Evolution of Vertebrate Endocrine and Neuronal Gene Families

Focus on Pituitary and Retina

DANIEL OCAMPO DAZA
The duplication of genes followed by selection is perhaps the most prominent way in which molecular biological systems gain multiplicity, diversity and functional complexity in evolution. Whole genome duplications (WGDs) therefore have the potential of generating an extraordinary amount of evolutionary innovation. It is now accepted that the vertebrate lineage has gone through two rounds of WGD in its early stages, after the divergence of invertebrate chordates and before the emergence of jawed vertebrates. These basal vertebrate WGDs are called 2R for two rounds of whole genome duplication. An additional WGD called 3R occurred early in the evolution of teleost fishes, before the radiation of this species-rich group. This thesis describes the evolution of several endocrine and neuronal gene families in relation to the vertebrate WGDs, through a comparative genomic approach including both phylogenetic analyses and chromosomal location data across a wide range of vertebrate taxa.

These results show that numerous endocrine gene families have expanded in 2R and in several cases also in 3R. These include the gene families of oxytocin and vasopressin receptors (OT/VP-R), somatostatin receptors (SSTR) and insulin-like growth factor binding proteins (IGFBP). For the OT/VP-R and SSTR families, previously undescribed subtypes were identified. The protein hormone family that includes growth hormone (GH), prolactin (PRL) and somatolactin (SL) acquired a new PRL gene in 2R, however the origins of GH, PRL and SL likely predate 2R. The corresponding family of receptors diversified during different time periods through a combination of local duplications and 3R.

Neuronal gene families of the visual system have also expanded in 2R and 3R. The results presented here demonstrate that the vertebrate repertoire of visual opsins genes arose in 2R as part of chromosomal blocks that also include the OT/VP-R genes. The gene families including the transducin alpha, beta and gamma subunits also arose in 2R, hinting at the importance of these events in the diversification and specialization of phototransduction cascades for rods and cones.

Thus, the whole genome duplications have been important contributors to the evolution of both vision and endocrine regulation in the vertebrates.

**Keywords:** phylogenetics, evolution, molecular evolution, gene family evolution, genome duplication, gene duplication, oxytocin receptor, vasopressin receptor, visual opsin, transducin, growth hormone, prolactin, somatolactin, growth hormone receptor, prolactin receptor, somatostatin receptor, SSTR, IGFBP
Here, in my place and time,
and here in my own skin,
I can finally begin.

Arcade Fire, "Deep Blue"
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


* These authors have contributed equally.

Reprint of Papers I and III with permission from Elsevier. Reprint of Paper VII with permission from The Endocrine Society. Paper VI is published under an Open Access license.
In addition to the papers included in the thesis, the author has published the following papers.


Contents

Introduction .................................................................................................................. 9

Methods ....................................................................................................................... 12
  Vertebrate genome databases ................................................................................. 12
  Sequence identification and curation ..................................................................... 13
  Sequence alignments and phylogenetic analyses .................................................... 14
  Analyses of conserved synteny .............................................................................. 16

Evolution of pituitary hormone systems ................................................................. 17
  Oxytocin and vasopressin receptors (Papers I and II) ........................................... 19
  Growth hormone, prolactin, somatolactin and their corresponding receptors (Papers IV and V) ................................................................. 22
  Somatostatin receptors (Paper VI) ....................................................................... 24
  Insulin-like growth factor binding proteins (Paper VII) .................................... 25

Evolution of vertebrate visual opsins and transducin subunits .................................. 28
  Vertebrate visual opsins (Paper II) ..................................................................... 29
  Transducin subunits (Paper III) ......................................................................... 30

Extensive chromosomal rearrangements in the teleost lineage after 3R ................. 32

Conclusions, ongoing research and perspectives .................................................. 34

Popular science summary in Swedish (Sammanfattning på svenska) ............... 38

Acknowledgements ................................................................................................. 41

Open science declaration and supporting information ........................................ 42

Corrections ............................................................................................................... 43

References ................................................................................................................ 44
Abbreviations

2R Two rounds of whole genome duplication
3R Third round of whole genome duplication
BLAST Basic local alignment search tool
CACNA1-L Voltage-gated calcium channel, L-type
GH Growth hormone
GHR Growth hormone receptor
GNAI G-protein alpha inhibiting subunit
GNAT G-protein alpha transducing subunit
GNB G-protein beta subunit
GNGT G-protein gamma transducing subunit
GPCR G-protein-coupled receptors
IGFBP Insulin-like growth factor binding protein
LWS Ancestral red opsin
NJ Neighbor joining (method)
OT Oxytocin
OT/VP-R Oxytocin and vasopressin receptors
PhyML Phylogenetic maximum likelihood (method)
PRL Prolactin
PRLR Prolactin receptor
Rh1 Rhodopsin
Rh2 Ancestral green opsin
SL Somatolactin
SSTR Somatostatin receptor
SWS1 Ancestral ultraviolet opsin
SWS2 Ancestral blue opsin
UTS2R Urotensin II receptor
VP Vasopressin
Introduction

“… over the course of the past decade, as genome sequences began to fill the literature, even the most molecular and computational of biologists have become like naturalists. They wander through diverse landscapes of As, Ts, Gs, and Cs, comparing genomes and wonder about the origin of the distinct classes of variation found there.” (Caporale 2009)

As vertebrates, we are united by a common ancestry spanning back some 500 million years to the Cambrian period (Holland and Chen 2001; Donoghue, Smith, and Sansom 2003; Donoghue and Purnell 2005). Since then we have evolved a great variety of different complex and specialized functions, and occupied diverse ecological niches. The availability of sequenced, mapped and annotated genomes from a broad selection of vertebrate species has made it possible to address questions about our shared vertebrate evolution on a genome-wide scale. Our common ancestry allows us to infer the homology of the contents of our genomes through comparative approaches, and the recent surge of genome sequencing projects allow us to do so with growing accuracy and resolution. Together with the advances in DNA sequencing technologies and bioinformatics methods, and together with the growing impetus to share the large quantity of obtained data openly, this development has made it ever more important for biologists of all fields to consult sequence databases and analyze genomic data in order to answer their research questions. For vertebrate evolution, the questions that can be addressed range from large-scale analyses of genome evolution to targeted comparative evolutionary studies of individual gene families.

There is now convincing evidence from large-scale genomic studies that the vertebrate lineage went through two rounds of whole genome duplication, called 1R and 2R, early in its evolution (Dehal and Boore 2005; Nakatani et al. 2007; Putnam et al. 2008). Subsequently a third round of whole genome duplication, called 3R, occurred early in teleost fish evolution (Jaillon et al. 2004; Meyer and Van de Peer 2005; Kasahara et al. 2007). The timing of the 2R events has been estimated to 500-430 million years ago, with one round occurring before the divergence of cyclostomes (lampreys and hagfishes) and one round after (Panopoulou and Poustka 2005). This was based on the presence of fewer gene orthologs in cyclostomes compared with jawed vertebrates. However, recent studies have challenged this and argued that the
2R events occurred before the divergence of cyclostomes and jawed vertebrates (Kuraku, Meyer, and Kuratani 2009; Kuraku 2013).

Duplication, be it of genes, segments of chromosomes or entire genomes, is a force to be reckoned with in evolution because it creates new genetic material that mutation and selection can act on to generate novel gene function and thus originate evolutionary novelties. Whole genome duplications in particular have the potential of generating significant innovation (Van de Peer, Maere, and Meyer 2009). This way of reasoning can be traced all the way back to Darwin who hinted at it by recognizing the contribution of repeated phenotypic elements to increased organismal complexity: “... it is quite probable that natural selection, during a long-continued course of modification, should have seized on a certain number of the primordially similar elements, many times repeated, and have adapted them to the most diverse purposes.” (Darwin 1859). The Japanese-American geneticist Susumu Ohno’s seminal work Evolution by Gene Duplication (Ohno 1970) is often cited as the first comprehensive argument for the significant impact of gene duplications in evolution, and as the first proposition that the vertebrate lineage had undergone genome duplication events.

Genes that derive from a common ancestral gene are said to form a gene family. Related genes within the same genome are said to be paralogous and the corresponding gene in different species are called orthologous. The study of gene family evolution mostly comprises the elucidation of the duplication events that gave rise to the gene family members. A prerequisite of this process is to study the molecular phylogenies of the gene families combined with chromosomal positional data for the gene family members. By studying gene family evolution in detail it is also possible to investigate how complex molecular systems have evolved through changes in their individual components. The comparative phylogenetic approach makes it possible to predict the functions of newly discovered gene family members, or of gene family members in species where they have not yet been studied, because related genes often have protein products that perform related functions.

