102-2	ENSDARG00000004598	15:30764422	15:30764422-30766836	none	none	15:30764546-30765529	1	
81	or102-3	ENSDARG00000074966	15:30769576	15:30769576-30772022	none	none	15:30769824-30770810	1	
82	or102-4	ENSDARG00000041003	15:30774345	15:30774345-30778783	none	none	15:30774365-30775345	1	
83	or128-10	ENSDARG00000042810	15:5016381	15:5016313-5016381	15:5016202-5016312	15:5014258-5016203	15:5015253-5016170	-1	
84	or126-7	ENSDARG00000058724	15:5023933	15:5021551-5023933	none	none	15:5022856-5023731	-1	
85	or126-5	ENSDARG00000042809	15:5031006	15:5030984-5031006	15:5030911-5030983	15:5028951-5030910	15:5029963-5030910	-1	+
86	or126-4	ENSDARG00000093536	15:5037740	15:5037716-5037740	15:5037633-5037715	15:5034338-5037634	15:5036586-5037530	-1	
87	or126-3	ENSDARG00000092581	15:5042923	15:5041735-5042923	none	none	15:5041868-5042815	-1	
88	or126-2	ENSDARG00000092415	15:5050499	15:5050473-5050499	15:5050327-5050472	15:5048345-5050326	15:5049372-5050307	-1	+
89	or128-9	ENSDARG00000093930	15:5056876	15:5055692-5056876	none	none	15:5055874-5056791	-1	
90	or128-7	ENSDARG00000095222	15:5066464	15:5064611-5066464	none	none	15:5065471-5066388	-1	
91	or128-6	ENSDARG00000094092	15:5071590	15:5070285-5071590	none	none	15:5070585-5071502	-1	
92	or128-5	ENSDARG00000087635	15:5076473	15:5076441-5076473	15:5076161-5076440	15:5073316-5076160	15:5075207-5076127	-1	
93	or128-4	ENSDARG00000090931	15:5081733	15:5081910-5081733	15:5081432-5081682	15:5079899-5081431	15:5080480-5081397	-1	
94	or128-3	ENSDARG00000058762	15:5089222	15:5089177-5089222	15:5088934-5089176	15:5086744-5088933	15:5087988-5088899	-1	
95	or128-2	ENSDARG00000058765	15:5098323	15:5098278-5098323	15:5098019-5098277	15:5097018-5098018	15:5097067-5097984	-1	
96	or128-1	ENSDARG00000058767	15:5104561	15:5104481-5104561	15:5104260-5104480	15:5103124-5104259	15:5103309-5104226	-1	
97	or126-1	ENSDARG00000091929	15:5111475	15:5111441-5111475	15:5110883-5111440	15:5109909-5110882	15:5109953-5110882	-1	+
98	or127-1	ENSDARG00000095060	15:5121733	15:5120597-5121733	none	none	15:5120644-5121588	-1	
99	**or127-2**	no Ensembl ID	15:5136709	15:5136709-5138390	none	none	15:5136788-5137657	1	
100	or121-1	no Ensembl ID	15:5139738	15:5139732-5140752	none	none	15:5139824-5140762	1	
101	or122-1	ENSDARG00000087761	15:5146628	15:5146628-5147986	none	none	15:5146846-5147769	1	
102	or122-2	no Ensembl ID	15:5158424	15:5158424-5159449	none	none	15:5158481-5159404	1	
103	or120-1	ENSDARG00000091358	15:5168352	15:5168332-5168352	15:5168250-5168331	15:5166719-5168249	15:5167274-5168245	-1	+
104	or113-1	no Ensembl ID	15:5180753	15:5180753-5182825	none	none	15:5181033-5181986	1	
105	or113-2	no Ensembl ID	15:5192055	15:5192055-5193320	none	none	15:5192391-5193344	1	
106	or113-3	ENSDARG00000077146	15:5202415	15:5202415-5202525	15:5202526-5202700	15:5202701-5203631	15:5202731-5203651	1	
107	or114-1	no Ensembl ID	15:5210786	15:5210786-5210944	15:5210945-5211038	15:5211039-5213509	15:5211057-5212001	1	+
108	or112-1	ENSDARG00000077211	15:5221511	15:5221511-5221745	15:5221746-5222317	15:5222318-5223452	15:5222344-5223327	1	+

Table 4.1. Mapping of OR gene TSSs and intron exon structure for 161 OR genes (cont.).

			chromosome 21						
109	or113-4	ENSDARG00000094919	21:20418088	21:20416246-20418068	none	none	21:20416973-20417929	-1	
110	or133-8	ENSDARG00000092756	21:20424658	21:20420896-20424658	none	none	21:20423567-20424517	-1	
111	or133-7	ENSDARG00000093379	21:20428237	21:20426237-20428323	none	none	21:20427283-20428233	-1	
112	or133-6	ENSDARG00000092852	21:20436978	21:20435078-20436978	none	none	21:20436118-20437071	-1	
113	or133-5	ENSDARG00000092193	21:20442246	21:20440451-20442246	none	none	21:20441184-20442149	-1	
114	or133-10	ENSDARG00000094819	21:20451202	21:20449599-20451202	none	none	21:20450393-20451178	-1	
115	or133-4	ENSDARG00000093399	21:20457531	21:20455825-20457531	none	none	21:20456257-20457210	-1	
116	or133-3	ENSDARG00000092982	21:20467376	21:20466003-20467376	none	none	21:20466327-20467274	-1	
117	or133-2	ENSDARG00000094913	21:20482026	21:20479483-20482026	none	none	21:20480973-20481902	-1	
118	or133-1	ENSDARG00000056911	21:20487925	21:20486421-20487925	none	none	21:20486796-20487758	-1	
119	or125-8	ENSDARG00000056940	21:20542947	21:20542947-20545622	none	none	21:20543103-20544198	1	
120	or125-7	ENSDARG00000093183	21:20550262	21:20550262-20551500	none	none	21:20550439-20551374	1	
121	or125-6	ENSDARG00000092577	21:20560923	21:20560869-20561909	none	none	21:20560865-20561818	1	
122	or125-5	ENSDARG00000094812	21:20566598	21:20566598-20568005	none	none	21:20566461-20567387	1	
123	or125-4	ENSDARG00000095417	21:20572442	21:20572442-20574335	none	none	21:20572641-20573564	1	
124	or125-3	ENSDARG00000095443	21:20579504	21:20579504-20580682	none	none	21:20579683-20580612	1	
125	or125-2	ENSDARG00000094714	21:20590022	21:20590022-20591231	none	none	21:20590011-20590931	1	
126	or125-1	ENSDARG00000093856	21:20596399	21:20596399-20598214	none	none	21:20596570-20597529	1	
127	or123-1	ENSDARG00000092558	21:20600505	21:20600505-20600607	21:20600608-20600737	21:20600738-20601644	21:20600768-20601703	1	+
128	or124-1	ENSDARG00000092707	21:20612520	21:20612520-20613479	none	none	21:20612569-20613495	1	
129	or124-3	ENSDARG00000093603	21:20614458	21:20614458-20615929	none	none	21:20614668-20615604	1	
130	or124-2	ENSDARG00000094326	21:20616874	21:20616874-20618052	none	none	21:20617028-20617960	1	
131	**or125-9**	no Ensembl ID	21:20635806	21:20635806-20637454	none	none	21:20636000-20636923	1	
132	**or125-10**	ENSDARG00000037147	21:20642271	21:20642271-20643428	None	None	21:20642429-20643358	1	
133	**or125-11**	no Ensembl ID	21:20658833	21:20658833-20659901	None	None	21:20658865-20659788	1	
134	**or124-5**	no Ensembl ID	21:20671267	21:20671267-20672225	none	none	21:20671316-20672242	1	
135	or124-4	ENSDARG00000096264	21:20690029	21:20690029-20691815	none	none	21:20690243-20691184	1	
136	or132-1	ENSDARG00000095656	21:24600041	21:24600041-24603787	none	none	21:24601059-24602030	1	
137	or132-2	ENSDARG00000094515	21:24613296	21:24613296-24613576	21:24613577-24613658	21:24613659-24617213	21:24613662-24614630	1	+
138	**or132-6**	ENSDARG00000092866	21:24622014	21:24622014-24625958	none	none	21:24622219-24623187	1	
139	or132-4	ENSDARG00000094153	21:24631787	21:24631787-24634286	none	none	21:24631955-24632923	1	
140	or132-3	ENSDARG00000095431	21:24642036	21:24642036-24643775	none	none	21:24642187-24643155	1	
141	or132-5	ENSDARG00000056277	21:24652904	21:24652904-24652971	21:24652972-24653097	21:24653098-24655974	21:24653101-24654069	1	+
142	or115-15	ENSDARG00000079294	21:39031841	21:39031841-39033388	none	none	21:39032140-39033090	1	
143	or115-1	ENSDARG00000018521	21:39038987	21:39038987-39039069	21:39039070-39039302	21:39039303-39040758	21:39039310-39040248	1	+
144	or115-14	ENSDARG00000094235	21:39051751	21:39051751-39051861	21:39051862-39052078	21:39052079-39054075	21:39052086-39053024	1	
145	or115-13	ENSDARG00000044343	21:39056150	21:39056150-39056249	21:39056250-39056465	21:39056466-39057968	21:39056473-39057411	1	
146	or115-12	ENSDARG00000044748	21:39062228	21:39062228-39062428	21:39062429-39062495	21:39062496-39064319	21:39062502-39063440	1	+
147	or115-11	ENSDARG00000053813	21:39068928	21:39068928	two introns	21:39070907	21:39069372-39070289	1	+
148	or115-10	ENSDARG00000035048	21:39072295	21:39072295-39072421	21:39072422-39073024	21:39073025-39074561	21:39073032-39073973	1	+
149	or115-9	ENSDARG00000069040	21:39075367	21:39075367-39075832	21:39075833-39075966	21:39075967-39077746	21:39075986-39076930	1	
150	**or115-16**	ENSDARG00000094505	21:39086885	21:39086772-39086900	21:39086411-39086771	21:39084010-39086410	21:39085466-39086407	-1	
151	or115-8	ENSDARG00000092406	21:39090818	21:39090818-39090890	21:39090891-39091085	21:39091086-39092249	21:39091105-39092049	1	+
152	**or115-17**	ENSDARG00000068937.4	21:39097398	21:39097398-39099672	none	none	21:39097599-39098528	1	
153	or115-7	ENSDARG00000044338	21:39104677	21:39104677-39105971	none	none	21:39104949-39105890	1	
154	**or115-18**	ENSDARG00000053779	21:39112953	21:39112953-39115438	none	none	21:39113241-39114182	1	
155	or115-6	ENSDARG00000044337	21:39116417	21:39116417-39116528	21:39116529-39117309	21:39117310-39118462	21:39117330-39118277	1	+
156	or115-5	ENSDARG00000022319	21:39123369	21:39123369-39126090	none	none	21:39123514-39124455	1	
157	or115-4p	ENSDARG00000044336	21:39129181	21:39129181-39129535	21:39129536-39129878	21:39129879-39132369	21:39129897-39130850	1	+
158	or115-2	ENSDARG00000091915	21:39132991	21:39132991-39134034	none	none	21:39133156-39134082	1	
159	**or115-19**	ENSDARG00000095143	21:39139131	21:39139131-39139178	21:39139179-39139517	21:39139518-39140523	21:39139533-39140471	1	
160	or115-3	ENSDARG00000053817	21:39143858	21:39143858-39143929	21:39143930-39144039	21:39144040-39144947	21:39144059-39145003	1	
161	or101-1	ENSDARG00000013014	21:39153416	21:39153227-39153416	21:39152877-39153226	21:39150604-39152465	21:39151877-39152827	-1	

4.1.4. Comparison of Transcript Prediction by RNA-seq to Annotated Transcripts

In order to estimate the quality of our transcript prediction, we compared the results obtained by RNA-Seq data to previously annotated OR gene transcripts. Of the identified 179 OR genes, 131 OR gene transcripts were annotated in the Zv9/DanRer7 genome or listed in the RefSeq database (Pruitt, 2014). However, for most OR genes only the coding sequence was annotated and for the few genes for which transcript structures were reported these sequences were not derived from olfactory tissue in some cases. A total of eight OR genes, for which transcript structures have been verified from olfactory tissue by RACE or direct sequencing of cDNA clones (Taştekin, 2012) were chosen to compare their reported transcript structure to the prediction by RNA-seq.

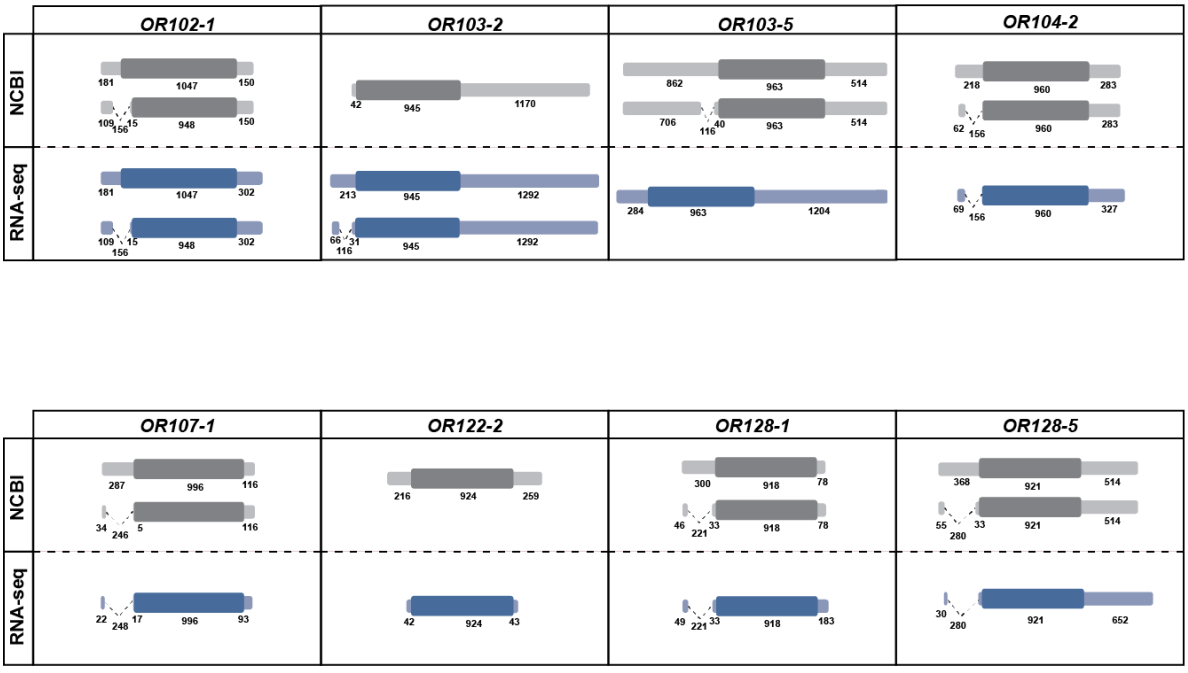

Figure 4. 7. Comparison of RNAseq transcript prediction to known OR transcripts. Comparison of eight previously reported OR gene transcript structures that were derived from zebrafish OSNs with transcript prediction. Top row shows the annotated transcripts structures as reported in NCBI, the bottom row shows our resolution of the corresponding transcript, including alternative splicing by intron retention

When the or102-1, or103-2, or103-5, or104-2, or107-1, or122-2, or128-1, or128-5 and or122-2 genes were analyzed, the structure of six genes (or102-1, or 104-2, or107-1, or128-1, or128-5 and OR122-2) determined by RNA-seq closely resembled previously reported structures (Figure 4.7). In one case (OR128-1), the reported intron retention (Taştekin, 2012) could not be observed and only split sequence reads distributed around the 5'-inton, but not within the intron, could be observed from the RNA-seq data. For the or103-2, transcript prediction by RNA-seq revealed a longer 5'-UTR and a novel intron for some transcripts. However, for the or103-5 gene the reported intron structure (Taştekin, 2012) could not be detected from the RNA-seq data.

For the further assesment of the quality of structure prediction by RNA-seq, transcript prediction was compared to 5'-RACE data previously generated by our group for four genes (Taştekin, 2012), including the members of the or103 gene family (Figure 4.8). Comparison of or101-1 gene 5'-RACE data and our transcript prediction showed in very high similarity between the two methods, whereas inconsistencies were observed between RNA-seq and 5'-RACE data for or103 family members. It can be speculated that the high sequence homology of 65% between or103-1 and or103-5 and mismapping of RNA-seq reads to or103 loci have caused these discrepancies. On the other hand, for the or103-2 gene, it is likely that transcript prediction by RNA-seq is more accurate than the 5'-RACE data considering the fact that the predicted transcripts have a longer 5'-UTR and that an alternatively spliced variant was found from RNA-seq, which might have been missed by 5'-RACE.