The studies presented in this thesis investigate the evolution of several endocrine and neuronal gene families within the context of whole genome duplications in the evolution of vertebrates, through a comparative genomic approach including a broad variety of vertebrate taxa. This work entails the bioinformatic identification and annotation of gene sequences (in many cases the de novo identification of genes), molecular sequence analyses, sequence-based phylogenetic analyses as well as analyses of conserved synteny and similarities between duplicated chromosomes within a genome. The main findings of these studies are that several gene families involved in neurobiological and endocrine processes expanded in the vertebrate whole genome duplications, giving rise to much of the present diversity in these systems. Thus, the process of whole genome duplication has likely been a significant component in the evolution of vertebrate nervous systems and vertebrate
endocrine regulation. In several cases, previously unrecognized or misidentified components of these systems have been identified. Additionally, these studies have made it possible to trace the evolution of the chromosomal regions harboring the gene families: There is evidence that there have been extensive chromosomal rearrangements within paralogous genome regions that arose in the vertebrate whole genome duplications, particularly within the teleost fish lineage.

There are two themes explored in the works included in this thesis. The first theme is the evolution of chromosomal regions harboring gene families involved in the visual signal transduction pathways of the retina, including the visual opsins and also the adjacent gene family of the receptors for the neurohypophyseal hormones oxytocin and vasopressin. The second theme is the evolution of several endocrine gene families involved in the hypothalamic-pituitary axis of growth regulation and metabolism.
Methods

Vertebrate genome databases

The Ensembl genome browser (www.ensembl.org) currently provides genome sequence databases for numerous animal species, including a series of integrated genome resources such as annotated gene predictions, orthology and paralogy predictions, comparative tools and data mining functions (Flicek, Ahmed, et al. 2012). For some species regulation and sequence variation data are also integrated. The available species span most vertebrate classes as well as two tunicate species and three invertebrate species. As of the latest version update (January 2013) there are over forty mammalian genomes, six avian genomes (three in preliminary release status), three reptilian genomes (one in preliminary release), one amphibian genome, eight teleost fish genomes and one cyclostome genome, that of the sea lamprey (Petromyzon marinus). The coverage of the genome sequences and the quality of their assemblies varies considerably between different species; about half of the genome sequences are low-coverage (Milinkovitch et al. 2010), which can make the automatic orthology and paralogy predictions unreliable.

Recently genome assemblies of the coelacanth (Latimeria chalumnae) and the spotted gar (Lepisosteus oculatus) have been released in Ensembl, providing important out-groups for tetrapods and teleost fishes, respectively. The spotted gar genome was particularly useful to sort out 3R-generated duplications and chromosome rearrangements in the teleost genomes due to its divergence before the 3R event in the teleost lineage (Figure 1).

The following basic setup of vertebrate genomes was used in this thesis: human (Homo sapiens), mouse (Mus musculus), grey short-tailed opossum (Monodelphis domestica), anole lizard (Anolis carolinensis), chicken (Gallus gallus), Western clawed frog (Xenopus tropicalis), zebrafish (Danio rerio), three-spined stickleback (Gasterosteus aculeatus), medaka (Oryzias latipes), green spotted pufferfish (Tetraodon nigroviridis) and fugu (Takifugu rubripes). This setup represents most vertebrate classes, including both lobe-finned fishes and ray-finned fishes. Additional species were used for some of the analyses in order to solve specific gaps in the data. See the respective papers for details. Notably, we have used sequences and positional data from the spotted gar and coelacanth genomes in Papers II, IV, V and VI and se-
quences from the draft genome of the elephant shark (*Callorhinichus milii*) (Venkatesh et al. 2007) were used in Papers I and II to resolve the phylogeny of oxytocin and vasopressin receptors. The elephant shark genome was accessed via *esharkgenome.imcb.a-star.edu.sg*. However, due to the shortness of the assembled genomic scaffolds, it was not practical to use this genome for all the analyses.

In addition to the Ensembl database, the database system maintained by the National Centre for Biotechnology Information (NCBI) (ncbi.nlm.nih.gov) was used to identify sequences when Ensembl’s automatic gene predictions were incomplete or when there seemed to be gaps in the Ensembl genome assemblies. The NCBI resources include a variety of genomic sequences, nucleotide sequences, protein sequences and expressed sequence tags (ESTs) (Sayers et al. 2010). For the first studies of the spotted gar genome described in Paper II and Paper VI the genome assembly was accessed via NCBI.

![Figure 1. Phylogenetic relationship between the spotted gar and the teleost fish species with sequenced genomes. Divergence times (million years ago) are estimates from timetree.org (Hedges, Dudley, and Kumar 2006), except those shown in italics which are minimum and maximum estimates by (Donoghue and Benton 2007).](image)

**Sequence identification and curation**

To identify the complete gene repertoire for the analyzed gene families, the automatic orthology predictions (Vilella et al. 2009) and protein family pre-
dictions (Enright, Van Dongen, and Ouzounis 2002) in the Ensembl genome browser were used (Flicek, Ahmed, et al. 2012). Amino acid sequences corresponding to the identified gene annotations were collected for sequence-based phylogenetic analyses (see below). To identify additional gene predictions that were not included in the automatic protein family predictions, Basic Local Alignment Searches (BLAST) (Altschul et al. 1990) were performed using the identified amino acid sequences (blastn) as search terms with standard settings in both the Ensembl and NCBI databases. When using this method, special consideration needs to be taken for each gene family with regard to the interpretation of the BLAST hits. If the family has a low degree of sequence conservation, or if it shares conserved domains or motifs with other gene families, the chances of mining false gene family members increases and careful selection criteria have to be applied. All identified gene predictions were tested for inclusion in the final phylogenetic analyses by either reciprocal BLAST searches in the NCBI databases or preliminary neighbor joining phylogenetic trees (see below).

For short, incomplete, or divergent gene annotations in the Ensembl genome databases, the corresponding genomic sequences (including full intron sequences and flanking regions) were retrieved. These sequences were inspected manually to identify erroneous automatic exon predictions, or exons that had not been predicted, by following consensus splice donor and acceptor sites as well as sequence identities to other family members. The Genscan gene prediction server (genes.mit.edu/GENSCAN.html) (Burge and Karlin 1997) was used to complement the manual curation. The prevalence of faulty and incomplete sequence predictions in the Ensembl databases has been reviewed by (Prosdocimi et al. 2012).

Additionally, some gene families were investigated for characteristic conserved domains and/or signature motifs using several of the available protein family and protein domain databases, mainly Pfam (pfam.sanger.ac.uk) (Finn et al. 2010) and InterPro (ebi.ac.uk/interpro) (Quevillon et al. 2005). The TMHMM server (cbs.dtu.dk/services/TMHMM) (Krogh et al. 2001) was used in Paper I to predict the transmembrane helices of the previously unrecognized oxytocin and vasopressin receptors. This approach allowed us not only to positively identify family members in the different gene families, but also to predict and curate previously unknown family members and correct the faulty sequence predictions in the Ensembl genome databases.

Sequence alignments and phylogenetic analyses

Amino acid sequences were aligned using the ClustalW algorithm (Larkin et al. 2007) with the following settings: Gonnet weight matrix, gap opening penalty 10.0 and gap extension penalty 0.20. The alignment algorithm was applied through different bioinformatic programs (see the respective papers
for details). All resulting alignments were inspected manually and curated carefully in order to ratify faulty or divergent predictions and correct mali
lied sequence stretches. Specific details on how the alignments have been curat
ed are given in the respective papers. As a general rule, we have adjust
ed conserved exon boundaries and all apparent individual misalignments 
within conserved sequence stretches or known protein domains. In some 
cases we have also removed the most variable regions from the alignments 
in order to reduce phylogenetic artifacts. This step was necessary in some 
cases since these regions often represent unalignable sequence stretches of 
varying length and contain many non-homologous positions that introduce 
errors into the phylogenetic analyses. The manual curation of the amino acid 
sequence predictions from the genomic databases and of the sequence 
alignments is a necessary step in order to focus on the relevant evolutionary 
and biological information in the dataset and avoid erroneous sequence pre
dictions.

The manually curated alignments were used to calculate phylogenetic 
trees using both the distance-based neighbor joining (NJ) method (Saitou 
and Nei 1987) and the Phylogenetic Maximum Likelihood (PhyML) method 
(Guindon and Gascuel 2003; Guindon et al. 2010). The topologies of both 
NJ trees and PhyML trees were in most cases supported by non-parametric 
bootstrap tests with 1000 and 100 replicates respectively. In Paper II the 
PhyML trees for neighboring gene families were supported by a non
parametric SH-like approximate likelihood ratio test (aLRT). This method is 
several times faster than both Bayesian and bootstrap tests of tree topology, 
and has been shown to produce comparable results with different kinds of 
sequence data (Anisimova and Gascuel 2006; Guindon et al. 2010; Anisi
mova et al. 2011). In Paper VI we have calculated concurrent PhyML 
trees for the somatostatin receptor family using both bootstrapping and 
aLRT.

For all PhyML analyses the settings were changed from the default in or
der to take advantage of the careful curation of each alignment. The amino 
acid frequencies (equilibrium frequencies), proportion of invariable sites and 
gamma shape parameters were estimated from the alignments. The exact 
settings for the different PhyML analyses have been detailed in the respec
tive papers. The best amino acid substitution models for the PhyML analyses 
were selected for each alignment using ProtTest (Abascal, Zardoya, and 
Posada 2005)

Careful consideration has also been given to the selection of out-group 
sequences for the rooting of the phylogenetic trees. In most cases the ge
nomes of the tunicates *Ciona intestinalis* and *Ciona savignyi* (accessed via 
Ensembl) have been used to provide the relative dating for the time-window 
of the 2R whole genome duplications, and fruit fly (*Drosophila melanogaster*) sequences have been used to root the trees. In some cases the ge
nome assembly of the Florida lancelet (*Branchiostoma floridae*) (accessed
via *genome.jgi-psf.org*) has been searched when no tunicate sequences could be found. For some gene families other invertebrate sequences were sought by Hidden Markov Model searches using the HMMER web server (*hmmer.janelia.org*) (Finn, Clements, and Eddy 2011) and its pHMMER implementation against the UniProtKB database and the NCBI non-redundant (NR) protein database.

**Analyses of conserved synteny**

Different strategies were used for the detection of conserved syntenies in the different papers. As a general method, gene lists corresponding to the putatively paralogous chromosome regions were downloaded for several species using Ensembl’s data-mining tool BioMart (Flicek, Amode, et al. 2012). These lists were then compared within the same species and between species with regard to the composition of Ensembl protein families (Enright, Van Dongen, and Ouzounis 2002). The different strategies have been detailed in the respective papers.
Evolution of pituitary hormone systems

Pituitary hormone axes regulate growth, homeostasis, metabolism, immune function, reproduction, to name a few broad categories, and integrate these physiological processes into adaptive responses to a changing environment, such as metabolic adaptations, stress responses and behavioral changes. The expansion of pituitary hormone systems constitutes one of the great innovations in vertebrate evolution and underlies the physiological adaptations that have allowed vertebrates to thrive and diversify in a large number of varied habitats. The pituitary gland consists of the anterior pituitary, derived from oral ectoderm tissue, and the posterior pituitary, which consists largely of neuronal projections from the hypothalamus. It is present in all vertebrates, but not in invertebrates; including the extant stem groups of vertebrates, the tunicates and the cephalochordates (lancelets). However, many individual components of the pituitary molecular systems originated before the emergence of vertebrates (Campbell, Satoh, and Degnan 2004).

Anterior pituitary

The adult mammal anterior pituitary consists of five different endocrine cell types. These cell-types are the somatotrophs, which secrete growth hormone (GH), lactotrophs, which secrete prolactin (PRL), gonadotrophs, which secrete the gonadotropins luteinizing hormone (LH) and follicle-stimulating hormone (FSH), corticotrophs, which secrete adrenocorticotropic hormone (ACTH), and thyrotrophs, which secrete thyroid stimulating hormone (TSH) (Le Tissier et al. 2012). The teleost anterior pituitary can more clearly be divided into smaller structures, an anterior lobe called pars distalis, containing the same endocrine cell types described above, as well as a posterior lobe called pars intermedia that contains corticotrophs, melanotrophs, which release alpha-melanocyte stimulating hormone (α-MSH) (Pogoda and Hammerschmidt 2009) as well as endocrine cells that release somatolactin. Somatolactin (SL) is a protein hormone related to GH and PRL that has only been identified in teleost fishes, see (Lopez et al. 2006), (Benedet et al. 2008) and references therein, as well as in sturgeon and lungfish (Amemiya et al. 1999). The anterior pituitary hormones belong to larger hormone families, based on amino acid sequence homology as well as structural and functional similarities. LH, FSH and TSH belong to the glycoprotein hormone family, GH, PRL and SL to the somatotropin family and ACTH to the
proopiomelanocortin (POMC)-derived peptides of the opioid peptide family. ACTH is cleaved from the pro-peptide POMC, same as the melanocyte-stimulating hormones (MSH) and beta-endorphins (β-END). The pituitary in cyclostomes (lamprey and hagfish) consists of the same two basic components as in jawed vertebrates and the presence of anterior pituitary hormones has been confirmed, primarily through physiological and histochemical studies. Lampreys have putative ACTH and MSH like hormones as well as putative GH and gonadotropin (GTH) (Kawauchi and Sower 2006), likewise hagfishes have ACTH, GTH and GH like hormones (Nozaki 2008). The vertebrate ancestor likely had components of all three anterior pituitary hormone families, but notably fewer of them. This suggests that these hormone families expanded during the early evolution of vertebrates. The family of opioid peptide precursors, which includes POMC that encodes the ACTH, MSH peptides in addition to the opioid peptide β-endorphin, expanded in 2R concomitantly with the opioid receptor family, giving rise to the vertebrate opioid system (Dreborg et al. 2008; Sundström, Dreborg, and Larhammar 2010), reviewed in (Dores and Baron 2011). The story is less clear regarding the pituitary glycoprotein hormones. These proteins are heterodimers, sharing a common alpha-subunit bound to one out of three respective beta-subunits specific for TSH, FSH and LH in jawed vertebrates. Cyclostomes seem to have only a single putative GTH beta-subunit; presumably GTH and TSH beta-subunits arose through gene duplication before the divergence of cyclostomes and subsequently the GTH beta-subunit gene generated FSH and LH through gene duplication in a gnathostome ancestor (Quérat et al. 2001; Quérat et al. 2004). However it has not been determined whether this duplication occurred in 2R.

The release of anterior pituitary hormones is under the control of the hypothalamus through several releasing factors, such as corticotrophin-releasing factor (CRF) and gonadotrophin-releasing hormone (GnRH), and includes many feedback mechanisms that fine-tune the endocrine communication between the hypothalamus, the pituitary and the peripheral tissues of the body. Recent evidence shows that several families of peptides that regulate pituitary functions also have expanded in both 2R and 3R. This includes the somatostatin family (Tostivint, Lihrmann, and Vaudry 2008; Liu et al. 2010; Tostivint et al. 2013) and the secretin superfamily, which includes growth hormone-releasing hormone (GHRH) among others (Lee et al. 2007; Roch, Wu, and Sherwood 2009; Cardoso et al. 2010). Somatostatin is the main inhibitor of GH secretion from somatotrophs in the anterior pituitary, via the somatostatin receptors (SSTRs). GHRHs are the main stimulatory factors of GH release. The interaction between the opposite actions of somatostatin and GHRH produce the pulsatile release of GH that is required for normal growth (Gillies 1997). Both of these peptides, together with GH itself and insulin-like growth factor 1 (IGF-1), are involved in the feedback mechanisms that integrate different levels of the regulation of somatic growth,
metabolism and fuel homeostasis. Additionally, the insulin-like growth factor binding proteins (IGFBPs) modulate the actions of IGF-1 and add another level of complexity to the regulation of the GH/IGF-1 axis. In summary, several multi-member vertebrate gene families involved with anterior pituitary functions have been shown to have expanded in the 2R and 3R events. This suggests that whole genome duplications have contributed to innovations on different levels of the vertebrate anterior pituitary endocrine axes as well as their feedback mechanisms. Concomitant expansions of intracellular signaling pathways might also have contributed to vertebrate endocrine innovations. A recent study shows that components of the Janus kinase (Jak) – Signal transducer and activator of transcription (STAT) pathways of signal transduction, which are used by the GH and PRL receptors among others, expanded in 2R (Liongue et al. 2012).

Posterior pituitary

The neurohypophyseal (posterior pituitary) hormone family is ancient, with identified members from many invertebrate species, both protostomes and deuterostomes (Beets et al. 2012; Garrison et al. 2012). However, only jawed vertebrates, excluding cyclostomes, have at least two neurohypophyseal hormones, one oxytocin (OT)-like and one vasopressin (VP)-like, whose genes are located adjacent to each other (Gwee et al. 2008; Larhammar et al. 2009). This setup evolved in an ancestor of all jawed vertebrates after a tandem gene duplication generated a neurohypophyseal gene locus with two genes. Lampreys have been found to have only one neurohypophyseal hormone gene, as has the cephalochordate lancelet (Gwee et al., 2009). This shows that while neurohypophyseal hormones seem to be a very ancient family of peptides, there was significant innovation in the neurohypophyseal system in early vertebrate evolution.