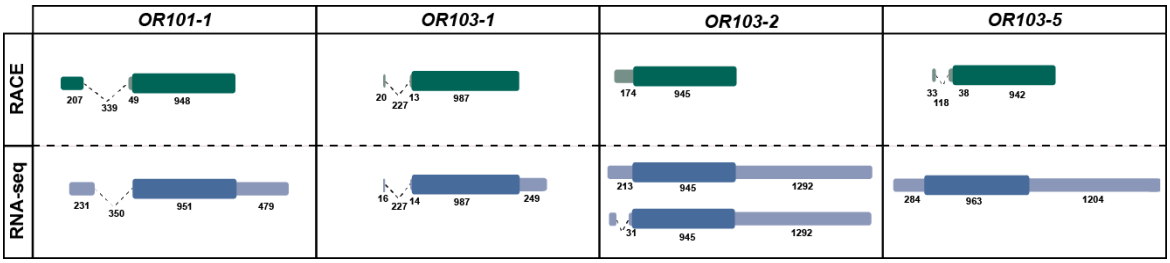

Figure 4. 8. Comparison of RNAseq transcript prediction to 5'-RACE data.
Comparison of four previously identified 5'-transcript structures (top row, green transcripts) with our trancript prediction (bottom row, blue transcripts).

Thus, transcript prediction by RNA-seq appears to be accurate in many cases when compared to transcript structures established by other experimental means. Yet, in some cases additional relevant information on the extend of 5'-sequences and additional alternative transcripts can be detected by RNA-seq. Especially the tight correlation between RNA-seq and 5'-RACE results strongly supports the validity of the RNA-seq approach taken here. Because the majority (75%) of OR genes did not contain any information on transcript structure at all, the RNA-seq-based TSS prediction proved feasible to generate the required data for further analysis of OR promoter structure.

4.2. Regulatory Motif Investigation

In an attempt to identify candidate regulatory sequences in OR gene promoters, sequences upstream of the TSS of the 161 OR genes for which the precise TSS location was identified, were further analyzed. To do so, the first 500 bp of sequence upstream of the TSS was extracted and bioinformatics analysis was performed to detect conserved or enriched DNA sequence motifs. For unguided detection of enriched DNA motifs the MEME Suite (Bayley *et al.*, 2009) and for prediction of known DNA binding motifs the RSAT suite Matrix Scan tool

(van Helden *et al*., 2008) wer used. In order to provide consistency for different analyses a statistical threshold for all motif scan analysis was chosen.

Default threshold in Matrix Scan was set to 1. Because, positional weight matrices (PWMs) typically contain core motifs and less stringent flanking sequences, certain DNA sequences will match the search motif even though only flanking sequences are matching and core similarity is weak. These kind of false positive hits can be eliminated by choosing proper statistical cutoff values. In order to empirically establish proper statistical thresholds, motif scans were performed at different cutoffs. Increasing the threshold causes a loss of true positives, while too low thresholds generate too many false positives. For of the PWMs used here, it was empirically determined that a threshold value of 4 prevents false positives in most cases. However, promoter regions contained sequences with true positive matches (similar to the core sequence determined by observation) under the threshold 4, false positive hits could still be observed.

4.2.1. Enrichment of TATA Box Binding Protein in OR repertoire

As an initial test for the validity of the general methodology and extracted promoter sequences it was tested whether the approach is capable of identifying previously established sequence motifs, which are typical for eukaryotic gene promoters in general or have been reported for OR gene promoters (Bulger *et al*., 2000). The canonical TATA box motif is present in approximately 27% of vertebrate / mammalian promoters and in 21% of mouse OR gene promoters (Plessy *et al*., 2012). Thus, the occurrence of the TATA box binding protein (TBP) motif was investigated in zebrafish OR upstream sequences by performing a bioinformatics analysis using a PWM obtained from the Universal PBM Resource for Oligonucleotide-Binding Evaluation (UniPROBE) database (http://www.uniprobe.org). The matrix-scan tool in RSAT and the P29037 (UniPROT) / MA0108 (JASPAR) matrix (TBP) was used against a fasta-file which contained 500 bp upstream sequences of 161 OR genes.

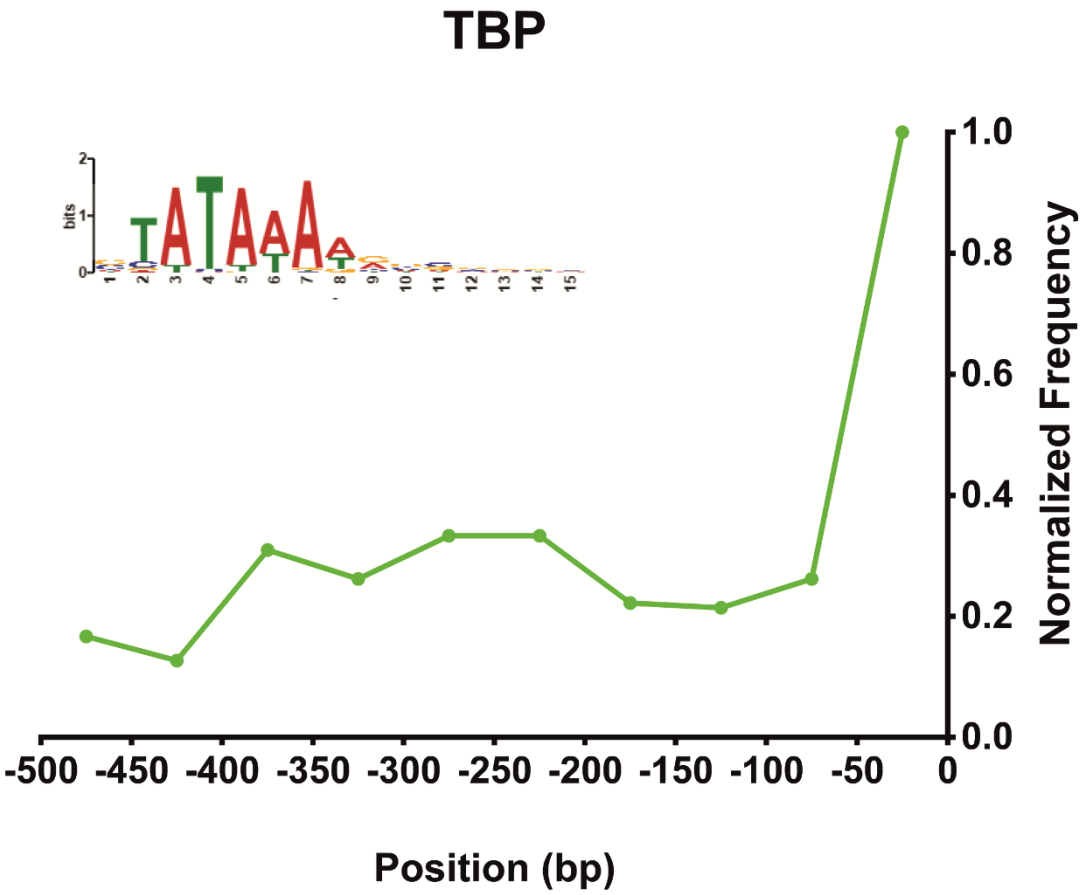

Figure 4. 9. Distribution of the TATA-box binding protein (TBP) motif in OR gene promoters. The first 500 bp upstream of Transcription Start Site (TSS) of OR genes are investigated. Accumulation of hits are observed between -50 bp upstream and the TSS. On the top left, positional weight matrix which used in analysis can be observed.

It has been shown that that 27% of mice promoters and 74.4% of zebrafish promoters contain a clearly identifiable TATA box (Shi and Zhou, 2006). A study, which investigated the presence of the TBP motif within mouse OR promoters showed the presence of TBP motif in 85% of OR promoters and 21% of these OR genes contain the canonical TATA-Box in the expected location (25-35 bp) (Plessy *et al*., 2012). Similar to the mouse, 149 of the161 zebrafish OR promoters contained a TBP motif in the first 500 bp upstream of their respective TSS. The TBP motif was located within the first 50 bp upstream of TSS for the 45.9% (74/161) of the ORs. This position is in good agreement with the location of the canonical TATA box between

25-35 bp upstream of TSS. Thus, it is likely that many of the extracted upstream sequences indeed represent sequences upstream of the TSS of OR genes.

4.2.2. Enrichment of EBF-like (O/E) Transcription Factor in OR repertoire

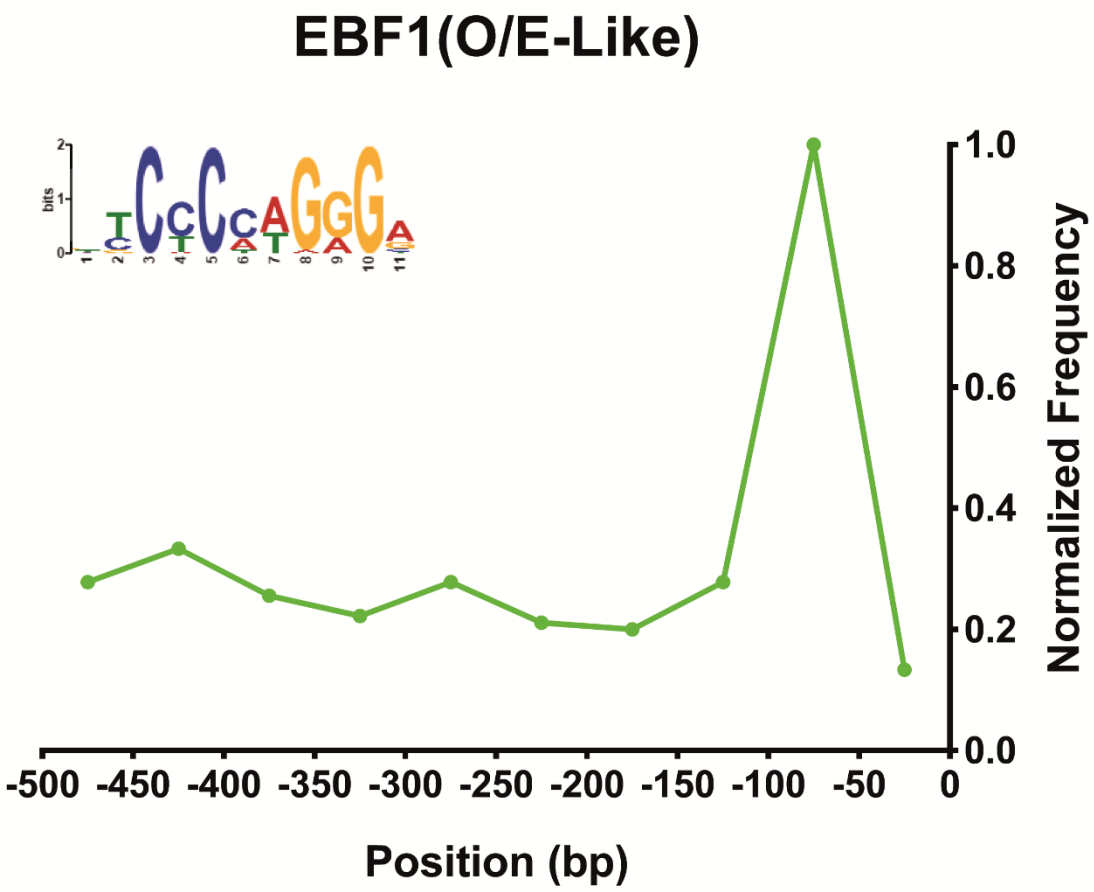

Figure 4. 10. Distribution of the EBF-like (O/E) motif in OR gene promoters. The first 500 bp upstream of Transcription Start Site (TSS) of OR genes are investigated. Accumulation of hits are observed betwen -100 and -50 bp positions. On the top left, postional weight matrix which used in analysis can be observed.

Previous studies indicated that both promoters of olfaction-related genes and OR genes contain Olf1/Ebf1-like (or O/E) motifs (Michalosky *et al*., 2006; Plessy *et al*., 2012), a matrix scan using a PWM for EBF1 (JASPAR database, MA0154.1) was performed against the 161 OR gene upstream regions. Curiously, a distribution pattern of the O/E motif with a tendency to accumulate around the TSS was observed for zebrafish OR genes as well. In 131 of 161 (81.3%) gene promoters the presence of O/E motif could be observed and in 54 of them (33.5%) an O/E-like motif was found between 50-100 bp upstream of TSS.

4.2.3. Enrichment of Homeodomain Transcription Factor in OR repertoire

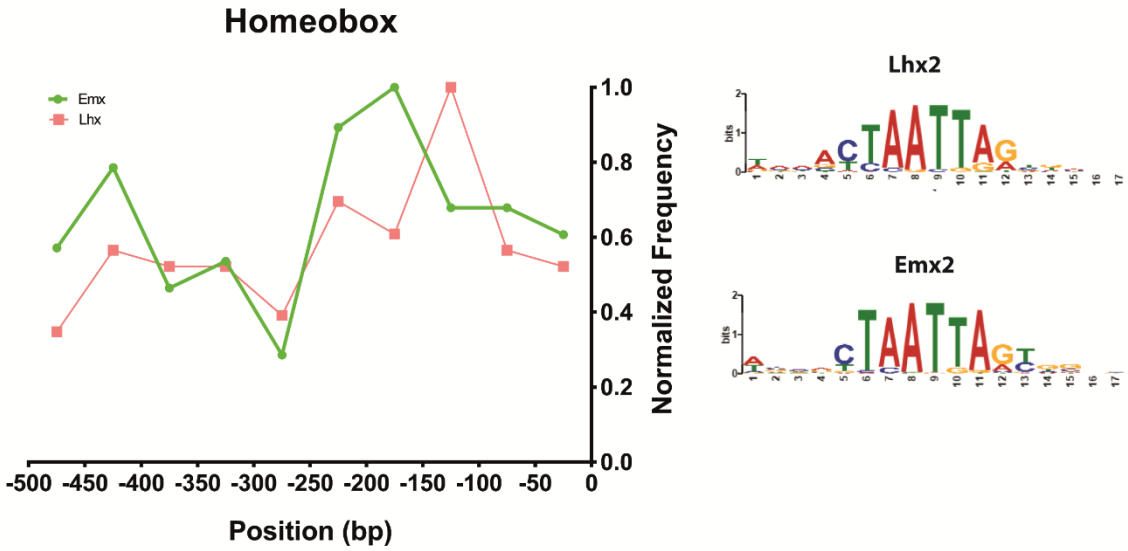

Figure 4. 11. Distribution of the Homeodomain motifs in OR gene promoters. The first 500 bp upstream of Transcription Start Site (TSS) of OR genes are investigated. Accumulation of hits are observed betwen -200 and -100 bp positions. On the right side, Positional Weight Matrices (PWMs) for Homeodomain motifs Lhx2 and Emx2.

Homeodomain motifs were also shown to be present in some murine OR gene promoters (Hirota *et al*., 2004; Rothmann *et al*., 2005; McIntyre *et al*., 2008). When the zebrafish

sequences were scanned using a PWMs for two homeodomain TFs Lhx2 (UP00115) and Emx2 (UP00201) a distribution of these motifs with a peaks within the first 200 bp upstream of the TSS could be observed. In general, the homeodomain-like motif showed a more distal positions than the O/E-like binding motif. A peak distribution of Lhx2 and Emx2 binding motifs was observed between -100 and -200 bp. The Lhx2-like motifs could be detected in 54% of OR gene upstream sequences whereas Emx2-like motifs were present in 60.8% of promoters.

In summary, zebrafish OR upstream sequences contain a strong representation of O/E-like and homeodomain motifs, at least in a significant subset of OR promoters, similar to rodent OR gene promoters. These sequences show spatial preferences with O/E-like sites located closer to the TSS while homeodomain-like sites occupy more distal prositions.

4.2.4. De novo Motif Search in or111 family

In order to perform an unbiased motif search to identify new DNA sequence motifs within OR gene promoters, the oligo-analysis algorithm (van Helden *et al*., 1998), which detects over-represented sequence motifs, was used. As a first test, the OR111 subfamily, which contains 11 family members and which are expressed at intermediate to high levels, was analyzed. Again, the first 500 bp upstream of the TSS was subjected to oligo-analyis and the algorithm resulted in PWMs and distribution pattern of the enriched motif.

Interestingly, a sequence motif with the core sequence CTCTCAAGAGATG could be identified. This motif corresponds to a DNA motif previously described by Dugas and Ngai (2001) and which was implicated in early onset of expression. A comnparison of the motif to known DNA binding motifs using the TOMTOM (Tanaka *et al*., 2011) tool revealed that the PWM obtained is most similar to the Ebf1 motif. The distribution of the motif on or111 family promoters (Figure 4.12) are quite similar to the distribution of Ebf1 binding sites in same region.

7 of the 12 occurences of the novel or111 motif in or111 promoters coincide with the Ebf1 motif in the same sequences.

When the analysis was repeated for the entire set of 161 OR gene upstream sequences, a similar, yet less varied PWM (Figure 11) was obtained. It is likely that the core motif represents the zebrafish-specific variant of the O/E binding site. however, this argument could be far reaching since the functional data for this enriched regions are lacking and some of the regions might be false positive and could be non-functional. The whole repertoire motif has the core region "CTCAAGAGA" whereas the murine Ebf1 motif's core sequence is "TCCCAGGA".

Nevertheless, de novo detection of the O/E-like motif in our data in principle proves the accuracy and feasibility of the methodology for detecting enriched motifs in an unbiased fashion.

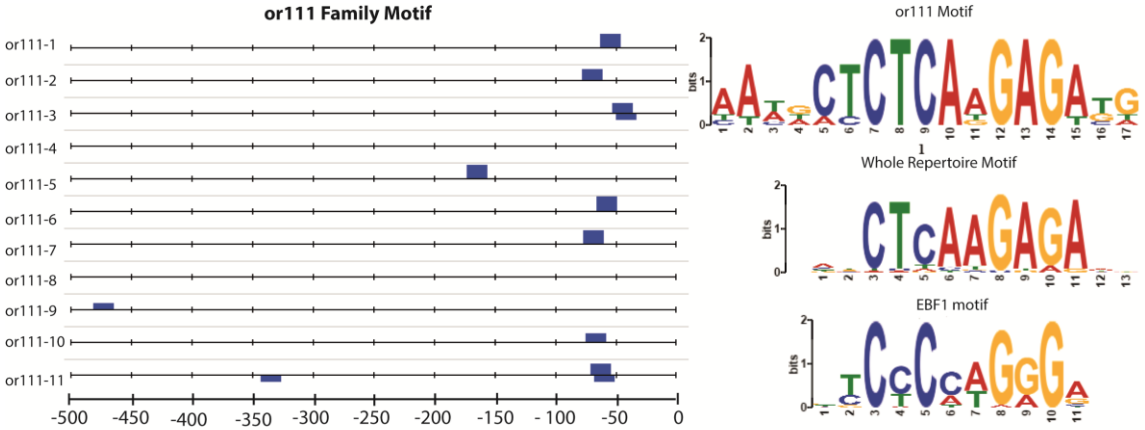

Figure 4. 12. Distribution of the de novo motif obtained from 500 bp upstream regions of or111 family. On the right, positional weight matrices obtained from promoters of or111 family and whole repertoire of OR genes are shown. At the bottom right, EBF1 motif is shown for comparison.

Next, the analysis was extended to identify additional motifs from the entire repertoire of OR genes by using de novo motif finding tools. Some of the oligomers such as CTCAAGAG, AGCAAACT, TCAAGAGA, CCCAAGA, ACCAATTC are enriched in whole repertoire. However, a specific pattern of distribution over the OR promoter repertoire or accumulation on promoters of the genes with highest or lowest expression levels could not be observed for these oligomers. Also, it was not possible to detect a motif or oligomer which is present in whole OR promoters. Therefore, the analysis was extended to the characterization of candidate regulatory factors and sequences that emerged from the transcriptome data and the literature.

4.2.5. Bromodomain PHD finger transcription factor

The comparison of gene expression levels of OE and brain samples revealed genes that are differentially expressed. Genes that were highly expressed in the OE tissue but not in the brain belonged to the group of olfaction-related genes, which were previously shown to have an established role in olfaction. In addition, the analysis also identified transcripts that were not previously implicated in olfactory function. Among those transcripts, a gene which was 1.300-fold enriched in the OE compared to the brain was identified as a gene coding for the Bromodomain PHD finger transcription factor 1 (BPTF). Bptf is capable of binding to the acetylated and methylated histone tails through bromo and PHD-finger domains and this capacity allows it to function as a NURF chromatin remodelling complex (Ruthenberg *et al*., 2011).

Given that epigenetic mechanisms were recently identified to play a significant role in the regulation of OR expression (Clowney *et al*., 2012; Lyons *et al*., 2013), we reasoned that this highly expressed nucleosome-remodeling factor subunit BPTF might have a role in OR gene regulation as well. In fact, it was very recently shown that BPTF has a dual function in OR gene regulation as a facilitator of both enhancer interactions and OR transcription (Markenscoff-Papadimitriou *et al*., 2014). In vivo footprinting experiments suggested that

BPTF binds to enhancer regions and BPTF knock out experiments resulted in significant decrease in OR expression (Markenscoff-Papadimitriou *et al*., 2014).

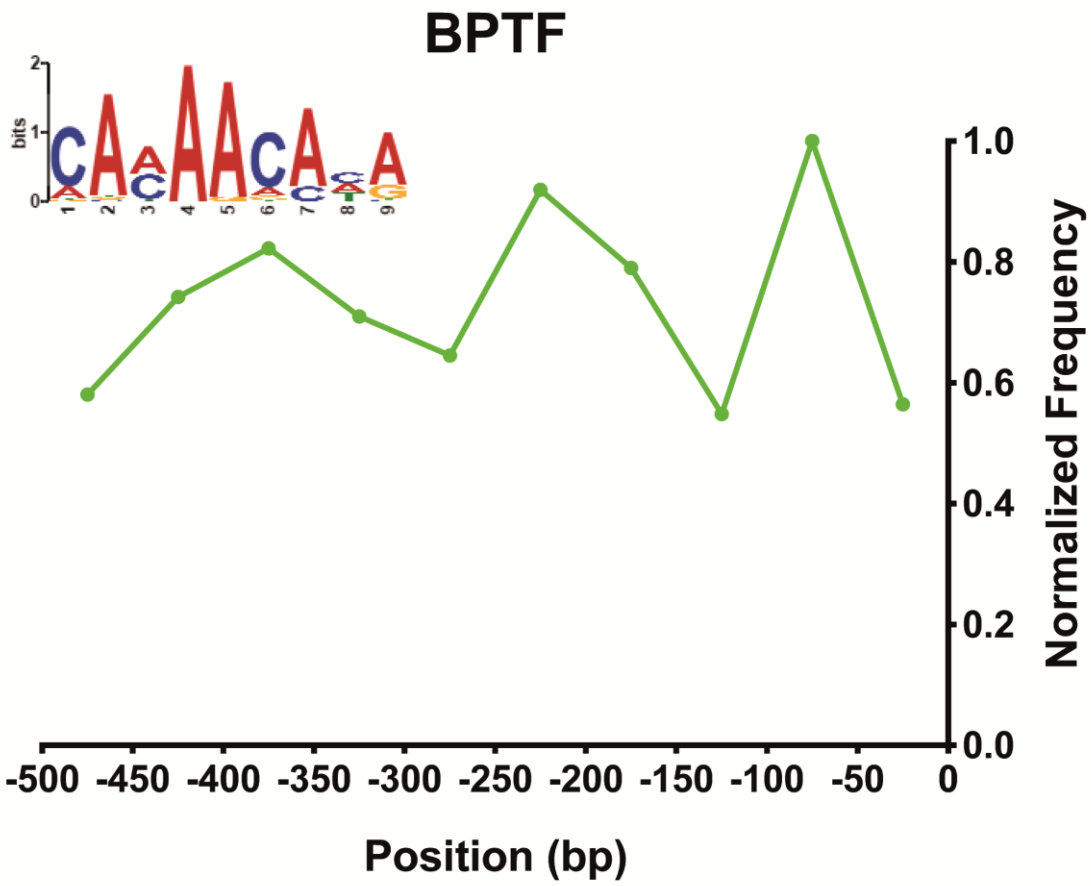

Figure 4. 13. Distribution of the BPTF motifs in OR gene promoters. The first 500 bp upstream of Transcription Start Site (TSS) of OR genes are investigated. Several maxima are observed with 150 bp intervals.

We obtained PWM for the mammalian binding motif from the ISMARA (Integrated System for MOTIF Activity Response Analysis) database and used the PWM in RSAT on the OR promoter data set. In 89.4% of the 161 promoters analyzed, at least one occurrence of BPTF motif was observed. The motif did not show any clear distribution, such as accumulation around the TSSs, but a modulated pattern of distribution with more than one peak. The highest

frequencies of the motif were observed between -50 and -100 bp, -200 and -250 bp and 350-400 bp. These peaks are identified in approximately 150 bp interval. This fixed interval might be related to nucleosome distribution along the OR gene promoters since nucleosome spacing typically varies between 150 and 180 bp (Clark, 2010).

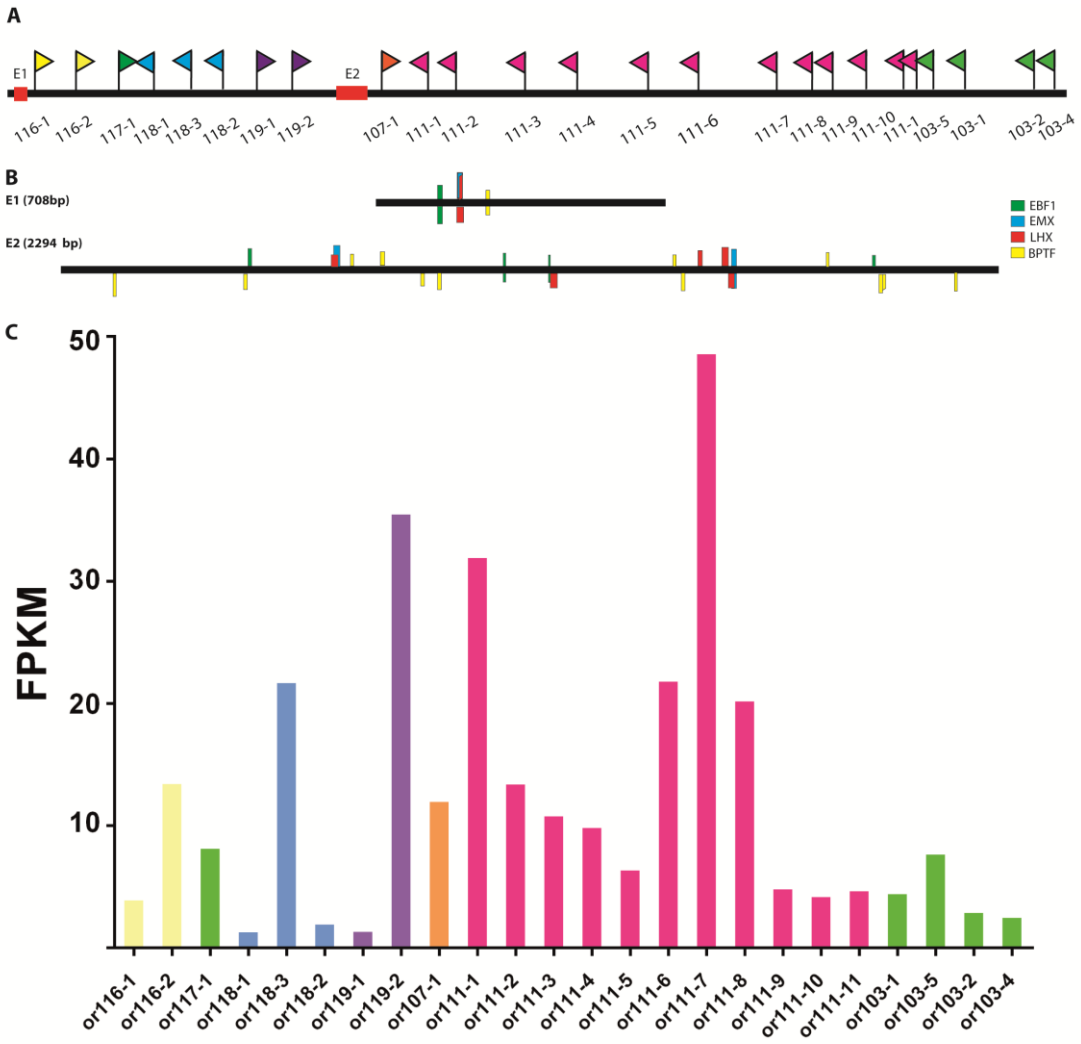

Figure 4. 14. Analysis of E regions A. Chromosomal positions of E1 and E2 regions and the OR gene cluster which surrounds them (Adapted from Nishizumi *et al*., 2007). B. Presence of EBF1, Homeodomain and BPTF motifs in E1 and E2 region. C. Expresssion levels OR gene clusters which located in proximal regions of LCR regions.

4.3. Characterization of LCR regions

As mentioned above, expression levels of OR genes varied along the dimension of chromosomes and within OR gene clusters. Interestingly, a pattern of modulations across OR gene clusters with local maxima and minima can be observed. Figure 9 shows the expression levels on the OR cluster on chromosome 15, with varying fpkm levels. Curiously, the candidate locus control region E2 is located very close to the peak maximum of the cluster or119-2, suggesting that increase in the expression levels within a cluster might be influenced by the presence of regions that have control function. We reasoned that investigating the presence of transcription factor binding motifs and their spatial preferences along these regions might be helpful for understanding the function of these regions.

Investigation of motif distribution within these elements may shed light on what makes them unique compared to other intergenic regiosn between ORs. Previously, the presence of O/E-like and homeodomain motifs was described in the mouse LCR regions H and P element (Nishizumi *et al*., 2007; Bozza *et al*., 2009). In addition, the presence of the BPTF motif might play a role for E1 and E2 function given its role in chromosome rearrangement. As shown in Figure 4.14 B, Ebf1, homeodomain and Bptf motifs are present in the E1 and E2 regions. The height of the colored boxes indicate their statistical significance whereas their upward or downward position indicates genomic orientation. In E1, two Ebf1, one Emx2, two Lhx2 and two Bptf motifs could be detected. Interestingly, all of these motifs are located tightly clustered with a region of only 200 bp. In E2, six Ebf1, three Emx2, five Lhx2 and twelve Bptf motifs could be recognized. In E2 the motifs are more dispersed, yet, two regions of high motif density could be observed. Even though speculative without further experimental support, these motif rich regions could constitute the core functional regions of these candidate long range elements similar to H and P, which also contain conserved motifs of homeodomain-like and O/E binding sites (Nishizumi *et al*., 2007; Vassali *et al.*, 2011)

4.3.1. Enrichment of TF Motifs in Intergenic Regions

If high density of homeodomain- and O/E-like motifs is a signature of cluster regulators of OR expression, density profiles of these motifs may reveal novel long range regulatory sequences. In order to perform the analysis, the 150 kb of OR cluster was divided into 1.000 bp seqments and the presence of Ebf1 or Bptf motifs was scored for each 1.000 bp bin. Then, the average number of motif per bin and the standard deviation for each motif was calculated. The bins, which contained more hits than the sum of average value and standard deviation were considered to be enriched for a motif. The graph in Figure 13 indicates the enriched regions for EBF and BPTF motif. The OR genes which have higher expression than their neighbor genes in a cluster (local maxima) is also indicated. Curiously, Bptf-rich regions of the cluster also contain the genes with the highest levels of expression level. The or111-7, which shows the highest overall level of expression within the cluster, is located in the highest density region of BPTF. Only one of the 1.000 bp bins was enriched with both Ebf1 and BPTF motifs. Interestingly, this region is located adjacent to the or111-7. Also, closely located Ebf1 and Bptf regions are observed around another highly expressed gene or119-2. This Ebf1- and Bptf-rich region is adjacent to E2. (Supplementary figure; Nishizumi *et al*., 2007). However, sequences that only showed enrichment for Ebf1 did not contain local maxima of OR transcription. Even though the data awaits further experimental support, it is an attractive hypothesis that Bptf binding may affect OR gene expression level, probably in combination with Ebf1.