Oxytocin and vasopressin receptors (Papers I and II)

The neurohypophyseal hormones oxytocin (OT) and vasopressin (VP) were among the first peptides to be sequenced, and homologs have been described in all gnathostome classes (Larhammar et al. 2009), albeit with different names such as vasotocin for orthologs of vasopressin, and isotocin and mesotocin for oxytocin orthologs (Acher 1980; Hoyle 1999). Oxytocin and vasopressin genes arose from a common ancestral gene through a local duplication in a gnathostome ancestor (Gwee et al. 2009).

The vertebrate OT and VP receptors form a family of G-protein-coupled receptors (GPCRs) that mediate a large variety of functions, including social behavior and the regulation of blood pressure, water balance and reproduction. For instance, see reviews by (Gimpl and Fahrenholz 2001), (Balment et
al. 2006), (Bourque 2008), (Donaldson and Young 2008) and (Insel 2010). In mammals four family members have been identified, three of which respond to vasopressin, named V₁A, V₁B and V₂, and one of which is activated by oxytocin, called the OT receptor (see references in Table 1, Paper I). Until recently only V₁-type receptors (Mahlmann et al. 1994; Conklin, Smith, and Olson 1999; Warne 2001; An, Kim, and Choi 2008) and OT receptors (Hausmann et al. 1995) had been identified in teleost fishes, although the presence of a V₂-type receptor had been suggested by the finding that vasotocin (VT) could activate the VT-type adenyl cyclase/cAMP signal transduction pathway in fish (Guibbolini, Pierson, and Lahlou 2000). With the identification of two V₁A-type and a V₂-type receptor in the Amargosa pupfish (*Cyprinodon nevadensis amargosae*) (Lema 2010), and the identification of V₂-type receptors in gray bichir (*Polypterus senegalus*) and medaka (*Oryzias latipes*) (Konno et al. 2010) the known repertoire of OT/VP receptors in ray-finned fishes was expanded.

The investigation of OT/VP receptor sequences in several gnathostome classes presented in Paper I allowed us to deduce a phylogenetic tree that provided some surprising findings. We concluded that there were at least five receptor subtypes present already in the gnathostome ancestor; V₁A, V₁B, V₂A, V₂B and OT receptors, thus there is a larger receptor repertoire in the vertebrate lineage than was previously known (Ocampo Daza, Lewicka, and Larhammar 2012). The existence of distinct V₂A and V₂B receptors had not been recognized prior to this publication. The V₂A-type receptors are orthologous to the human V₂ receptor, while the receptor subtype we called V₂B appeared to be orthologous to the chicken V₂ receptor (Baeyens and Cornett 2006), also called VT1 for vasotocin receptor 1. We were able to identify both V₂A- and V₂B-type receptor sequences in all examined teleost genomes as well as in the genomes of the Western clawed frog and anole lizard. Thus reciprocal losses seem to have occurred in some lineages, namely V₂B-type receptors from mammals and V₂A-type receptors from birds. We also identified an additional V₂-like receptor type in the zebrafish and stickleback, but since no putative orthologs were found in any other genomes, the orthology relationship of this receptor subtype was inconclusive at the time. Subsequent analyses of the OT/VP receptor family by (Yamaguchi et al. 2012) helped clarify the orthology relationships within the family. Through their phylogenetic, genomic and functional analyses of the gene family it became clear that there were in fact three different V₂ receptor types: V₂A, V₂B1 and V₂B2, with both types of V₂B receptors signaling via Ca²⁺ rather than through cAMP like V₂A-type receptors. The V₂B2-type receptor clade includes the chicken V₂ receptor as well as the “V₂-like” receptor sequences we had identified in zebrafish and stickleback.
Figure 2. Updated view of the evolution of the oxytocin and vasopressin receptor family in gnathostomes. The upper panel shows a cladogram of the phylogenetic maximum likelihood analysis presented in Paper II. The side panel shows a timetree of chordate evolution with divergence estimates from timetree.org (Hedges, Dudley, and Kumar 2006) and (Near et al. 2012). Arrowheads: (2R) Expansion of the OT/VP-R family in 2R giving rise to the six subtypes seen in the upper panel. (1) Loss of V1B-type receptors from the ray-finned fish lineage, before the divergence of the spotted gar and teleosts. (2) Duplications of V2A-type receptors in 3R, local duplications of V1A-type and OTR receptors. (3) Loss of V1B-type receptors from the tetrapod lineage after the divergence of amphibians. (4) Independent losses of V2A-type receptors from the lineages leading to elephant shark and chicken. (5) Independent losses of V2B2-type receptors from several lineages, including both ray-finned fishes and lobe-finned fishes.

Our studies of conserved synteny between the OT/VP receptor gene-bearing chromosome regions were underway when the article by Yamaguchi et al. (2012) was published. As part of Paper II we present the results of these conserved synteny analyses as well as an updated phylogeny of the OT/VP receptors, including identified sequences from the genomes of the coelacanth (Latimeria chalumnae), spotted gar (Lepisosteus oculatus) and Southern platyfish (Xiphophorus maculatus). Our current evolutionary scheme for the evolution of OT/VP receptors is that two ancestral genes were present on the same ancestral vertebrate chromosome before 2R, one giving rise to V1A-type, V1B-type and OT receptors through 2R, and one giving rise to V2A-type, V2B1-type and V2B2-type receptors. In the teleost lineage, the 3R event gave rise to teleost-specific duplicates of V1A-type receptors, while the duplicate OT and V2A-type receptor genes arose through local duplications. See Figure 5 in Paper II for a summary of the duplication events. A V1B-type receptor sequence could not be found in the spotted gar, which indicat-
ed that these receptors were lost early in the ray-finned fish lineage. We can also conclude that there have been several independent losses of \( V_2 \)-type receptor genes, especially of \( V_2B2 \). Figure 2 summarizes our current view of the OT/VP receptor evolution in vertebrates.

Our comprehensive phylogeny of the OT/VP receptors, coupled with our analysis of the chromosome regions, helps resolve orthology relationships and corrects previous cases of receptor subtype misidentification. These evolutionary clarifications also provide a coherent framework for comparative studies of the many functions of neurohypophysyal hormones and their receptors.

Growth hormone, prolactin, somatolactin and their corresponding receptors (Papers IV and V)

Growth hormone (GH), prolactin (PRL), and somatolactin (SL) belong to the same gene family within the large family of class-1 helical cytokines (Huising, Kruiswijk, and Flik 2006). While GH and PRL, including some lineage-specific homologs, have been identified in a variety of vertebrate species (Forsyth and Wallis 2002), SL has only been identified in teleost fishes (Ono et al. 1990; Rand-Weaver et al. 1991; Benedet et al. 2008) as well as in a sturgeon (Acipenser transmontanus) and a lungfish species (Protopterus annectens) (Amemiya et al. 1999). Duplicate SL genes called SL\( \alpha \) and SL\( \beta \) have been characterized in some teleost fishes, including salmonid fishes (Benedet et al. 2008) and cyprinid fishes (Zhu et al. 2004; Jiang et al. 2008; Azuma et al. 2012). Growth hormone and prolactin have a variety of effects in development, osmoregulation, metabolism and stimulation of growth through a family of interrelated receptors (Liongue and Ward 2007). Prolactin is particularly pleiotropic, with over 300 different reported functions, including reproductive functions in mammals (Freeman et al. 2000). In teleost fishes GH and PRL mainly have osmoregulatory and reproductive functions (Sudo et al. 2013). The functions of SL are only partially known and include skin pigmentation (Fukamachi et al. 2009), electrolyte balance (Uchida et al. 2009) and lipid metabolism (Sasano, Yoshimura, and Fukamachi 2012). See also (Jiang and Wong 2013) and references therein.

Both the hormone and the receptor family have been proposed to have expanded early in vertebrate evolution through gene duplications, although not concomitantly. The protein hormone family seems to have expanded early in vertebrate evolution, but different proposals for how this happened have been launched (Bole-Feysot et al. 1998; Forsyth and Wallis 2002; Kawauchi and Sower 2006). The growth hormone receptor (GHR) duplicated in the 3R event early in teleost evolution (Fukamachi and Meyer 2007) and although the evolution of prolactin receptor (PRLR) genes has not been
examined in detail, duplicate receptors have been reported in teleost fish species (Huang et al. 2007). The different suggested time-windows for the protein hormone and receptor families raises questions about possible hormone-receptor pairings and how they emerged. It is therefore of interest to identify the relative time points for the expansion of both families as well as the possible correlation with the 2R and 3R events.

In Paper IV, GH, PRL and SL gene family members were identified in a broad range of vertebrate genomes and investigated using a combination of sequence-based phylogenetic analyses and comparative genomic analyses of synteny. The chromosomal regions of the GH, PRL and SL genes show many lineage-specific rearrangements with only a small region of conserved synteny between tetrapod and teleost genomes. No family member could be identified in any invertebrate deuterostome lineage, which leaves the origin of the gene family in obscurity. Nonetheless, the most parsimonious interpretation of our results is that the origin of GH and PRL lies in a local gene duplication before 2R. GH and PRL genes are located on the same chromosomes in several of the investigated teleost genomes, and the observed pattern of conserved synteny suggests that this is the ancestral organization of these genes. In tetrapod genomes this linkage is broken by the translocation of the PRL gene. The emergence of the SL gene could also predate 2R, since very little conserved synteny could be found between the chromosomal regions bearing SL on the one hand, and GH and PRL on the other hand. Subsequently the 2R events generated two PRL subtypes: The second, a newly found PRL homolog called PRL2, is present in all teleost, avian and reptilian genomes investigated, as well as in the genomes of the coelacanth (Latimeria chalumnae) and the spotted gar (Lepisosteus oculatus). We first reported this PRL duplicate as PRLb, based on its identification in teleost genomes, and suggested it had arisen through the duplication of a large chromosome block (Ocampo Daza et al. 2009). Subsequently it has been identified in both ray-finned and lobe-finned fishes, and been characterized biologically with regard to tissue expression and receptor activation in independent studies. In contrast to PRL, PRL2 does not seem to be expressed in the pituitary, but expression has been detected in the retina and in extrapituitary brain regions (Huang et al. 2009; Wang et al. 2010).

Phylogenetic analyses of SL sequences and chromosomal analyses of SL-bearing chromosome regions in teleost genomes support the duplication of SL in 3R, giving rise to SLα and SLβ genes. However, out of the teleost species with sequenced genomes, only the zebrafish (Danio rerio) has both SLα and SLβ genes.

The corresponding analyses of the receptor family described in Paper V confirm that it is mainly 3R that contributed new family members, giving rise to two pairs of GH and PRL receptor genes located on paralogous chromosome regions. Thus the protein hormone diversity arose millions of years before the diversity of the corresponding receptors. The GHR and PRLR
genes did not arise as a result of the 2R rounds of whole genome duplication, since they likely were located on the same chromosome in the jawed vertebrate ancestor. The time point for the duplication within the same chromosome that gave rise to GHR and PRLR is difficult to deduce since no homologous invertebrate sequence could be found. Until there is a complete cyclostome or cartilaginous fish genome assembly available, the earliest divergence time between GHR and PRLR genes that can be deduced is before the divergence of lobe-finned fishes and ray-finned fishes.

Somatostatin receptors (Paper VI)

Somatostatin and its related neuroendocrine peptides have a wide variety of physiological functions that are mediated by five somatostatin receptors subtypes (SSTR1-5). SSTs are the main inhibitors of GH secretion from somatotrophs in the anterior pituitary (Gillies 1997). In fact it was this action that first led to the isolation and identification of the 14-amino acid somatostatin peptide from the hypothalamus of sheep (Brazeau et al. 1973). Somatostatin is widely distributed and serves as a neuroendocrine peptide regulating the anterior pituitary, a neuropeptide acting on neurons and as an endocrine peptide. Accordingly, somatostatin has been reported to have many physiological effects (Viollet et al. 2008). The different somatostatin receptor subtypes have diverse and partially overlapping expression patterns across multiple target organs, with all five subtypes being expressed in the brain. They also differ in their ligand-binding affinities to the different somatostatin analogues (Patel 1999; Möller et al. 2003; Gahtete et al. 2010). SSTR2 and SSTR5 are the most abundant subtypes expressed on somatotroph cells (Thoss et al. 1996) and seem to be the main subtypes mediating the inhibition of GH release (Geris et al. 2003; Ren et al. 2003; Sheridan and Hagemeister 2010). The vertebrate somatostatin family of peptides expanded in the 2R events (Tostivint et al. 2013) and it has been suggested that also the SSTRs originated in 2R (Gahtete et al. 2010).

In Paper VI we have investigated the SSTR gene family and 47 adjacent gene families by phylogeny and conserved synteny analyses in a broad range of vertebrate species in order to resolve the evolution of SSTRs. In addition to the five known subtypes we were able to identify a previously unrecognized member of the receptor family in the teleost genomes as well as in the genome of the coelacanth. We have called this subtype SSTR6. Our phylogenetic analyses of the SSTR gene family provide strong support for expansion and diversification in both the 2R and 3R events, giving rise to the six different SSTR subtype genes early in vertebrate evolution and subsequently expanding the SSTR2, -3 and -5 branch in the teleost lineage. The chromosomal locations of the SSTR genes as well as the early divergence of two ancestral SSTR branches in the phylogenetic analysis suggested that the
SSTR1, -4 and -6 genes derive from one ancestral SSTR gene and the SSTR2, -3 and -5 genes from a separate ancestral SSTR gene and that these two ancestral genes were located in distinct ancestral chromosome regions. Therefore two separate analyses of conserved synteny were done. These analyses support the following scenario: two ancestral SSTR-bearing chromosome regions were quadruplicated in 2R. One of these ancestral SSTR genes generated SSTR2, -3 and -5 and the other gave rise to SSTR1, -4 and -6. Taken together these results indicate that all six SSTR subtype genes were ancestral to both lobe-finned and ray-finned fishes, but that reciprocal losses have occurred. The SSTR6 subtype was lost in tetrapods and the SSTR4 subtype in teleosts. Subsequently SSTR2, -3 and -5 conserved duplicates from generated in the 3R event, generating a remarkable multiplicity. Although there have been losses of SSTR subtype genes, the paralogous genome regions could be identified in both tetrapod and teleost genomes. However, our study shows that extensive chromosomal rearrangements have taken place between related chromosome regions in teleosts, especially between SSTR2, -3 and -5-bearing chromosome regions. These rearrangements make it exceedingly difficult to resolve orthology relationships for gene families located within these chromosomal regions, but we show that this can be resolved by investigating several distantly related species using both phylogenetic analyses and chromosomal data. The implications of these chromosomal rearrangements are detailed below under the heading *Extensive chromosomal rearrangements in the teleost genomes after 3R.*

**Insulin-like growth factor binding proteins (Paper VII)**

The insulin-like growth factor binding proteins (IGFBP) have multiple functions in the regulation of the growth hormone (GH) and insulin-like growth factor (IGF) hormonal axis. IGFBPs modulate the GH/IGF actions on cellular proliferation, diversification, migration and survival, as well as the anabolic effects on metabolism. Six distinct IGFBP subtypes numbered IGFBP-1 through -6 have been characterized in various vertebrate species. They act as high-affinity carrier proteins to IGF-1 and IGF-2 and regulate the turnover, availability, transport and half-life of IGFs both in circulation and in local tissues (Duan and Xu 2005; Duan, Ren, and Gao 2010). Different IGFBP subtypes have been found to either potentiate or inhibit IGF actions, or both. The expression of different IGFBPs is tissue-specific in some cases, but most cells express several subtypes that can have compensatory or opposite actions (Firth and Baxter 2002; Duan, Ren, and Gao 2010). Some IGFBPs also have IGF-independent actions, determined by the presence of specific conserved motifs of amino acid residues. For example, human IGFBP-1 interacts with integrins through a specific amino acid motif and IGFBP-3 and -5 have functional nuclear localization motifs (Jogie-Brahim,
Teleost-specific duplicates of several IGFBP subtypes have been identified and studied, and there is evidence that the duplicates have evolved specific functions (Kamei et al. 2008; Zhou et al. 2008; Wang et al. 2009; Dai et al. 2010). It is clear that there are many and diverse biological outcomes of IGFBP actions, both in the adult organism and in development, and IGFBP evolution seems to be a good model for the neo- and subfunctionalization of duplicate genes.

The IGFBP gene family was proposed early on to have expanded by chromosome duplications. This was based on their location within the same chromosome regions as the homeobox (HOX) gene clusters in the human genome (Lundin 1993). The HOX clusters as well as multiple syntenic gene families, including IGFBP, were subsequently found to have expanded in 2R and 3R (Sundström, Larsson, and Larhammar 2008). However, the phylogeny of the IGFBP family has been difficult to resolve and several reports have favored independent duplications or more complicated chromosome duplication scenarios (Hughes, Da Silva, and Friedman 2001; Abbasi and Grzeschik 2007; Gordon and Marcinkiewicz 2008). In Paper VII we identified and annotated IGFBP family genes in a wide variety of vertebrate species and done phylogenetic analyses of the IGFBP gene family. The phylogenies and chromosomal locations of three adjacent gene families, the epidermal growth factor receptors (EGFR) (Stein and Staros 2006), Ikaros transcription factors (John, Yoong, and Ward 2009) and distal-less clusters (Dlx) were also analyzed. Our results support the following scenario: an ancestral chordate IGFBP gene underwent a local gene duplication, resulting in a gene pair adjacent to a chordate ancestral HOX cluster. Subsequently, the gene family expanded in 2R, resulting in the six IGFBP types that presently constitute the gene family. In teleost fishes differential losses appear to have occurred in different lineages after the 3R event duplicated the gene repertoire to 12 genes. We also describe the identification of a single IGFBP-like sequence in the genome of the lancelet Branchiostoma floridae. Our detailed sequence comparisons show that several important structural components in the IGFBPs are ancestral vertebrate features that have been maintained in all orthologs, for instance the integrin interaction motif Arg-Gly-Asp in IGFBP-2. In contrast, the Arg-Gly-Asp motif in IGFBP-1 has arisen independently in mammals. The large degree of retention of IGFBP genes after the ancient expansion of the gene family strongly suggests that each gene evolved distinct and important functions early in vertebrate evolution. This likely includes other functions than the regulation of growth and metabolism via the growth hormone/IGF-1-axis, but distinct preferences for either IGF-1 or IGF-2 are also a possibility, as are differing expression patterns.