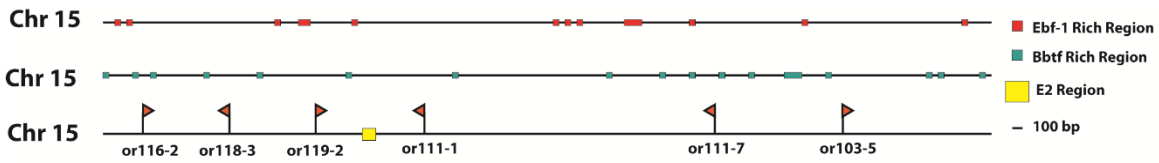

Figure 4. 15 Enriched regions for EBF1 and BPTF motifs. Red regions indicates the EBF1 rich 1000 bp segments whereas blue regions indicate BPTF rich 1000 bp segments. Indicated OR genes are highly expressed genes relative to their respective clusters.

4.4. Investigation of a Highly Expressed OR gene or132-5

Expression of OR genes was not uniform and the cluster comprising the or132 family stood out with the highest levels of expression. The cluster is an isolated cluster of six OR genes located on chromosome 21 containing only members of the or132 family. The gene with the highest level of expression, or132-6, stood out with an fpkm value of 325, which was more than 10-fold higher than the average OR gene. The other members were expressed with fpkm values of 14,6 for or132-1, 57,3 for or132-2, 18,1 for or132-3, 75,8 for or 132-4 and 85,6 for 132-5. Thus, the possibility exists that this gene family contains unique regulatory sites that underlie the observed high level of expression.

4.4.1. *In Situ* Hybridization on or132-5 gene

The transcriptome data suggested that the OR132 family is more abundantly expressed than other OR genes. However, it is not clear whether high levels of expression revealed by FPKM values are a reflection of high levels of expression on an individual cell level or if they represent a high number of OSNs expressing the OR gene. To discriminate between these two possibilities, in situ hybridization for OR132 family members was performed.

Members of or132 gene family are highly similar both in their coding and upstream regions. Therefore, probes for in situ hybridization were designed against the 3'-UTR of the genes to avoid cross-hybridization among family members. Probe sequences were cloned into the pGEM-T Easy vector and RNA probes were obtained by in vitro transcription (Figure 4.17). However, only the or132-5 probe worked successfully *in situ* hybridization experiment.

An in situ hybridization probe against the or132-5 gene, which is only 46.6% similar to the corresponding sequence in other member of or132 family was generated by in vitro transcription. Characteristically, olfactory receptors are expressed in a sparse and ring like pattern (Weth *et al*., 1996). However, or132-5 was expressed in a more dispersed pattern along the OE. As suggested by the RNA-seq data, or132-5 was expressed by an unprecedented high number of OSNs compared to other OR genes. Number of cells or132-5 gene are relatively high (252,5 cell per section) compared to the previous studies (Weth *et al*.,1996; Sato 2007).. Thus, high levels of expression observed by RNA-seq appears not to reflect high level of expression on an individual cell level but rather the number of OSNs expressing a given OR gene.

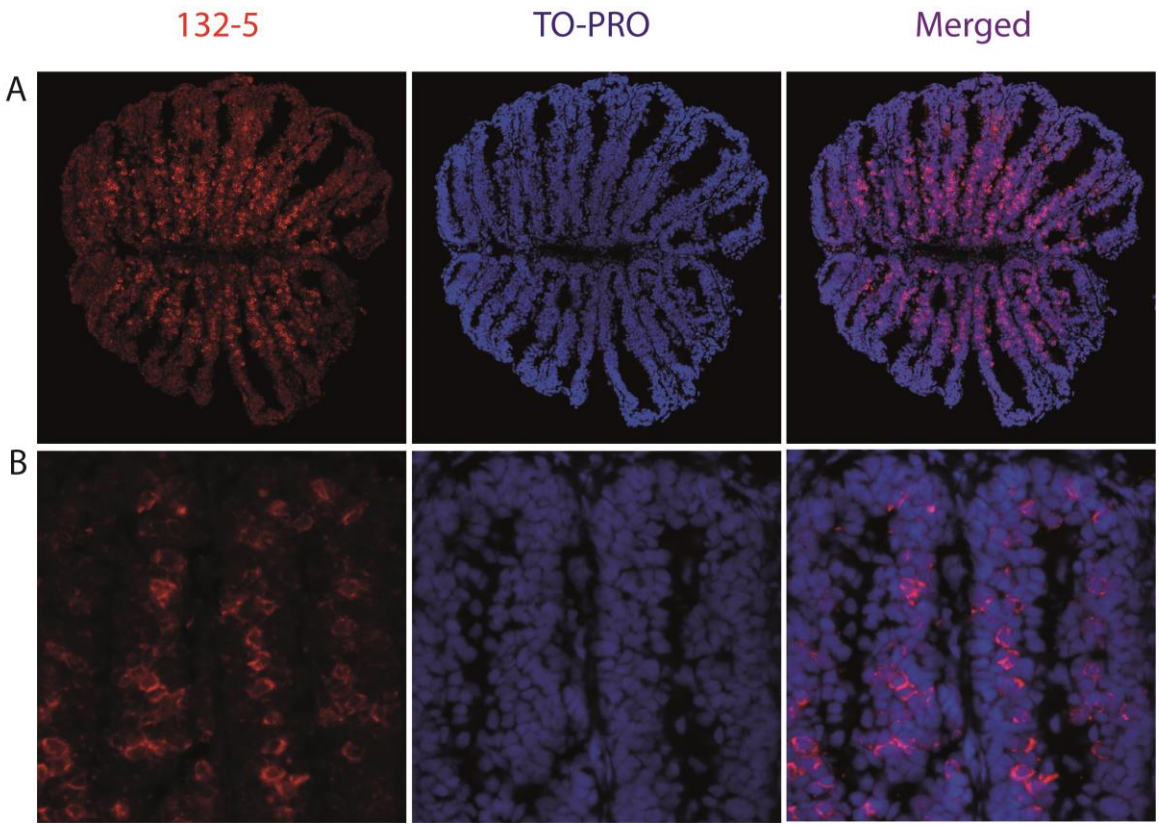

Figure 4. 16. In Situ Hybridization results of or132-5 gene. Red cells indicate OSNs which express or132 gene. Blue cells indicated the whole cells stained by TO-PRO. A.Whole section view B. Close view shows two lamellae.

An in situ hybridization probe against the or132-5 gene, which is only 46.6% similar to the corresponding sequence in other member of or132 family was generated by in vitro transcription. Characteristically, olfactory receptors are expressed in a sparse and ring like pattern (Weth *et al*., 1996). However, or132-5 was expressed in a more dispersed pattern along the OE. As suggested by the RNA-seq data, or132-5 was expressed by an unprecedented high number of OSNs compared to other OR genes. Number of cells or132-5 gene are relatively high (252,5 cell per section) compared to the previous studies (Weth *et al*.,1996; Sato 2007).. Thus, high levels of expression observed by RNA-seq appears not to reflect high level of expression on an individual cell level but rather the number of OSNs expressing a given OR gene.

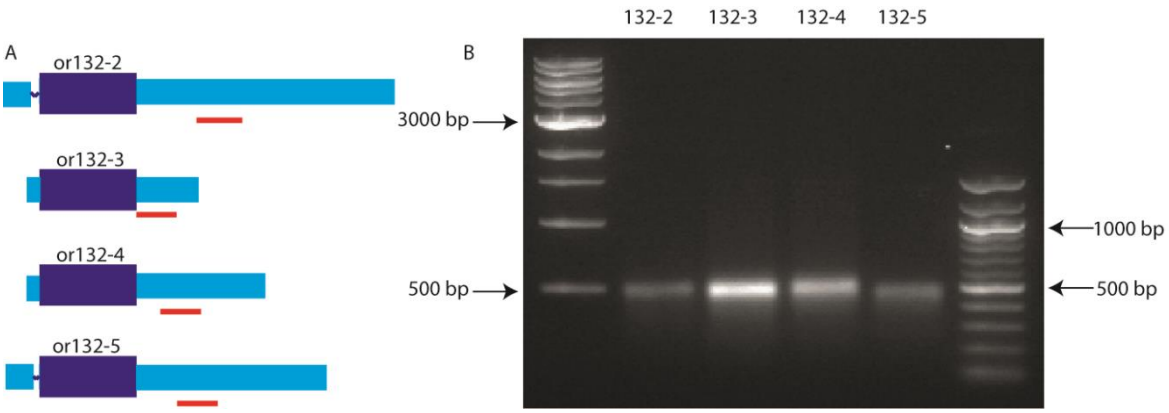

Figure 4. 17. A. Transcript structures of or132-2, or132-3, or132-4 and or132-5 are depicted in blue. B. Agarose gel photo of in situ probes designed for or132-2 (455 bp), or132-3 (400 bp), or132-4 (405 bp) and 132-5 (407 bp).

An in situ hybridization probe against the or132-5 gene, which is only 46.6% similar to the corresponding sequence in other member of or132 family was generated by in vitro transcription. Characteristically, olfactory receptors are expressed in a sparse and ring like pattern (Weth *et al*., 1996). However, or132-5 was expressed in a more dispersed pattern along the OE. As suggested by the RNA-seq data, or132-5 was expressed by an unprecedented high number of OSNs compared to other OR genes. Number of cells or132-5 gene are relatively high

(252,5 cell per section) compared to the previous studies (Weth *et al.*,1996; Sato 2007).. Thus, high levels of expression observed by RNA-seq appears not to reflect high level of expression on an individual cell level but rather the number of OSNs expressing a given OR gene.

4.5. Correlation of Expression values and Cell Counts

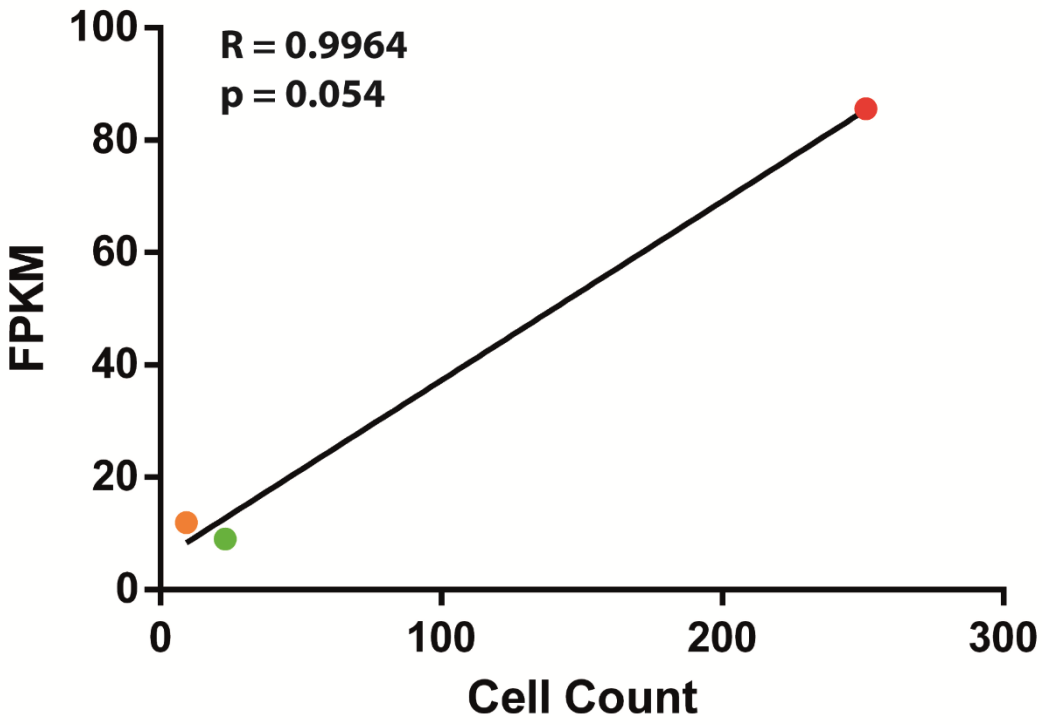

Figure 4. 18. Average cell numbers for whole OE sections and expression levels are depicted for or107-1, or101-1 and or132-5. R indicates strength of association, p indicates significance level for linear regression analysis.

To further substantiate that expression levels revealed by RNA-seq and number of OSNs expressing a given gene correlate, the number of cells which can be detected by *in situ* hybridization were compared with their corresponding expression (fpkm) levels. In one set of genes, the OR101-1, OR107-1 and OR132-5 were analyzed. The average number of cell counts

for individual 12µm sections through the OE and their corresponding FPKM values were analyzed (Figure 4.18.).

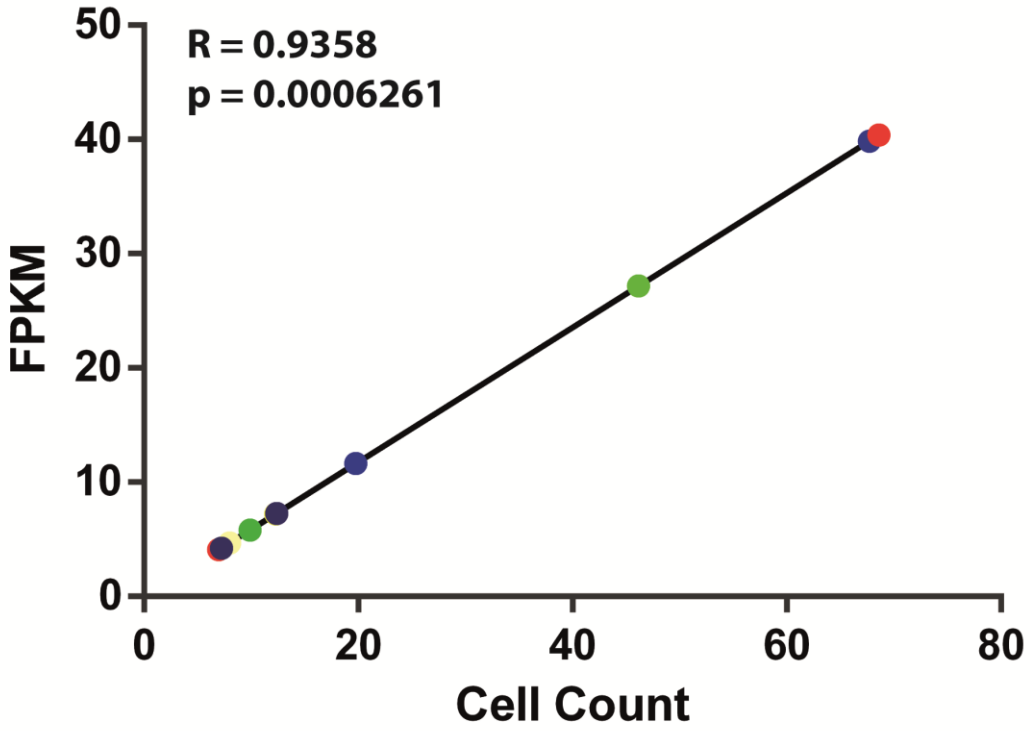

Figure 4. 19. Average cell numbers for whole OE sections and expression levels are depicted for or102-1, or103-1, or111-10, or 111-7, or111-5, or111-3, or111-2, or107-1 and or119-2. R indicates strength of association, p indicates significance level for linear regression analysis.

To extend the data set for this analysis, cell counts of the or102-1, or103-1, or111-10, or 111-7, or111-5, or111-3, or111-2, or107-1 and or119-2 genes, which were previously quantified by Sato *et al*. (2007) were compared to RNA-seq-derived FPKM values (Figure 4.19.). Linear regression analysis provided a strong correlation between RPKM value and cell number ($R=0.9358$, $p<0.05$), suggesting that the expression value of given OR might depend on the number of OSNs which express it.

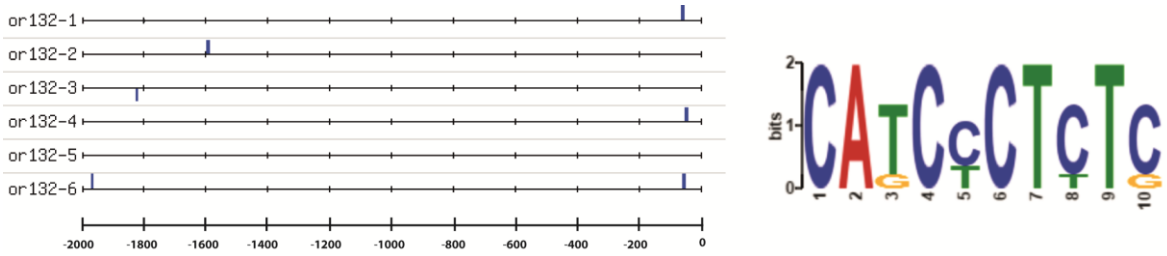

Figure 4. 20. Distribution of the de novo motif obtained from 2000 bp upstream regions of or132 family. On the right, positional weight matrix obtained from promoters of or132 family are shown.

Candidate promoter sequences of the or132 family were further analyzed in detail for the presence of conserved motifs, which may confer high level of expression. We analyzed the 2.000 bp and 500 bp promoter regions with Consensus and Oligo Analysis tools. When the full 2.000 bp sequences upstream of or132 genes was used for analysis, Oligo Analysis failed to detect any over-represented motif which could be represented as a PWM, while Consensus detected a motif with the core sequence "CATCCCTCTC". This motif is present in all of the or132 family 2.000 bp promoters except or132-5 and it is located in first 100 bp upstream regions of or132-1, or132-4 and 132-6 (Figure 4.20). Interestingly, two genes, or132-2 and or 132-5, contain introns in their 5'-sequence and the motif is present within the intronic sequence.

When 500 bp regions of or132 family were investigated, Consensus tool identified another motif with the core region "ACTGGCCCAG". This motif is present in all or132 promoters with at least one occurence.When 500 bp regions of or132 family were investigated with Oligo Analysis, a motif which was present in three members of or or132 promoters was obtained. Core region of this motif was "GGCTGCCC".Thus, the best candidate for an or132-specific motif that may confer high levels of expression was found with the CATCCCTCTC sequence.

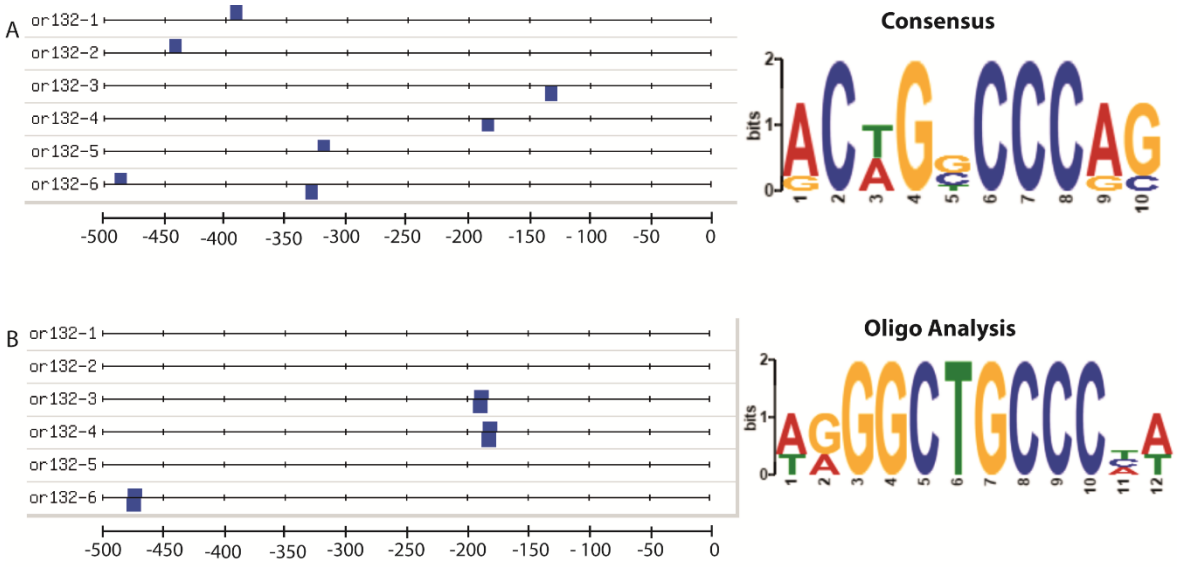

Figure 4. 21. A. Distribution of the Consensus tool de novo motif obtained from 500 bp upstream regions of or132 family (left). Positional weight matrix obtained by Consensus tool (right) B. Distribution of the Oligo Analysis tool de novo motif obtained from 500 bp upstream regions of or132 family (left). Positional weight matrix obtained by Oligo Analysis tool (right).

Thus, a matrix scan for the CAtCcCTcTc motif was performed on the first 2.000 bps upstream of all 161 OR genes and scored for the presence of the motif within close proximity to the TSS, either on the direct or reverse strand. The presence of the CAtCcCTcTc sequence could be identified within 200 bp distance from the TSS in a total of 31out of the 161 OR genes. In 24 OR genes the motif was located within the first 100 bps upstream of the TSS.

To quantify the result, ORs were grouped according to the presence or absence of the motif within the first 100bp and the average FPKM values for the different groups of genes was calculated (Figure 4.22). The average expression level of OR genes that contained the motif proximal to the TSS was about 3-times higher than in OR genes in which a similar motif is located more distally from the TSS. The difference between these two groups of OR genes was highly significant (two-tailed Student's t-test) while there was no statistical difference between the average expression levels of OR genes in which the motif was represented on the direct or

reverse strand. Thus, the results suggest that CAtCcCTcTc motif might be causative for the observed high-level expression of these OR genes.

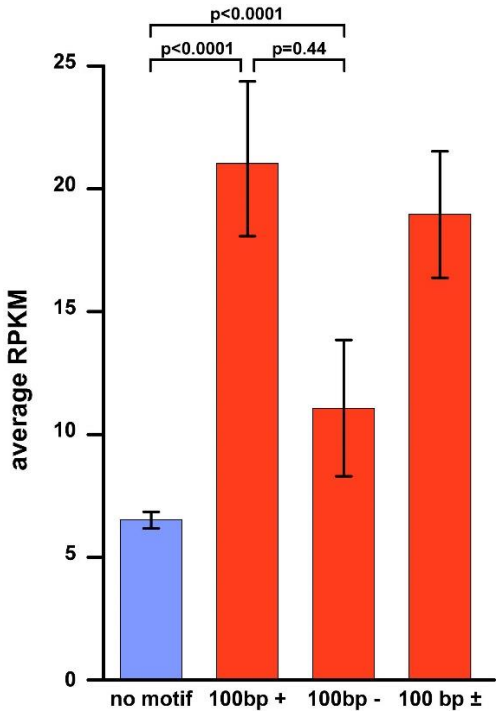

Figure 4. 22. Comparison of average RPKM values for the genes which contain or132 motif and the remining ones.

The identification of a high expression-related motif is interesting. To provide functional data on the motif, the 1.000 bp promoter sequences of the or132-5 and or132-6 genes were cloned into pGEM-T Easy vector and verified by sequencing. However, completion of an expression construct and injection into zebrafish oocytes for expression analysis awaits further experimentation.

In summary, various aspects of the transcriptome data obtained by RNA-seq from OE and brain were analyzed. The analysis successfully pinpointed the TSS of 161 OR genes and

the corresponding transcript structure of the genes was identified. When regulatory motifs within the candidate promoter sequences of these genes was analyzed, the presence of the Olf1/EBF1 motif could be detected by unbiased search. The distribution of this and the TBP, Lhx2 and Emx2 motifs appeared to be similar to the occurrence of these sites in other systems. Likewise, the presence of a BPTF motif, which was recently shown to have a function in OR gene regulation, emerged from the analysis. Analysis of a highly expressed or gene, revealed that FPKM values correlate better with cell numbers than expression per cell, suggesting that RNA expression levels can predict the number of OSNs which express a given OR.

5.DISCUSSION

Evolution devised a complex mechanism to generate the cellular diversity in the nasal epithelium that underlies odor perception. The main attribute of this diversity is the singular expression of OR genes at the level of individual OSNs, where each OSN expresses only one out of many possible OR genes (Malnic *et al*., 1996) and only a single allele of it (Chess *et al*., 1994). This "one neuron-one receptor rule" is ensured by a combination of a variety of cellular mechanisms: the interaction of transcription factors with specific DNA-Binding motifs, the activity of Locus Control Regions (Serizawa *et al*., 2003; Khan *et al*., 2012), epigenetic mechanisms, such as histone (de-) methylation (Lyons *et al*., 2013, Dalton *et al.*, 2013), and GPCR-specific signaling pathways (Ferreira *et al*., 2014). Even though, some parts of this puzzle have been studied in detail, our understanding of how OR gene expression is controlled is still not complete.

The enigma of one neuron-one receptor rule becomes even more striking when the size of the OR gene repertoire is considered: the OR family comprises 1400 OR genes in the mouse, 1000 in dogs, 800 in humans and 171 in the zebrafish. (Malnic *et al*., 2004; Olender *et al*., 2004; Zhang *et al*., 2004; see Table 4.1.). In addition to the initial choice of an OR gene for expression, each OSN must ensure that expression of the OR persists throughout the life span of the cell. Thus, the initial choice of the OR must be tightly regulated, so must be the mechanisms that prevent the expression of additional ORs in the same OSN. OSNs that express the same OR form synaptic relay structures in the olfactory bulb called glomeruli (Mombaerts *et al*., 1996; Wang *et al*., 1998). Thus, odorant-specific input to glomeruli is preserved by stable OR gene choice demonstrating the importance of the one neuron - one receptor rule at the physiological level.

Several cis-activating elements have been identified in proximal OR gene promoters (Plessy *et al*., 2012; Rothman *et al*, 2005, Vassalli *et al*., 2002, 2011; Qasba and Reed, 1998). Of those Olf1 / EBF1 recognition sites and homeodomain-like sites are the most prominent and best studied motifs in OR promoters (Hirota and Mombaerts 2004; Hirota *et al*,. 2007; Rothman *et al*., 2005; Michaloski *et al*., 2006). It has been shown that the homeodomain transcription factors Emx2 and Lhx2 bind to homeodomain-like sites and instruct expression of specific OR subsets (Hirota *et al*., 2007; McIntyre *et al*., 2008). Thus, binding of specific combinations of transcription factors may be an early upstream event that guides downstream mechanisms, such as epigenetic modification to specific OR subsets.

In addition to proximal promoter elements, two long-range *cis*-acting genomic elements, the H and P elements, have been identified experimentally in the mouse. (Serizawa *et al*., 2003; Fuss *et al*., 2007; Nishizumi *et al*., 2007; Khan *et al*., 2012). It has been shown that these enhancer-like sequences regulate expression of OR genes from adjacent gene clusters. Recently, a large number of similar elements has been functionally identified in the mouse genome, all of which are associated with OR gene clusters (Markenscoff-Papadimitriou *et al*., 2014). It has been suggested that the physical interaction of these enhancer regions in the nucleus could act as a hub for OR gene choice and that these nuclear aggregates facilitate the interaction of LSD1 with OR gene loci to regulate their expression.

In this model of OR gene regulation, epigenetic repression, de-repression, and subsequent re-repression of OR gene loci plays a crucial role (Lyons *et al*., 2013). In immature OSNs repressive histone marks, indicative of constitutive heterochromatin, are enriched around all OR loci, suggesting that OR subgenome wide repression is a crucial first step in OR gene choice. The specific lysine demethylase 1, LSD1, then releases individual OR gene loci from repression at low rate through H3K9 demethylation. This would allow for the expression of a single or a few OR genes per OSNs, provided LSD1 activity is shut down quickly after the first OR gene is expressed at sufficiently high levels. It has been suggested that LSD1 activity is downregulated through components of the unfolded protein response pathway downstream of

baseline OR signaling, which acts as a negative feedback signal of OR gene choice and prevents further OR expression (Dalton *et al*., 2013).

This powerful model may explain singular OR expression and coordination among the 46 different OR gene clusters in the mouse genome, however, it cannot explain more detailed aspects of OR expression. For instance, OR genes are expressed in restricted spatial domains in the OE (Ressler *et al*., 1993; Vassar *et al*., 1993, Weth *et al*., 1996) and different ORs are expressed with vastly different frequencies (Bressel *et al*., 2015; Weth *et al*., 1996; Sato *et al*., 2007). Thus, additional mechanisms, such as regulation by zone-specific transcription factors, or combinations thereof, must be n place to guide LSD1 activity to specific OR gene subsets. It is therefore highly likely that the epigenetic mechanism works in concert with other, less well understood pathways and regulatory factors, such as proximal DNA-binding elements and long-range regulators to generate the full range of diversity of OR expression.

5.1. A bioinformatic approach to uncover proximal regulatory sequences

Here, I set out to analyze genomic upstream regions of OR genes in the zebrafish to uncover new candidate regulators of OR expression that may act upstream of the epigenetic control mechanisms described above. In order to detect conserved motifs across the zebrafish OR promoter repertoire, knowledge of their precise TSSs was required. RNA-seq offers distinct advantages over alternative strategies, such as 5'-RACE. Because the OR gene family is very large, obtaining structural information on a large number of OR transcripts by 5'-RACE is labor intensive and time consuming (Hoppe *et al*., 2006; Michaloski *et al*,. 2006). The recently developed nanoCAGE technology is an efficient alternative method for TSS detection, however, it does not report expression levels of the respective transcripts (Plessy *et al*. 2010; Plessy *et al*., 2011). On the other hand, alternative quantitative methods, such as custom-tiling microarray analysis, lacks sufficient detail at the TSS and intron-exon junctions (Clowney *et al*. 2011).

RNA-seq, however, provides high-resolution structural data at a quantitative level, which allows for a comprehensive understanding of the olfactory transcriptome (Trapnell *et al*., 2012).

A biological drawback for all of these analysis is the fact that each OSN only express a single OR gene from the entire repertoire, thus each OR transcript is represented only at a low level within the complex OE tissue. The zebrafish is estimated to express around 450 different chemoreceptor genes (ORs, TAARs, VRs), potentially in a singular fashion, thus the expression level of individual ORs are expected to be low, when compared to other, more widely expressed genes that play a role in olfactory function.

Thus, a critical concern for the approach taken here was to generate sufficient sequencing depth to obtain structural data even for OR genes that are expressed at relatively low levels. A post-hoc analysis of the sequencing data and the number of OSNs expressing a variety of ORs revealed a rather tight correlation and linear relationship between FPKM values obtained from RNA-seq and the number of OSNs. Similar results were obtained in two recent studies where mouse OR transcripts were compared to OSN number using nanostring (Khan *et al*., 2012) and in zebrafish using RNA-seq (Saraiva *et al*., 2015). Thus, low expression levels revealed by RNA-seq may represent OR genes that are only expressed by few OSNs and less than the average.

From the initial sequencing data at 4 GB depth, 179 loci of transcription corresponding to OR genes could be revealed in the zebrafish genome. However, transcript structures could not be revealed for all of these genes due to low coverage of certain ORs. The enlarged data set with 12 GB of clean reads allowed for high quality identification of transcripts for 161 OR genes. For 10 additional ORs RNA-seq reads could be mapped to specific loci, however, the mapping was sparse and there were not a sufficient coverage to resolve transcript structures. Those genes may be expressed by a very low number of OSNs, the reason for which remains unclear. It could be that these ORs represent receptors with very high affinity for their ligand

and that they are able to detect specific compounds in a labeled line fashion (Lemon and Katz, 2007) where even low concentrations of the ligand triggers specific behavioral responses. For instance, a comparably small glomerulus has been shown to respond to the mating pheromone prostaglandin (Friedrich and Korsching, 1998). Alternatively, these ORs could have lost their relevance over evolutionary time and deterioration of the respective promoter region. Interestingly, the greatest distinction between expression levels was observed across chromosomes or OR clusters and the variability between clusters appeared to be more consistent than within clusters. This could indicate that cluster-specific locus control regions may be absent or deteriorated within certain regions of the zebrafish OR subgenome.

For the remaining 8 of the 179 OR genes, no RNA-seq reads were observed at all. Thus, from a transcriptional point of view, these genes can be considered pseudogenes. Interestingly, no nonsense mutations could be detected in the underlying genomic sequence, which would be expected if disrupting mutations were acquired within the promoter sequence. Thus, the (unknown) mutations in the promoter of these genes could have been acquired rather recently, or these ORs may serve other functions in tissues outside the OE. Such a extra-olfactory function has been shown for some ORs in the mouse kidney, where these receptors may be involve in sensing specific signals within the filtrate at the level of the macula densa (Pluznick *et al*., 2009).

A limited comparison of OR transcripts revealed by RNA-seq with previously annotated transcript structures in Ensembl and RefSeq database and those obtained by 5’ RACE indicated that the RNA-seq approach resulted in structures with better resolution. Thus, in sum, the approach taken here is plausible and, for a large fraction of the OR repertoire, revealed reliable transcript structures including the TSSs and alternatively spliced products.