More recently we identified all six IGFBP subtype sequences in the genomes of the anole lizard, coelacanth and spotted gar, and updated our phylogenetic analyses of the gene family (Ocampo Daza, Bergqvist, and
Larhammar 2012). These analyses support the conclusions drawn in Paper VII and provide a higher-resolution phylogeny of IGFBPs.
Evolution of vertebrate visual opsins and transducin subunits

In the vertebrate retina there are two distinct types of photoreceptor cells, rods used for scotopic (dim light) vision and cones used for photopic (bright light) color vision. These two types of ciliary photoreceptor cells share distinct but related components in their cellular phototransduction cascades, starting with the expression of different photoreceptors, or visual opsins, each characterized by different spectral sensitivities (Larhammar, Nordström, and Larsson 2009; Shichida and Matsuyama 2009). Rods express the scotopic visual opsin rhodopsin while cones express different photopic color opsins.

Opsins are ubiquitous in the animal kingdom and constitute the largest subgroup of G-protein coupled receptors by far (Fredriksson et al. 2003; Nordström, Fredriksson, and Schiöth 2008). Depending on taxonomic sampling and method, the opsins have been grouped according to different systems of classification that roughly correspond to expression in different cell-types and different signal transduction pathways (Terakita 2005; Shichida and Matsuyama 2009; Porter et al. 2012). Together rhodopsin and the cone opsins form the vertebrate family of visual opsins. This family is characterized by the expression in vertebrate ciliary photoreceptor cells and by the activation of transducin (Gt)-mediated intracellular signaling.

Previous studies have suggested that most of the gene families in the phototransduction cascade expanded in 2R, including the vertebrate visual opsins and transducin subunit gene families (Nordström, Larsson, and Larhammar 2004; Larhammar, Nordström, and Larsson 2009). This concomitant expansion of phototransduction cascade components has implications for the diversification and specialization of rods and cones early in vertebrate evolution. However, due to chromosomal rearrangements and a low sampling of adjacent gene families for analyses of conserved synteny made conclusions uncertain.
Vertebrate visual opsins (Paper II)

The characterization of visual opsins in various vertebrate lineages has shown that the vertebrate visual opsin family consist of five members: rhodopsin (Rh1), ancestral green opsin (Rh2), ancestral ultraviolet opsin (SWS1), ancestral blue opsin (SWS2) and ancestral red opsin (LWS) (Davies et al. 2009; Collin 2010; Davies, Collin, and Hunt 2012). All five visual opsins have been described in the pouched lamprey (Geotria australis) (Collin et al. 2003; Davies et al. 2007), which implies that all five opsins existed before the divergence of lampreys and jawed vertebrates early in vertebrate evolution. Subsequently different vertebrate lineages have gained or lost visual opsin genes. For example, placental mammals have lost both Rh2 and SWS2. Extant mammals have instead adapted the ancestral ultraviolet-sensitive opsin SWS1 to detect light in the blue part of the spectrum and primates have a duplicate of the red-sensitive opsin LWS that has adapted its wavelength sensitivity to the green part of the spectrum. In teleost fishes many additional visual opsin duplicates have arisen, mainly through tandem duplications (Rennison, Owens, and Taylor 2012). Notably, teleosts have two Rh1 genes located on the same chromosome, one with introns which is expressed outside of the retina, called exo-rhodopsin, and one without introns which is expressed in rods, called rhodopsin (Mano, Kojima, and Fukada 1999). The zebrafish (Danio rerio) has two copies of the rhodopsin gene called rho and rhol. These are located on two different chromosomes (Morrow, Lazic, and Chang 2011).

In Paper II we describe the evolution of the chromosomal regions containing the genes for vertebrate visual opsins. During the analyses of conserved synteny, we realized that the chromosomal regions of the visual opsin genes overlapped with similar ongoing analyses of the oxytocin and vasopressin receptor gene family (OT/VP-R), the G-protein alpha transducing subunit (GNAT) and G-protein alpha inhibiting subunit (GNAI) gene families, as well as the gene family of L-type voltage-gated calcium channels (CACNA1-L). Therefore we used these gene families as starting points for extensive analyses of conserved synteny in species representing several vertebrate classes. We found that these families and 34 adjacent gene families comprise large paralogous chromosomal regions with extensive similarities to one another. This allowed us to confirm the previously proposed scenario for vertebrate visual opsin evolution (Nordström, Larsson, and Larhammar 2004; Larhammar, Nordström, and Larsson 2009): a local duplication occurred before 2R giving rise to an ancestral SWS gene and an ancestral LWS gene. These genes and their chromosomal region later duplicated in 2R so that the ancestral SWS gene gave rise to the SWS1, SWS2, Rh1 and Rh2 genes and the ancestral LWS gene gave rise to four copies, out of which only one has been retained (see Figure 5 in Paper II). The phylogenetic analyses of the visual opsin gene family presented in Paper II supports this scenario.
Several of the gene families that were used as a starting point for the conserved synteny analysis also acquired duplicates as part of the same quadrupled chromosome regions. The evolution of the OT/VP receptors has been discussed previously in this thesis, and the evolution of the GNAT gene family will be discussed below. In pouched lamprey all five vertebrate visual opsins genes have been previously identified. Their positions in our phylogenetic analyses, taken together with the robust analyses of conserved synteny, suggest that the lineage leading to lampreys diverged after 2R, as has previously been proposed by (Kuraku, Meyer, and Kuratani 2009).

We also find that major rearrangements have taken place in the teleost genomes, which obscure the view of the whole genome duplications. By comparing with the corresponding chromosomal regions in the spotted gar (Lepisosteus oculatus), which diverged prior to 3R, we could time these rearrangements to post-3R (see Extensive chromosomal rearrangements in the teleost genomes after 3R below).

The inclusion of the spotted gar in the phylogenetic and positional analyses also allowed us to time the local Rh1 gene retrotransposition giving rise to the intron-less rhodopsin and exo-rhodopsin. Four out of six identified spotted gar visual opsin genes represent the LWS, SWS1, SWS2 and Rh2 types, while the other two represent the Rh1 type. The two Rh1 genes are located on the same linkage group, one with introns and one without, suggesting that the intron-less rhodopsin and exo-rhodopsin arose early in ray-finned fish evolution and were preserved in both spotted gar and teleost genomes. There is also evidence from some adjacent gene families that the duplicate rho and rhol genes found in zebrafish were generated in 3R.

Transducin subunits (Paper III)

Transducins are the heterotrimeric G proteins (guanine nucleotide binding proteins) in the phototransduction cascade of vertebrate visual photoreceptor cells. The trimers consist of $G\alpha$, $G\beta$, and $G\gamma$ subunits, encoded by the unrelated GNAT, GNB and GNGT gene families respectively (Birnbaumer 2007). Each gene family has multiple members in vertebrate, out of which not all participate in the phototransduction cascade.

The GNAT genes and the related G-protein alpha inhibiting subunit (GNAI) genes are located adjacent to the vertebrate visual opsin genes in many vertebrate genomes. The phylogenetic analyses presented in Paper III together with the conserved synteny analyses of the vertebrate visual opsin gene-bearing chromosome regions in Paper II support the following scenario. An ancestral GNAT-GNAI gene pair arose through a local duplication before 2R. This gene pair subsequently expanded in 2R giving rise to a GNAT3-GNAI1 gene pair adjacent to SWS1, a GNAT1-GNAI2 gene pair adjacent to Rh1 and a GNAT2-GNAI3 gene pair adjacent to Rh2. GNAT3
genes have been lost at least twice independently, from the amphibian and teleost lineages. In Paper III phylogenetic analyses of the GNB gene family were combined with conserved synteny analyses. The phylogenetic and positional data from eight adjacent gene families support the expansion of the transducin Gβ subunit family in both 2R and 3R. The GNGT genes are located in the same genomic regions as the HOX gene clusters, which have duplicated in both 2R and 3R (Sundström, Larsson, and Larhammar 2008). The short amino acid sequence of the Gγ subunits precludes reliable sequence-based phylogenetic analyses. However, the location of Gγ subunit genes in the HOX regions, as well as their species representation, supports duplications in both 2R and 3R.

Thus, we have found that all three transducin subunit families expanded in 2R, and that the transducin Gβ and Gγ subunit families expanded further in the teleost-specific 3R event. Thus, the early vertebrate tetraploidizations provided the basis for the subsequent specialisation of transducin subunits leading to differential visual specializations in vertebrates.
Extensive chromosomal rearrangements in the teleost lineage after 3R

The results presented for the paralogous chromosomal regions bearing visual opsins and OT/VP receptor genes (Paper II), and the paralogous chromosomal regions bearing SSTR genes (Paper VI), show that there have been extensive chromosomal rearrangements in teleost genomes. The rearrangements involve homologous chromosome regions generated in 2R but occurred in the teleost lineage after 3R. For the SSTR gene-bearing chromosome regions there is also evidence from several gene families that three of the 2R-generated chromosomes fused before 3R, and that one of these fused and duplicated chromosomes split through a fission after 3R (see Figure 6 in Paper VI for a summary).

These rearrangements made the evolutionary analysis of several gene families presented in this thesis difficult since they suggested other mechanisms than duplications in 2R and obscured the orthology relationships between gene family members. It was only by considering both phylogenetic and positional data in a broad selection of vertebrate genomes that these events could be sorted out. These analyses also demonstrate the importance of having the appropriate out-groups to determine the (relative) time points of the events. In both Paper II and Paper VI it was possible to confirm the likely ancestral paralogy relationships between the rearranged chromosome regions in teleosts by comparing our findings against the genome of the spotted gar.

Our results are consistent with previously published large-scale genomic analyses (Kasahara et al. 2007; Nakatani et al. 2007), but resolve the fusions and rearrangements in greater detail. They are also consistent with the finding that linkage between genes in the same chromosomal regions was disrupted at a faster rate after 3R (Sémon and Wolfe 2007a). These results support the idea presented by (Sémon and Wolfe 2007b) that chromosome rearrangements accelerated as a consequence of 3R. The analysis of the spotted gar genome (Amores et al. 2011) also supports this, finding that it has more conserved synteny with the human genome than compared with teleost genomes due to post-3R rearrangements in the teleost lineage. In our analyses the exchange of (2R-generated) paralogs occurred specifically between paralogous chromosome regions, likely due to recombination facilitated by homology. However, we can not discard other rearrangements in the ances-
tral teleost genome after 3R because our method only considered conserved synteny between putatively paralogous chromosome regions.
Conclusions, ongoing research and perspectives

There is now convincing evidence that the vertebrate lineage went through two rounds of whole genome duplication (called 1R and 2R) early in its evolution (Dehal and Boore 2005; Nakatani et al. 2007; Putnam et al. 2008). Subsequently a third round of whole genome duplication (called 3R) occurred early in teleost fish evolution (Jaillon et al. 2004; Meyer and Van de Peer 2005; Kasahara et al. 2007). Numerous endocrine and neuronal gene families have expanded as part of these vertebrate whole genome duplications (Dreborg et al. 2008; Larsson et al. 2008; Sundström, Larsson, and Larhammar 2008; Braasch, Volf, and Schartl 2009; Sundström, Dreborg, and Larhammar 2010; Dos Santos et al. 2011; Widmark et al. 2011; Hultqvist et al. 2012). This thesis resolves the evolution of several neuronal and endocrine gene families in the framework of vertebrate whole genome duplications. The gene families of oxytocin and vasopressin receptors (OT/VP-R), somatostatin receptors (SSTR) and insulin-like growth factor binding proteins (IGFBP) expanded in 2R and in several cases also in 3R. For the OT/VP-R and SSTR families, previously undescribed subtypes were identified. The protein hormone family that includes growth hormone (GH), prolactin (PRL) and somatolactin (SL) acquired a new PRL gene in 2R, PRL2, however the origins of GH, PRL, and SL likely predate 2R. The corresponding family of receptors diversified during different time periods through a combination of a local duplication and 3R. This raises interesting questions regarding ligand-receptor relationships in this system.

Neuronal gene families of the visual system also expanded in 2R and 3R. The results presented in this thesis demonstrate that the ancient vertebrate repertoire of visual opsin genes arose in 2R as part of chromosomal blocks that also include the OT/VP-R genes. The gene families including the transducin alpha (GNA), beta (GNB) and gamma (GNG) subunits also arose in 2R, hinting at the importance of these events in the diversification and specialization of phototransduction cascades for rods and cones.

Papers II, IV and V are presented as manuscripts and will be submitted shortly for publication. In addition to the papers included in this thesis, studies have been carried out resolving the evolution of the voltage-gated sodium channel alpha subunits (SCNa) (Widmark et al. 2011; Ocampo Daza et al. 2012) and the paralemmins (PALM) (Hultqvist et al. 2012).
The chromosomal analyses of the visual opsin gene-bearing chromosome regions and the SSTR-bearing chromosome regions have allowed the elucidation of extensive chromosomal rearrangements between paralogous regions in the teleost fish genomes. The only model that is consistent with these results is that the rearrangements occurred after 3R. This in accordance with previous studies of teleost genomes (Kasahara et al. 2007; Nakatani et al. 2007) and supports the hypothesis that chromosome rearrangements were accelerated as a result of 3R (Sémon and Wolfe 2007a). For the SSTR-bearing chromosome regions we could also detect chromosome fusions preceding 3R. These large-scale chromosomal events, and the resulting loss of conserved synteny, made the evolutionary history of the visual opsin, OT/VP-R and SSTR gene families particularly difficult to resolve. It was only through the combination of curated sequence-based phylogenetic analyses and chromosomal studies of conserved synteny across a broad sample of vertebrate taxa that it could be done.

Studies are ongoing on the evolution of the urotensin II receptor (UTS2R) family of GPCRs, which are located in the chromosomal regions of the SSTR gene family described in Paper VI. Mammals retain only one UTS2R receptor gene (Liu et al. 1999; Tostivint et al. 2013), however we have found that the ancient setup of this gene family consists of five genes. Four UTS2R genes arose from a common ancestor in the 2R chromosome duplications and subsequently one of the subtype genes underwent a local duplication. The spotted gar (Lepisosteus oculatus) has retained all five ancestral subtypes and the teleost fish genomes we have investigated conserve four to five of them. In the tetrapod lineage there seem to have been several differential losses that complicate the view: we have investigated reptilian genomes and conclude that only non-avian reptiles have conserved all five ancestral UTS2R genes. The analyzed reptilian genomes include the newly available genome assemblies of Chinese softshell turtle (Pelodiscus sinensis) and painted turtle (Chrysemys picta bellii), and drafts of the Burmese python (Python molurus) (Castoe et al. 2011) and American alligator (Alligator mississippiensis) (St John et al. 2012) genomes. Placental mammals and marsupials have conserved only one UTS2R gene, however each lineage has conserved a different subtype. This is evidenced by comparison to the monotreme platypus (Ornithorhynchus anatinus) which has conserved both UTS2R subtypes. The evolution of the UTS2R family has been challenging to sort out due to the many gene losses within the tetrapod lineage and the relatively high degree of sequence conservation between different UTS2R subtypes, which makes phylogenetic analyses unreliable. However, our careful analyses of the SSTR-bearing chromosome regions which also house the UTS2R genes, have allowed us to resolve the evolutionary history of both gene families. Analyses of conserved synteny in the available reptilian genomes as well as in the spotted gar genome assembly are planned in order to bring more clarity to the evolution of this gene family.
In the studies presented in Paper II we have carried out orthology-prediction based analyses of conserved synteny in the genome of the spotted gar. This allowed us to time the rearrangements of the chromosomal regions carrying visual opsin genes and OT/VP receptor genes to the early evolution of teleosts, post-3R. Sequences and positional data from the spotted gar were also useful in the analyses of the visual opsin gene family (Paper II), the GH, PRL and SL gene family (Paper IV) and of the gene family of their corresponding receptors GHR and PRLR (Paper V). This is true for several reasons: 1) The spotted gar genome assembly has been mapped to linkage groups (Amores et al. 2011). This makes it possible to analyze conserved synteny, which is valuable both for comparative orthology assignments of genes and for the comparative analysis of genome rearrangements. 2) The phylogenetic position of the spotted gar as part of an early-diverging lineage among ray-finned fishes (Inoue et al. 2003; Near et al. 2012) makes it an appropriate out-group for the analysis of teleost evolution, in particular the impact of 3R (Amores et al. 2011).