5.2. Motif search for candidate regulators

The analysis obtained candidate promoter sequences for 161 OR genes for which enriched sequence motifs were searched using a variety of publically available tools. Genomic upstream sequences with various lengths ranging from 500 bp to 2000 bp were investigated using the RSAT suite. To narrow down the analysis it was decided to focus on the first 500 bp upstream of the TSS since previously identified motifs in the mouse were concentrated within this interval (Plessy *et al*., 2012).

Even though, several enriched oligomers (6-8 bp sequences) could be obtained, no consisted pattern emerged when positional preference of these motifs in OR upstream sequences was analyzed. The presence, absence, or density of these motifs was not different for highly expressed OR genes or OR genes with low expression levels. On the other hand, some of these oligomers could be considered as a source for a PWM. However, a clear motif that came out of the analysis was the zebrafish Olf1/Ebf1 motif. Interestingly, this motif has a similar distribution around the TSS as the mouse homologue and is found predominantly within the first 100 bp upstream of the TSS (Plessy *et al*., 2012). Similary, a homeodomain-like site appears to be present in a large number of OR promoters, although with variable distance to the TSS. Again, this is similar to observations in the mouse (Plessy *et al*., 2012), where the exact distance between Olf1/Ebf1 and homeodomain sites has been speculated to pose a specific instruction for OR expression (Vassalli *et al*., 2001; Vassalli *et al*., 2011). However, because of the lack of functional data these interpretations are purely speculative.

5.3. The Olf1/Ebf1 and HD sites: general or specific regulators of OR expression?

The Olf1/Ebf1 site has been shown to be enriched around OR TSSs, yet it is also present in the promoters of a variety of genes that are expressed in all OSNs, such as the G-protein subunit Gaolf (Jones and Reed 1989), the adenylate cyclase type III (Bakalyar and Reed 1990), or the olfactory specific nucleotide channel OCNC1 (Wang and Reed 1993). What function, other than instructing genral OE-wide expression could the Olf1/Ebf1 motif have? Interestingly, Olf1/Ebf1 sites have also been identified in the two characterized cluster control regions H and P (Nishizumi *et al*., 2007; Vassalli *et al.*, 2011). Thus Olf1/Ebf1 sites could be important DNA motifs for the interaction between enhancers and promoters from the associated OR gene cluster. As seen in other species (Weth *et al*., 1996; Sato *et al*., 2007) and the analysis presented here, different ORs are expressed with highly different frequency. Thus, stronger or weaker interactions between OR promoters and cluster control elements could influence the probability of choice of any given OR gene. It has been shown in the mouse that OR gene clusters from different chromosomes locate to a specific nuclear hub and that elements similar to H and P are important for the recruitment to this nuclear position (Markenscoff-Papadimitriou *et al*., 2014). Thus, distance of an OR TSS to the nuclear hub at which enhancer interaction takes place my influence OR gene choice. Indeed distance-dependent effects of promoter-enhancer interactions on OR expression have been described (Serizawa *et al*., 2003; Fuss *et al*., 2007). In this light, it may be meaningful to correlate the distance of Olf1/Ebf1 sites to the TSS with level of OR expression, a route not followed in this thesis.

In the mouse, OR genes are expressed in specific zones (Ressler *et al*., 1993; Vassar *et al*., 1993) but the functional significance of these zones and the mechanisms that generate the OR expression pattern remain elusive. Combinatorial codes of transcription factors could account for zonal expression, provided that these factors are expressed in non-congruent patterns across the OE. The Olf1/Ebf1 sites could be part of this code. Interesting in this light is the conservation of Olf1/Ebf1 sites around OR TSSs in zebrafish. Even though a zonal pattern of

OR expression, reminiscent of the expression zones in the mouse, has been described in the literature (Weth *et al*., 1996), recent work from our laboratory has shown that the restricted pattern of OR expression in zebrafish is an epiphenomenon (Bayramli, unpublished). OSNs are born from distinct regions of high proliferation and migrate across the OE. Thus, the sites of onset of OR expression (OR gene choice) and final position of OSNs expressing this OR are not identical. Therefore, it is unlikely that the Olf1/Ebf1 site itself, or in combination instructs spatial aspects of OR expression, at least within the zebrafish OE.

The Lhx2 and Emx2 transcription factors that bind to the homeodomain site in mouse OR promoters have been shown to regulate expression of specific subsets of ORs. In Lhx2-deficient mice, class II ORs are not expressed, while class I ORs are largely spared (Bozza *et al*., 2009). In Emx2 null mice, specific class II OR subsets are downregulated or missing while others compensate for the loss (McIntyre *et al*., 2008). About 60 % of zebrafish OR genes have a homeodomain binding motif within the first 200 bp upstream of the TSS, suggesting that the motif could serve a similar function in zebrafish. However, all but one of the 171 zebrafish OR genes are more related to class I ORs, while OR101-1 may be an exception as it is more related to mammalian class II genes (Niimura and Nei, 2005; Alioto and Ngai, 2005). Thus, if specific OR subsets were specified by homeodomain transcription factors, the distinction would not be between class I and class II ORs, but within the class I repertoire. However, no clear correlation between genomic location and presence / absence of homeodomain sites emerged from this analysis, suggesting that the dependence on the binding factor is not a result of simple gene duplication and subfamily expansion during OR evolution. Nevertheless, the OR genes affected in the Emx2 knock out background are also largely unlinked (McIntyre *et al*., 2008).

It remains unclear, which factors actually bind to the homeodomain motif in zebrafish. Specific morpholino oligonucleotides against the zebrafish homologues of Olf1, Lhx2, and Emx2 have been designed but have not revealed conclusive results and the identification of factors, along with the specific phenotypes in loss-of-function studies awaits further experimentation.

5.4. De novo search for motifs

For the *de novo* identification of motifs one very important aspect should be considered in more detail. In order to calculate the significance of the occurrence of a motif an appropriate background model is required. Thus, the choice of the background model may affect the sensitivity of the approach. The RSAT tool allows for the choice of different background models, such as custom background models, equiprobable residues models, Markov models, and genome subset based on background. To minimize the number of false positive hits, the *Danio rerio* genome subset was chosen under "upstream-noorf" conditions since only investigated upstream regions of OR TSSs were investigated. Since zebrafish intergenic regions are AT-rich, background model calibrated on zebrafish sequences

5.5. Candidate factor analysis

Likewise the statistical threshold for weight scores must be chosen properly when PWM are used. Since PWMs mostly contain a core region and the more varied flanking sequences, Matrix Scan might yield false positive matches, which are only similar to flanking sequences but not the core region without proper weight score threshold. Thus, low thresholds result in high number of false positive hits, however increasing the threshold might also cause loss of true positives. Various values of weight score threshold were applied in Matrix Scan in order to empirically determine the proper cutoff, which prevent false positives in most cases. It was observed that threshold value of 4 prevents false positives in most cases.

When the 500 bp uostream sequences were analyzed, TBP (TATA-box binding protein) motifs could be identified within the first 50 bp upstream of 46% of promoters. The detection of TATA-box motifs in a subset of ORs further supports the validity of TSS identification by RNA-seq. Yet, there was not difference in the level of expression of OR genes that did or did not contain a TBP motif at the group level. Similar observations were made for eukaryotic genes

in general (Shi and Zhou, 2006) and for OR genes in particular (Plessy *et al*., 2012). In the mouse, about XXX% of OR genes contained a significant TBP motif, while the remaining OR genes lacked the motif. The TBP is a component of TFIID, which loads the RNA polymerase II onto the promoter at the initiation of transcription (Lee and Young, 2000). Thus, more efficient initiation of transcription should result in higher transcriptional activity. This, however, should be true at the individual cell level. Our analysis revealed that quantitative levels of transcription are tightly correlated with the number of OSNs expressing a given OR gene. The variability at the individual cell level is unknown and could only be revealed by single cell apporaches, such as single cell RNA-seq, nanostring, or qRT-PCR. It is currently unknown is as far OR transcription levels per cell affect OR gene choice at the OSN population level. Considering the LSD1-driven mechanism, OR expression levels per cell should not result in biased OR gene choice, although it could result in increased OR coexpression due to delayed feedback signaling.

When the presence and positional preferences of Ebf1, Lhx2 and Emx2 motifs were investigated, all of them could be identified in most of OR promoters. The Ebf1 motifs was present in 81.3 % of OR genes and tend to accumulate between -50 and -100 bp, while Lhx2 and Emx2 motifs mostly located in between -100 and -150 bp in the 60% of ORs in which the motif could be detected. Presence of these motifs and their positional preference indicated the similarity of zebrafish OR promoter architecture to mouse OR promoters (Plessy *et al*., 2012). These results suggested that regulation of OR gene expression in zebrafish is similar in terms of TFs and promoter motifs to its mouse counterpart (see above).

5.6. Long range interaction

When expression levels of OR genes across the genome is considered, OR expression levels are variegated between chromosomes and within clusters. This could be due to the uneven distribution of LCR-like elements and the preferential expression of OR genes located in close proximity to these sequences. Thus, modulation of OR expression along a cluster may pinpoint

sequences with enhancer function. It will be interesting to test this possibility functionally in transgenic expression constructs, which have been shown to be sensitive assays in zebrafish, even across species (Nishizumi *et al*., 2007; Markenscoff-Papadimitriou *et al*., 2014).

Investigation a BPTF motif, suggested by the unususally high expression of BPTF in the OE relative to brain tissue also resulted in a striking pattern. It was observed that the regions with the highest occurrence of BPTF motifs were spaced 150 bp apart from each other. BPTF is a part of NURF chromatin remodelling complex and it binds to the acetylated and methylated histone tails through bromo and PHD-finger domains (Ruthenberg *et al*., 2011). This spatial organization of BPTF suggested that nucleosome reorganizing activity of BPTF might have a role in OR gene regulation. Interestingly, BPTF has recently been implicated in OR gene choice (Markenscoff-Papadimitriou *et al*., 2014), where it appears to coordinate the clustering of enhancer elements at specific nuclear hubs. In vivo footprinting experiments showed that BPTF interacts with enhancer regions. Furthermore, reduction in OR expression observed in BPTF knock-outs (Markenscoff-Papadimitriou *et al*., 2014). Investigation of BPTF motif in long range regulatory elements detected occurences of the motif. However, it should be expected that strong BPTF binding sites should occur more frequently within long range regulatory elements, as those are coordinately recruited to nuclear sites.

Thus, the entire 150 kb region on chromosome 15, for which two candidate enhancer sequences have been pinpointed (Nishizumi *et al*., 2007) was investigated for the presence of EBF1 and BPTF motif to check whether enrichment for these motif could indicate the presence of LCRs and/or highly expressed genes within the cluster. Interestingly, the regions containing highly expressed genes in a cluster coincides with BPTF-rich regions. In the genomic location which are enriched for BPTF binding sites, the gene with the highest expression level of the cluster, or111-7, was located. Sites of neighboring Ebf1 and BPTF sites were located in close proximity of another highly expressed gene, or119-2, and the E2 enhancer region. However, no highly expressed gene (local maxima) were observed in the region with only Ebf1 enrichment. These result suggested the possibility that BPTF binding may influence OR gene expression

levels or OR gene choice, probably in concert with Ebf1 binding. Mutation analysis in promoter-enhancer expression constructs and morpholino knockdown of BPTF could shed further light onto this possibility.

5.7. A candidate motif of high probability of choice

When the promoter regions of the highly expressed or132 family was investigated, family-specific motifs could be revealed. The motif was present in first 100 bp upstream regions of four members of this family and in the intronic sequence of the remaining two. At least for one member of the family, or132-4, an unprecedented pattern of expression could be observed and the number of OSNs expressing or 132-4 was up to 10 times higher than for any other highly expressed OR. Furthermore, at the level of the entire OR repertoire, genes which contain the motif in close proximity to the TSS showed an on average higher expression levels than genes in which the motif was absent. Thus, this motif is probably the strongest candidate for further functional studies. Another factor might also be causative for high level expression of or132 family is the presence of a LCR. Since this family is organized into a compact cluster of 60 kb size and surrounded by non-OR neighboring genes, the presence of adjacent long-range elements might influence this family strongly. Thus, an LCR might be affecting expression in concert with more proximal elements that endow the OR genes to interact strongly with the LCR.

In summary, TSS and transcript structure of 161 OR genes were identified by analysis of transcriptome data obtained from OE tissue. Investigation for regulatory motifs on the promoter regions of these genes resulted in identification of the Ebf1, TBP, Lhx2, Emx2 and BPTF motifs and their corresponding distribution among these regions. Furthermore, a candidate motif was identified in highly expressed OR gene family. To find out the definite role of these motifs in OR gene regulation expression reporter assays and yeast one-hybrid screenings could be performed.

APPENDIX A: EQUIPMENT

Table A.1. Equipment.

4 °C Room	Birikim Elektrik, Turkey
Autoclaves	Astell Scientific, UK
Centrifuge	Eppendorf, Germany (5417R)
Confocal Microscope	Leica SP5-AOBS, USA
Electronic Balance	Sartorius, Germany (TE412)
Electrophoresis Supplies	Bio-Rad Labs, USA (ReadySub-Cell GT Cells)
Fluorescence Microscope	Leica Microsystems, USA (MZ16FA)
Freezer 1 -20 °C	Arçelik, Turkey
Freezer 2 -80 °C	Thermo Electron Corp., USA (Farma 723)
Gel Documentation	Bio-Rad Labs, USA (GelDoc XR)
Glass Bottles	Isolab, Germany
Incubator 1	Weiss Gallenkamp, UK
Incubator 2	Nuve, Turkey
Incubating Shaker	Thermo Electron Corp., USA
Micropipetters	Eppendorf, Germany (Research)
Microwave Oven	Vestel, Turkey
Microinjector	Eppendorf, Germany (FemtoJet)
Luminometer	Fluroskan Ascent Fl (Thermo Scientific)
Refrigerator	Arçelik, Turkey
Softwares	Vector NTI (Invitrogen, USA)
Thermal Cyclers	Bio-Rad Labs, USA (C1000)
Vortex	Scientific Industries, USA

APPENDIX B: SUPPLIES

Table B.1. List of Supplies.

1 kb DNA Ladder	New England Biolabs, U.S.A. (N3232)
100 bp DNA Ladder	New England Biolabs, U.S.A. (N3231)
5X GoTaq Flexi Buffer	Clontech, U.S.A. (639201)
Advantage 2 Polymerase Mix	Promega, U.S.A. (M890A)
BamHI	New England Biolabs, U.S.A. (R0136 L)
EcoRI	New England Biolabs, U.S.A. (R0101 M)
EcoRV	New England Biolabs, U.S.A. (R0195 L)
Ethanol Absolute	Sigma-Aldrich, U.S.A. (34870)
Ethidium Bromide	Sigma Life Sciences, U.S.A. (E1510-1 ml)
EDTA Disodium Salt	Sigma-Aldrich., U.S.A. (E5134 - 1 kg).
Glycerol	Sigma-Aldrich, U.S.A. (G5516-500 ml)
GoTaq Flexi DNA Polymerase	Promega, U.S.A. (M830B)
LB Agar	Sigma Life Sciences, U.S.A. (SL08394)
LB Broth	Sigma-Aldrich, U.S.A. (L7658- 1 kg)
Magnesium Chloride, 25 mM	Promega, U.S.A. (A3511)
Magnesium Sulfate	Sigma-Aldrich, U.S.A. (M7506)
NcoI	New England Biolabs, U.S.A. (R0193 L)
NotI	New England Biolabs, U.S.A. (R0189 L)
pGEM®-T Easy Vector System	Promega, U.S.A. (A1360)

Table B.2. List of Supplies (cont.).

Potassium Chloride	Sigma-Aldrich, U.S.A. (P9541)
PstI	New England Biolabs, U.S.A. (R0140 L)
SalI	New England Biolabs, U.S.A. (R0138 L)
SeaKem® Agarose	Cambrex, U.S.A. (50004)
Sodium Acetate	Sigma-Aldrich, U.S.A. (S8625)
Sodium Chloride	Sigma-Aldrich, U.S.A. (S7653 - 1 kg)
Sodium Hydroxide	Sigma-Aldrich, U.S.A. (S8045 - 1 kg)
SpeI	New England Biolabs, U.S.A (R0133 L)
SphI	New England Biolabs, U.S.A (R0182 L)
T4 DNA Ligase	New England Biolabs, U.S.A (M0202L)
Trizma® Base	Sigma-Aldrich, U.S.A. (T6066)
XhoI	New England Biolabs, U.S.A. (R0146 L)

REFERENCES

Adler, J., 1966, "Chemotaxis in Bacteria", *Science,* Vol. 153, No. 3737l, pp. 708-716.

Ahuja, G., S. Bozorg Nia, V. Zapilko, V. Shiriagin, D. Kowatschew, Y. Oka, and S. I. Korsching, 2014, "Kappe Neurons, a Novel Population of Olfactory Sensory Neurons", *Scientific Reports,* Vol. 4, No. 4037.

Alioto, T. S., and J. Ngai, 2005, "The Odorant Receptor Repertoire of Teleost Fish", *BMC Genomics,* Vol. 6, No. 173.

Baier, H., and S. Korsching, 1994, "Olfactory Glomeruli in the Zebrafish Form an Invariant Pattern and Are Identifiable across Animals", *Journal of Neuroscience,* Vol. 14, No. 1l, pp. 219-230.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, 2009, "Meme Suite: Tools for Motif Discovery and Searching", *Nucleic Acids Research,* Vol. 37, No. Web Server issue, pp. W202-208.

Bakalyar, H. A., and R. R. Reed, 1990, "Identification of a Specialized Adenylyl Cyclase That May Mediate Odorant Detection", *Science,* Vol. 250, No. 4986l, pp. 1403-1406.

Belluscio, L., G. Koentges, R. Axel, and C. Dulac, 1999, "A Map of Pheromone Receptor Activation in the Mammalian Brain", *Cell,* Vol. 97, No. 2l, pp. 209-220.

Ben-Shaul, Y., L. C. Katz, R. Mooney, and C. Dulac, 2010, "In Vivo Vomeronasal Stimulation Reveals Sensory Encoding of Conspecific and Allospecific Cues by the Mouse Accessory Olfactory Bulb", *Proceedings of the National Academy of Sciences U S A,* Vol. 107, No. 11l, pp. 5172-5177.

Bozza, T., P. Feinstein, C. Zheng, and P. Mombaerts, 2002, "Odorant Receptor Expression Defines Functional Units in the Mouse Olfactory System", *Journal of Neuroscience,* Vol. 22, No. 8l, pp. 3033-3043.

Bozza, T., A. Vassalli, S. Fuss, J. J. Zhang, B. Weiland, R. Pacifico, P. Feinstein, and P. Mombaerts, 2009, "Mapping of Class I and Class Ii Odorant Receptors to Glomerular Domains by Two Distinct Types of Olfactory Sensory Neurons in the Mouse", *Neuron,* Vol. 61, No. 2l, pp. 220-233.

Braubach, O. R., A. Fine, and R. P. Croll, 2012, "Distribution and Functional Organization of Glomeruli in the Olfactory Bulbs of Zebrafish (Danio Rerio)", *Journal of Comparative Neurology,* Vol. 520, No. 11l, pp. 2317-2339, Spc2311.

Bressel, O. C., M. Khan, and P. Mombaerts, 2015, "Linear Correlation between the Number of Olfactory Sensory Neurons Expressing a Given Mouse Odorant Receptor Gene and the Total Volume of the Corresponding Glomeruli in the Olfactory Bulb", *Journal of Comparative Neurology,* Vol.

Buck, L., and R. Axel, 1991, "A Novel Multigene Family May Encode Odorant Receptors: A Molecular Basis for Odor Recognition", *Cell,* Vol. 65, No. 1l, pp. 175-187.

Bulger, M., M. A. Bender, J. H. van Doorninck, B. Wertman, C. M. Farrell, G. Felsenfeld, M. Groudine, and R. Hardison, 2000, "Comparative Structural and Functional Analysis of the Olfactory Receptor Genes Flanking the Human and Mouse Beta-Globin Gene Clusters", *Proceedings of the National Academy of Sciences U S A,* Vol. 97, No. 26l, pp. 14560-14565.

Catania, S., A. Germana, R. Laura, T. Gonzalez-Martinez, E. Ciriaco, and J. A. Vega, 2003, "The Crypt Neurons in the Olfactory Epithelium of the Adult Zebrafish Express Trka-Like Immunoreactivity", *Neuroscience Letters,* Vol. 350, No. 1l, pp. 5-8.

Celik, A., S. H. Fuss, and S. I. Korsching, 2002, "Selective Targeting of Zebrafish Olfactory Receptor Neurons by the Endogenous Omp Promoter", *Eur Journal of Neuroscience,* Vol. 15, No. 5l, pp. 798-806.

Chess, A., I. Simon, H. Cedar, and R. Axel, 1994, "Allelic Inactivation Regulates Olfactory Receptor Gene Expression", *Cell,* Vol. 78, No. 5l, pp. 823-834.

Cloonan, N., A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, *et al*., 2008, "Stem Cell Transcriptome Profiling Via Massive-Scale Mrna Sequencing", *Nature Methods,* Vol. 5, No. 7l, pp. 613-619.

Clowney, E. J., A. Magklara, B. M. Colquitt, N. Pathak, R. P. Lane, and S. Lomvardas, 2011, "High-Throughput Mapping of the Promoters of the Mouse Olfactory Receptor Genes Reveals a New Type of Mammalian Promoter and Provides Insight into Olfactory Receptor Gene Regulation", *Genome Research,* Vol. 21, No. 8l, pp. 1249-1259.

Dalton, R. P., D. B. Lyons, and S. Lomvardas, 2013, "Co-Opting the Unfolded Protein Response to Elicit Olfactory Receptor Feedback", *Cell,* Vol. 155, No. 2l, pp. 321-332.

Dulac, C., and R. Axel, 1995, "A Novel Family of Genes Encoding Putative Pheromone Receptors in Mammals", *Cell,* Vol. 83, No. 2l, pp. 195-206.

Emes, R. D., S. A. Beatson, C. P. Ponting, and L. Goodstadt, 2004, "Evolution and Comparative Genomics of Odorant- and Pheromone-Associated Genes in Rodents", *Genome Research,* Vol. 14, No. 4l, pp. 591-602.

Ferreira, T., S. R. Wilson, Y. G. Choi, D. Risso, S. Dudoit, T. P. Speed, and J. Ngai, 2014, "Silencing of Odorant Receptor Genes by G Protein Betagamma Signaling Ensures the Expression of One Odorant Receptor Per Olfactory Sensory Neuron", *Neuron*, Vol. 81, No. 4l, pp. 847-859.

Firestein, S., 2001, "How the Olfactory System Makes Sense of Scents", *Nature*, Vol. 413, No. 6852l, pp. 211-218.

Freitag, J., J. Krieger, J. Strotmann, and H. Breer, 1995, "Two Classes of Olfactory Receptors in Xenopus Laevis", *Neuron*, Vol. 15, No. 6l, pp. 1383-1392.

Freitag, J., G. Ludwig, I. Andreini, P. Rossler, and H. Breer, 1998, "Olfactory Receptors in Aquatic and Terrestrial Vertebrates", *The Journal of Comparative Physiology A*, Vol. 183, No. 5l, pp. 635-650.

Friedrich, R. W., and S. I. Korsching, 1998, "Chemotopic, Combinatorial, and Noncombinatorial Odorant Representations in the Olfactory Bulb Revealed Using a Voltage-Sensitive Axon Tracer", *Journal of Neuroscience*, Vol. 18, No. 23l, pp. 9977-9988.

Fuss, S. H., M. Omura, and P. Mombaerts, 2007, "Local and Cis Effects of the H Element on Expression of Odorant Receptor Genes in Mouse", *Cell*, Vol. 130, No. 2l, pp. 373-384.

Fuss, S. H., Y. Zhu, and P. Mombaerts, 2013, "Odorant Receptor Gene Choice and Axonal Wiring in Mice with Deletion Mutations in the Odorant Receptor Gene Sr1", *Molecular and Cellular Neuroscience*, Vol. 56, No. 212-224.

Garber, M., M. G. Grabherr, M. Guttman, and C. Trapnell, 2011, "Computational Methods for Transcriptome Annotation and Quantification Using Rna-Seq", *Nature Methods*, Vol. 8, No. 6l, pp. 469-477.

Germana, A., G. Montalbano, R. Laura, E. Ciriaco, M. E. del Valle, and J. A. Vega, 2004, "S100 Protein-Like Immunoreactivity in the Crypt Olfactory Neurons of the Adult Zebrafish", *Neuroscience Letters*, Vol. 371, No. 2-3l, pp. 196-198.

Gilad, Y., O. Man, and G. Glusman, 2005, "A Comparison of the Human and Chimpanzee Olfactory Receptor Gene Repertoires", *Genome Research*, Vol. 15, No. 2l, pp. 224-230.

Glusman, G., I. Yanai, I. Rubin, and D. Lancet, 2001, "The Complete Human Olfactory Subgenome", *Genome Research*, Vol. 11, No. 5l, pp. 685-702.

Go, Y., and Y. Niimura, 2008, "Similar Numbers but Different Repertoires of Olfactory Receptor Genes in Humans and Chimpanzees", *Molecular Biology and Evolution*, Vol. 25, No. 9l, pp. 1897-1907.

Grus, W. E., P. Shi, Y. P. Zhang, and J. Zhang, 2005, "Dramatic Variation of the Vomeronasal Pheromone Receptor Gene Repertoire among Five Orders of Placental and Marsupial Mammals", *Proceedings of the National Academy of Sciences U S A*, Vol. 102, No. 16l, pp. 5767-5772.

Hamdani el, H., and K. B. Doving, 2007, "The Functional Organization of the Fish Olfactory System", *Progress in Neurobiology*, Vol. 82, No. 2l, pp. 80-86.

Hansen, A., and E. Zeiske, 1998, "The Peripheral Olfactory Organ of the Zebrafish, Danio Rerio: An Ultrastructural Study", *Chemical Senses*, Vol. 23, No. 1l, pp. 39-48.

Hirota, J., and P. Mombaerts, 2004, "The Lim-Homeodomain Protein Lhx2 Is Required for Complete Development of Mouse Olfactory Sensory Neurons", *Proceedings of the National Academy of Sciences U S A*, Vol. 101, No. 23l, pp. 8751-8755.

Hoppe, R., H. Frank, H. Breer, and J. Strotmann, 2003, "The Clustered Olfactory Receptor Gene Family 262: Genomic Organization, Promotor Elements, and Interacting Transcription Factors", *Genome Research*, Vol. 13, No. 12l, pp. 2674-2685.

Hoppe, R., H. Frank, H. Breer, and J. Strotmann, 2003, "The Clustered Olfactory Receptor Gene Family 262: Genomic Organization, Promotor Elements, and Interacting Transcription Factors", *Genome Research*, Vol. 13, No. 12l, pp. 2674-2685.

Hoppe, R., M. Weimer, A. Beck, H. Breer, and J. Strotmann, 2000, "Sequence Analyses of the Olfactory Receptor Gene Cluster Mor37 on Mouse Chromosome 4", *Genomics*, Vol. 66, No. 3l, pp. 284-295.

Hussain, A., L. R. Saraiva, and S. I. Korsching, 2009, "Positive Darwinian Selection and the Birth of an Olfactory Receptor Clade in Teleosts", *Proceedings of the National Academy of Sciences U S A*, Vol. 106, No. 11l, pp. 4313-4318.

Ibarra-Soria, X., M. O. Levitin, L. R. Saraiva, and D. W. Logan, 2014, "The Olfactory Transcriptomes of Mice", *PLoS Genetics*, Vol. 10, No. 9l, pp. e1004593.

Ihara, S., K. Yoshikawa, and K. Touhara, 2013, "Chemosensory Signals and Their Receptors in the Olfactory Neural System", *Neuroscience*, Vol. 254, No. 45-60.

Ishii, T., S. Serizawa, A. Kohda, H. Nakatani, T. Shiroishi, K. Okumura, Y. Iwakura, F. Nagawa, A. Tsuboi, and H. Sakano, 2001, "Monoallelic Expression of the Odourant Receptor Gene and Axonal Projection of Olfactory Sensory Neurones", *Genes to Cells*, Vol. 6, No. 1l, pp. 71-78.

Iwema, C. L., and J. E. Schwob, 2003, "Odorant Receptor Expression as a Function of Neuronal Maturity in the Adult Rodent Olfactory System", *Journal of Comparative Neurology*, Vol. 459, No. 3l, pp. 209-222.

Jones, D. T., and R. R. Reed, 1989, "Golf: An Olfactory Neuron Specific-G Protein Involved in Odorant Signal Transduction", *Science,* Vol. 244, No. 4906l, pp. 790-795.

Kajiya, K., K. Inaki, M. Tanaka, T. Haga, H. Kataoka, and K. Touhara, 2001, "Molecular Bases of Odor Discrimination: Reconstitution of Olfactory Receptors That Recognize Overlapping Sets of Odorants", *Journal of Neuroscience,* Vol. 21, No. 16l, pp. 6018-6025.

Kanageswaran, N., M. Demond, M. Nagel, B. S. Schreiner, S. Baumgart, P. Scholz, J. Altmuller, C. Becker, J. F. Doerner, H. Conrad, *et al*., 2015, "Deep Sequencing of the Murine Olfactory Receptor Neuron Transcriptome", *PLoS One,* Vol. 10, No. 1l, pp. e0113170.

Khan, M., E. Vaes, and P. Mombaerts, 2011, "Regulation of the Probability of Mouse Odorant Receptor Gene Choice", *Cell,* Vol. 147, No. 4l, pp. 907-921.

Kobayakawa, K., R. Kobayakawa, H. Matsumoto, Y. Oka, T. Imai, M. Ikawa, M. Okabe, T. Ikeda, S. Itohara, T. Kikusui, *et al*., 2007, "Innate Versus Learned Odour Processing in the Mouse Olfactory Bulb", *Nature,* Vol. 450, No. 7169l, pp. 503-508.

Kolterud, A., M. Alenius, L. Carlsson, and S. Bohm, 2004, "The Lim Homeobox Gene Lhx2 Is Required for Olfactory Sensory Neuron Identity", *Development,* Vol. 131, No. 21l, pp. 5319-5326.

Kubick, S., J. Strotmann, I. Andreini, and H. Breer, 1997, "Subfamily of Olfactory Receptors Characterized by Unique Structural Features and Expression Patterns", *Journal of Neurochemistry*, Vol. 69, No. 2l, pp. 465-475.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009, "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome", *Genome Biology,* Vol. 10, No. 3l, pp. R25.

Le Hir, H., A. Nott, and M. J. Moore, 2003, "How Introns Influence and Enhance Eukaryotic Gene Expression", *Trends in Biochemical Sciences,* Vol. 28, No. 4l, pp. 215-220.

Lee, T. I., and R. A. Young, 2000, "Transcription of Eukaryotic Protein-Coding Genes", *Annual Review of Genetics*, Vol. 34, No. 77-137.

Lemon, C. H., and D. B. Katz, 2007, "The Neural Processing of Taste", *BMC Neuroscience*, Vol. 8 Suppl 3, No. S5.

Lewcock, J. W., and R. R. Reed, 2004, "A Feedback Mechanism Regulates Monoallelic Odorant Receptor Expression", *Proceedings of the National Academy of Sciences U S A,* Vol. 101, No. 4l, pp. 1069-1074.

Li, J., T. Ishii, P. Feinstein, and P. Mombaerts, 2004, "Odorant Receptor Gene Choice Is Reset by Nuclear Transfer from Mouse Olfactory Sensory Neurons", *Nature,* Vol. 428, No. 6981l, pp. 393-399.

Liberles, S. D., and L. B. Buck, 2006, "A Second Class of Chemosensory Receptors in the Olfactory Epithelium", *Nature,* Vol. 442, No. 7103l, pp. 645-650.

Liu, A. H., X. Zhang, G. A. Stolovitzky, A. Califano, and S. J. Firestein, 2003, "Motif-Based Construction of a Functional Map for Mammalian Olfactory Receptors", *Genomics,* Vol. 81, No. 5l, pp. 443-456.

Lyons, D. B., A. Magklara, T. Goh, S. C. Sampath, A. Schaefer, G. Schotta, and S. Lomvardas, 2014, "Heterochromatin-Mediated Gene Silencing Facilitates the Diversification of Olfactory Neurons", *Cell Reports*, Vol. 9, No. 3l, pp. 884-892.

Magklara, A., A. Yen, B. M. Colquitt, E. J. Clowney, W. Allen, E. Markenscoff-Papadimitriou, Z. A. Evans, P. Kheradpour, G. Mountoufaris, C. Carey, G. Barnea, M.Kellis,

S.Lomvardas, 2011, "An Epigenetic Signature for Monoallelic Olfactory Receptor Expression", *Cell,* Vol. 145, No. 4l, pp. 555-570.

Malnic, B., J. Hirono, T. Sato, and L. B. Buck, 1999, "Combinatorial Receptor Codes for Odors", *Cell,* Vol. 96, No. 5l, pp. 713-723.

Markenscoff-Papadimitriou, E., W. E. Allen, B. M. Colquitt, T. Goh, K. K. Murphy, K. Monahan, C. P. Mosley, N. Ahituv, and S. Lomvardas, 2014, "Enhancer Interaction Networks as a Means for Singular Olfactory Receptor Expression", *Cell,* Vol. 159, No. 3l, pp. 543-557.

Mathelier, A., X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Y. Chen, A. Chou, H. Ienasescu, *et al*., 2014, "Jaspar 2014: An Extensively Expanded and Updated Open-Access Database of Transcription Factor Binding Profiles", *Nucleic Acids Research,* Vol. 42, No. Database issue, pp. D142-147.

McIntyre, J. C., S. C. Bose, A. J. Stromberg, and T. S. McClintock, 2008, "Emx2 Stimulates Odorant Receptor Gene Expression", *Chemical Senses,* Vol. 33, No. 9l, pp. 825-837.

Meisami, E., 1989, "A Proposed Relationship between Increases in the Number of Olfactory Receptor Neurons, Convergence Ratio and Sensitivity in the Developing Rat", *Brain Research. Developmental Brain Research,* Vol. 46, No. 1l, pp. 9-19.

Michaloski, J. S., P. A. Galante, and B. Malnic, 2006, "Identification of Potential Regulatory Motifs in Odorant Receptor Genes by Analysis of Promoter Sequences", *Genome Research,* Vol. 16, No. 9l, pp. 1091-1098.

Michaloski, J. S., P. A. Galante, M. H. Nagai, L. Armelin-Correa, M. S. Chien, H. Matsunami, and B. Malnic, 2011, "Common Promoter Elements in Odorant and Vomeronasal Receptor Genes", *PLoS One,* Vol. 6, No. 12l, pp. e29065.

Miyamichi, K., S. Serizawa, H. M. Kimura, and H. Sakano, 2005, "Continuous and Overlapping Expression Domains of Odorant Receptor Genes in the Olfactory Epithelium Determine the Dorsal/Ventral Positioning of Glomeruli in the Olfactory Bulb", *Journal of Neuroscience,* Vol. 25, No. 14l, pp. 3586-3592.

Mombaerts, P., 1999, "Seven-Transmembrane Proteins as Odorant and Chemosensory Receptors", *Science,* Vol. 286, No. 5440l, pp. 707-711.

Mombaerts, P., 2004, "Odorant Receptor Gene Choice in Olfactory Sensory Neurons: The One Receptor-One Neuron Hypothesis Revisited", *Current Opinions in Neurobiology,* Vol. 14, No. 1l, pp. 31-36.

Mombaerts, P., F. Wang, C. Dulac, S. K. Chao, A. Nemes, M. Mendelsohn, J. Edmondson, and R. Axel, 1996, "Visualizing an Olfactory Sensory Map", *Cell,* Vol. 87, No. 4l, pp. 675-686.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008, "Mapping and Quantifying Mammalian Transcriptomes by Rna-Seq", *Nature Methods,* Vol. 5, No. 7l, pp. 621-628.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, 2008, "The Transcriptional Landscape of the Yeast Genome Defined by Rna Sequencing", *Science,* Vol. 320, No. 5881l, pp. 1344-1349.

Nedelec, S., I. Foucher, I. Brunet, C. Bouillot, A. Prochiantz, and A. Trembleau, 2004, "Emx2 Homeodomain Transcription Factor Interacts with Eukaryotic Translation Initiation Factor 4e (Eif4e) in the Axons of Olfactory Sensory Neurons", *Proceedings of the National Academy of Sciences U S A,* Vol. 101, No. 29l, pp. 10815-10820.

Newburger, D. E., and M. L. Bulyk, 2009, "Uniprobe: An Online Database of Protein Binding Microarray Data on Protein-DNA Interactions", *Nucleic Acids Research,* Vol. 37, No. Database issuel, pp. D77-82.

Ngai, J., M. M. Dowling, L. Buck, R. Axel, and A. Chess, 1993, "The Family of Genes Encoding Odorant Receptors in the Channel Catfish", *Cell,* Vol. 72, No. 5l, pp. 657-666.

Niimura, Y., and M. Nei, 2005, "Evolutionary Dynamics of Olfactory Receptor Genes in Fishes and Tetrapods", *Proceedings of the National Academy of Sciences U S A,* Vol. 102, No. 17l, pp. 6039-6044.

Niimura, Y., and M. Nei, 2007, "Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution", *PLoS One,* Vol. 2, No. 8l, pp. e708.

Niimura, Y., and M. Nei, 2007, "Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution", *PLoS One,* Vol. 2, No. 8l, pp. e708.

Nishizumi, H., K. Kumasaka, N. Inoue, A. Nakashima, and H. Sakano, 2007, "Deletion of the Core-H Region in Mice Abolishes the Expression of Three Proximal Odorant Receptor Genes in Cis", *Proceedings of the National Academy of Sciences U S A,* Vol. 104, No. 50l, pp. 20067-20072.

Norlin, E. M., M. Alenius, F. Gussing, M. Hagglund, V. Vedin, and S. Bohm, 2001, "Evidence for Gradients of Gene Expression Correlating with Zonal Topography of the Olfactory Sensory Map", *Molecular and Cellular Neuroscience,* Vol. 18, No. 3l, pp. 283-295.

Oka, Y., L. R. Saraiva, and S. I. Korsching, 2012, "Crypt Neurons Express a Single V1r-Related Ora Gene", *Chemical Senses,* Vol. 37, No. 3l, pp. 219-227.

Plessy, C., G. Pascarella, N. Bertin, A. Akalin, C. Carrieri, A. Vassalli, D. Lazarevic, J. Severin, C. Vlachouli, R. Simone, *et al*., 2012, "Promoter Architecture of Mouse Olfactory Receptor Genes", *Genome Research,* Vol. 22, No. 3l, pp. 486-497.

Pluznick, J. L., D. J. Zou, X. Zhang, Q. Yan, D. J. Rodriguez-Gil, C. Eisner, E. Wells, C. A. Greer, T. Wang, S. Firestein, *et al*., 2009, "Functional Expression of the Olfactory Signaling System in the Kidney", *Proceedings of the National Academy of Sciences U S A,* Vol. 106, No. 6l, pp. 2059-2064.

Prasad, B. C., and R. R. Reed, 1999, "Chemosensation: Molecular Mechanisms in Worms and Mammals", *Trends in Genetics,* Vol. 15, No. 4l, pp. 150-153.

Quignon, P., M. Giraud, M. Rimbault, P. Lavigne, S. Tacher, E. Morin, E. Retout, A. S. Valin, K. Lindblad-Toh, J. Nicolas, *et al*., 2005, "The Dog and Rat Olfactory Receptor Repertoires", *Genome Biol,* Vol. 6, No. 10l, pp. R83.

Rawson, N. E., J. Eberwine, R. Dotson, J. Jackson, P. Ulrich, and D. Restrepo, 2000, "Expression of Mrnas Encoding for Two Different Olfactory Receptors in a Subset of Olfactory Receptor Neurons", *Journal of Neurochemistry,* Vol. 75, No. 1l, pp. 185-195.

Ressler, K. J., S. L. Sullivan, and L. B. Buck, 1993, "A Zonal Organization of Odorant Receptor Gene Expression in the Olfactory Epithelium", *Cell,* Vol. 73, No. 3l, pp. 597-609.

Ressler, K. J., S. L. Sullivan, and L. B. Buck, 1994, "Information Coding in the Olfactory System: Evidence for a Stereotyped and Highly Organized Epitope Map in the Olfactory Bulb", *Cell,* Vol. 79, No. 7l, pp. 1245-1255.

Rothman, A., P. Feinstein, J. Hirota, and P. Mombaerts, 2005, "The Promoter of the Mouse Odorant Receptor Gene M71", *Molecular and Cellular Neuroscience,* Vol. 28, No. 3l, pp. 535-546.

Rouquier, S., S. Taviaux, B. J. Trask, V. Brand-Arpon, G. van den Engh, J. Demaille, and D. Giorgi, 1998, "Distribution of Olfactory Receptor Genes in the Human Genome", *Nature Genetics,* Vol. 18, No. 3l, pp. 243-250.

Royal, S. J., and B. Key, 1999, "Development of P2 Olfactory Glomeruli in P2-Internal Ribosome Entry Site-Tau-Lacz Transgenic Mice", *Journal of Neuroscience,* Vol. 19, No. 22l, pp. 9856-9864.

Saraiva, L. R., G. Ahuja, I. Ivandic, A. S. Syed, J. C. Marioni, S. I. Korsching, and D. W. Logan, 2015, "Molecular and Neuronal Homology between the Olfactory Systems of Zebrafish and Mouse", *Scientific Reports,* Vol. 5, No. 11487.

Sato, Y., N. Miyasaka, and Y. Yoshihara, 2007, "Hierarchical Regulation of Odorant Receptor Gene Choice and Subsequent Axonal Projection of Olfactory Sensory Neurons in Zebrafish", *Journal of Neuroscience,* Vol. 27, No. 7l, pp. 1606-1615.

Serizawa, S., K. Miyamichi, H. Nakatani, M. Suzuki, M. Saito, Y. Yoshihara, and H. Sakano, 2003, "Negative Feedback Regulation Ensures the One Receptor-One Olfactory Neuron Rule in Mouse", *Science,* Vol. 302, No. 5653l, pp. 2088-2094.

Shi, W., and W. Zhou, 2006, "Frequency Distribution of Tata Box and Extension Sequences on Human Promoters", *BMC Bioinformatics,* Vol. 7 Suppl 4, No. S2.

Shykind, B. M., S. C. Rohani, S. O'Donnell, A. Nemes, M. Mendelsohn, Y. Sun, R. Axel, and G. Barnea, 2004, "Gene Switching and the Stability of Odorant Receptor Gene Choice", *Cell,* Vol. 117, No. 6l, pp. 801-815.

Sosinsky, A., G. Glusman, and D. Lancet, 2000, "The Genomic Structure of Human Olfactory Receptor Genes", *Genomics,* Vol. 70, No. 1l, pp. 49-61.

Stephan, A. B., E. Y. Shum, S. Hirsh, K. D. Cygnar, J. Reisert, and H. Zhao, 2009, "Ano2 Is the Cilial Calcium-Activated Chloride Channel That May Mediate Olfactory Amplification", *Proceedings of the National Academy of Sciences U S A*, Vol. 106, No. 28l, pp. 11776-11781.

Strotmann, J., I. Wanner, T. Helfrich, A. Beck, C. Meinken, S. Kubick, and H. Breer, 1994, "Olfactory Neurones Expressing Distinct Odorant Receptor Subtypes Are Spatially Segregated in the Nasal Neuroepithelium", *Cell Tissue Research*, Vol. 276, No. 3l, pp. 429-438.

Sullivan, S. L., M. C. Adamson, K. J. Ressler, C. A. Kozak, and L. B. Buck, 1996, "The Chromosomal Distribution of Mouse Odorant Receptor Genes", *Proceedings of the National Academy of Sciences U S A*, Vol. 93, No. 2l, pp. 884-888.

Tanaka, E., T. Bailey, C. E. Grant, W. S. Noble, and U. Keich, 2011, "Improved Similarity Scores for Comparing Motifs", *Bioinformatics*, Vol. 27, No. 12l, pp. 1603-1609.

Taştekin, İ., 2012, *The Role of Odorant Receptor Coding Sequence in the Regulation of Odorant Receptor Transgene*, M. S. Thesis, Boğaziçi University

Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov, 2013, "Integrative Genomics Viewer (Igv): High-Performance Genomics Data Visualization and Exploration", *Briefings in Bioinformatics*, Vol. 14, No. 2l, pp. 178-192.

Tian, H., and M. Ma, 2008, "Activity Plays a Role in Eliminating Olfactory Sensory Neurons Expressing Multiple Odorant Receptors in the Mouse Septal Organ", *Molecular and Cellular Neuroscience*, Vol. 38, No. 4l, pp. 484-488.

Touhara, K., S. Sengoku, K. Inaki, A. Tsuboi, J. Hirono, T. Sato, H. Sakano, and T. Haga, 1999, "Functional Identification and Reconstitution of an Odorant Receptor in Single Olfactory

Neurons", *Proceedings of the National Academy of Sciences U S A,* Vol. 96, No. 7l, pp. 4040-4045.

Trapnell, C., L. Pachter, and S. L. Salzberg, 2009, "Tophat: Discovering Splice Junctions with Rna-Seq", *Bioinformatics,* Vol. 25, No. 9l, pp. 1105-1111.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, 2012, "Differential Gene and Transcript Expression Analysis of Rna-Seq Experiments with Tophat and Cufflinks", *Nature Protocols,* Vol. 7, No. 3l, pp. 562-578.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, 2010, "Transcript Assembly and Quantification by Rna-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation", *Nature Biotechnology,* Vol. 28, No. 5l, pp. 511-515.

Turatsinze, J. V., M. Thomas-Chollier, M. Defrance, and J. van Helden, 2008, "Using Rsat to Scan Genome Sequences for Transcription Factor Binding Sites and Cis-Regulatory Modules", *Nature Protocols,* Vol. 3, No. 10l, pp. 1578-1588.

van Helden, J., B. Andre, and J. Collado-Vides, 1998, "Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies", *Journal of Molecular Biology,* Vol. 281, No. 5l, pp. 827-842.

van Helden, J., B. Andre, and J. Collado-Vides, 2000, "A Web Site for the Computational Analysis of Yeast Regulatory Sequences", *Yeast,* Vol. 16, No. 2l, pp. 177-187.

van Helden, J., A. F. Rios, and J. Collado-Vides, 2000, "Discovering Regulatory Elements in Non-Coding Sequences by Analysis of Spaced Dyads", *Nucleic Acids Research,* Vol. 28, No. 8l, pp. 1808-1818.

Vassalli, A., P. Feinstein, and P. Mombaerts, 2011, "Homeodomain Binding Motifs Modulate the Probability of Odorant Receptor Gene Choice in Transgenic Mice", *Molecular and Cellular Neuroscience*, Vol. 46, No. 2l, pp. 381-396.

Vassar, R., S. K. Chao, R. Sitcheran, J. M. Nunez, L. B. Vosshall, and R. Axel, 1994, "Topographic Organization of Sensory Projections to the Olfactory Bulb", *Cell*, Vol. 79, No. 6l, pp. 981-991.

Volz, A., A. Ehlers, R. Younger, S. Forbes, J. Trowsdale, D. Schnorr, S. Beck, and A. Ziegler, 2003, "Complex Transcription and Splicing of Odorant Receptor Genes", *Journal of Biological Chemistry*, Vol. 278, No. 22l, pp. 19691-19701.

Wang, M. M., and R. R. Reed, 1993, "Molecular Cloning of the Olfactory Neuronal Transcription Factor Olf-1 by Genetic Selection in Yeast", *Nature*, Vol. 364, No. 6433l, pp. 121-126.

Weth, F., W. Nadler, and S. Korsching, 1996, "Nested Expression Domains for Odorant Receptors in Zebrafish Olfactory Epithelium", *Proceedings of the National Academy of Sciences U S A*, Vol. 93, No. 23l, pp. 13321-13326.

Young, J. M., B. M. Shykind, R. P. Lane, L. Tonnes-Priddy, J. A. Ross, M. Walker, E. M. Williams, and B. J. Trask, 2003, "Odorant Receptor Expressed Sequence Tags Demonstrate Olfactory Expression of over 400 Genes, Extensive Alternate Splicing and Unequal Expression Levels", *Genome Biology*, Vol. 4, No. 11l, pp. R71.

Youngentob, S. L., J. E. Schwob, P. R. Sheehe, and L. M. Youngentob, 1997, "Odorant Threshold Following Methyl Bromide-Induced Lesions of the Olfactory Epithelium", *Physiology and Behavior*, Vol. 62, No. 6l, pp. 1241-1252.

Zhang, X., O. De la Cruz, J. M. Pinto, D. Nicolae, S. Firestein, and Y. Gilad, 2007, "Characterizing the Expression of the Human Olfactory Receptor Gene Family Using a Novel DNA Microarray", *Genome Biology*, Vol. 8, No. 5l, pp. R86.

Zhang, X., and S. Firestein, 2002, "The Olfactory Receptor Gene Superfamily of the Mouse", *Nature Neuroscience*, Vol. 5, No. 2l, pp. 124-133.