In the same way, assembled and mapped whole genome sequences from a cyclostome species, for instance a lamprey, and a cartilaginous fish, would be useful to increase the resolution of several of our analyses. Because the cyclostomes represent the earliest extant lineage of vertebrates, and the cartilaginous fishes diverged before the lobe-finned fishes and ray-finned fishes (Benton 2005), they constitute highly relevant reference points for the study of basal vertebrate evolution. Currently there is a sea lamprey (Petromyzon marinus) genome assembly available in the Ensembl genome browser. However, due to its incomplete and unmapped status the searches in this genome assembly have largely had negative results (see Paper IV for example), and it has not been possible to do synteny analyses. The study of conserved synteny to infer orthology relationships would be of particular importance for lamprey genes since phylogenetic analyses that include lamprey sequences often give ambiguous branching order (Qiu et al. 2011). There is also an incomplete and unmapped survey sequence of the elephant shark (Callorhinchus milii) genome available (Venkatesh et al. 2007). OT/VP receptor sequences were identified from this cartilaginous fish genome for the analyses presented in Paper I, however several of these were partial sequences due to the shortness of the assembled genomic scaffolds. Elephant shark sequences have been useful for the evolutionary studies of the OT and VP pro-peptide genes (Gwee et al. 2009), the OT/VP receptors (Ocampo Daza, Lewicka, and Larhammar 2012; Yamaguchi et al. 2012) and the neuropeptide Y-family peptides and receptors (Larsson et al. 2009).

Complete high coverage mapped genome assemblies from cyclostome and cartilaginous fish species would allow for more accurate gene searches, phylogenetic analyses and studies of conserved synteny in the perspective of early vertebrate evolution. This would, for instance, help answer remaining questions regarding the evolution of the growth hormone, prolactin and so-
matolactin gene family, and of the gene family of their corresponding receptors. For these families local duplication events seem to have occurred before the divergence of lobe-finned and ray-finned fishes at the latest, but it was not possible to say how early in evolution they could have occurred. The origin of somatolactin is also still unknown. The identification of the complete gene family repertoire in cyclostome and cartilaginous fish species would help disentangle the relative time-points for these events.

Det är nu känt att hela arvsmassan fördubblades två gånger tidigt i ryggradsdjurens evolution. Detta var en avgörande händelse i vår evolutionära historia och alla nu levande ryggradsdjur bär på spår av denna fyrdubbling. I teleosterna fördubblades arvsmassan en tredje gång. Dessa genomduplikationer skapade kopior av många gener hos många genfamiljer. En genfamilj är en grupp besläktade gener som har uppkommit från en gemensam förfader.

I detta avhandlingsarbete undersökt evolutionen hos flera genfamiljer inblandade i ryggradsdjurens nervsystem och hormonsystem. Närmare bestämt undersökt det om nya gener i dessa genfamiljer uppkom i de tre genomduplikationerna i ryggradsdjurens evolution. Genom att söka efter genfamiljernas medlemmar i arvsmassan från många olika arter ryggradsdjur och jämföra både deras sekvenser, med hjälp av fylogenetiska analyser, och deras positioner i arvsmassan mellan arterna, är det möjligt att dra slutsatser om genfamiljernas evolution. Resultaten visar att flera genfamiljer som deltar i våra hormonsystem samt i vårt syn-system fick nya gener både i de genomduplikationer som skedde tidigt i ryggradsdjurens evolution och i den tredje genomduplikationen som skedde i de egentliga benfiskarna.

Hormonerolyt och vasopressin frisätts till kroppen från hypofysen, men är också viktiga signalsubstanser inne i hjärnan. Vasopressin kontrollerar bl. a. kroppens vätskebalans och blodtryck, och oxytocin frisättningen av mjölk hos däggdjur men också sociala beteenden när den agerar i hjärnan. De utför sina funktioner i mål-vävnaderna genom att binda till proteiner som kallas receptorer på cellernas yta. Resultaten i denna avhandling visar att det finns flera typer av vasopressin-receptorer i ryggradsdjur som tidigare inte har beskrivits, samt att den uppsättning oxytocin- och vasopressin-receptorer som är gemensam för ryggradsdjur har uppkommit genom duplikationerna av arvsmassan tidigt i ryggradsdjurens evolution.

Vidare visar resultaten att flera genfamiljer som deltar i hypofysens reglering av tillväxt och ämnesomsättning också har fått nya medlemmar i genomduplikationerna. Detta innefattar genfamiljen för tillväxthormon och dess släktingar prolaktin och somatolaktin, samt genfamiljen för deras motsvarande receptorer. Just de här genfamiljerna har varit svåra att analysera eftersom tillväxthormon och prolaktin inte verkar ha uppkommit i genomduplikationerna. Däremot gav de tidiga genomduplikationerna upphov till en extra tidigare okänd variant av prolaktin. Tillväxthormonets och prolaktinets receptorer fick nya kopior senare i evolutionen, i teleosternas genomduplikation. En familj proteiner som kallas IGF-bindande proteiner fick ytterligare medlemmar både i de tidiga genomduplikationerna och i teleosternas genomduplikation. Vissa fiskarter har därmed upp till tolv olika medlemmar medan vi människor bara har sex. IGF-bindande proteiner regle-
rar tillväxthormonets effekter på tillväxt och ämnesomsättning i kroppens vävnader.

En annan familj som deltar i tillväxthormonets reglering av tillväxt och metabolism är genfamiljen för somatostatin-receptorer. Somatostatin är ett hormon som frisätts till hypofysen från hjärnan och minskar frisättningen av tillväxthormon till kroppen. Resultaten i denna avhandling visar att det finns ytterligare en typ av somatostatin-receptorer som inte har beskrivits förut, utöver de fem typer som redan var kända. Dessa sex olika typer av somatostatin-receptorer uppkom i de tidiga genomduplicationerna, och flera av dem fick ytterligare kopior i de egentliga benfiskarnas genomduplikation. Studierna av de delar av arvsmassan som innehåller generna för somatostatin-receptorer visade också att arvsmassan omstrukturerades väldigt mycket både före och efter den tredje genomduplikationen i teleosterna. Forskare har tidigare föreslagit att denna genomduplikation ledde till att denna grupp fisikars arvsmassor arrangerades om mycket mer än hos andra grupper av ryggradsdjur.


Sammantaget visar denna avhandling att ryggradsdjurens hormonsystem och syn-system troligtvis nådde stora delar av sin nuvarande form tidigt i evolutionen, och att den huvudsakliga mekanismen för detta var fördubblingarna av hela arvsmassan i flera omgångar. Detta visar vilken viktig roll genomduplikationer har haft för uppkomsten av ryggradsdjurens specialiserade anpassningar.
Acknowledgements

I would like to acknowledge the contributions of the following people to the work presented in this thesis. My thesis supervisor Dan Larhammar, whom I will always thank first and foremost. Görel Sundström, who took me in from day one, and taught me everything I needed to know about the practicalities of genome databases and phylogenies. Christina Bergqvist did all the preliminary sequence and phylogenetic analyses for Papers VI and VII. Our talks, pouring over countless gene predictions, sequence alignments and phylogenetic trees over and over again have been some of the most fun times of this work. Görel, David Lagman and Jenny Widmark contributed in equal parts to the work in Paper II together with me. The projects I have shared with you three have always been the most rewarding, and our papers together were the most fun to write and discuss. David also did the analyses for Paper III with my support and supervision. Susanne Dreborg helped initiate the visual opsin evolution project in Paper II together with Görel. Michalina Lewicka did the preliminary analyses for Paper I together with me.

I also want to acknowledge the contributions of co-authors and collaborators. Cunming Duan for his contributions to Paper VII. Xesus Abalo for his input to Papers II and III. Lars-Gustav “Ludde” Lundin for his invaluable input to our general discussions about paralogons and genome evolution. Greta Hultqvist for our excellent collaboration. Although our paper is not part of this thesis I am very proud and happy for the work we did together.

Thank you…

… to my wonderful colleagues, to teachers I’ve had, to my sweet friends & a mi familia; todo lo mío es suyo.

Daniel Ocampo Daza
Uppsala, January 2013
Open science declaration and supporting information

The analyses presented as in this thesis were made possible thanks to the organized effort to make genetic sequence data as well as whole-genome data from diverse genome sequencing projects publicly available.

In the same spirit, it is my goal to share all data files and supporting information underlying the scientific papers of which I am the principal author under Open Science principles. These principles include the freedom for anyone to use, reuse and redistribute the data and supporting information – subject only, at most to the requirement to attribute and/or share-alike (adapted from the Open Knowledge Foundation definition, see opendefinition.org).

The phylogenetic data and supporting information used for Papers I, II, VI and VII have been shared in figshare (figshare.com) under a Creative Commons license. They can be accessed through the stable link goo.gl/1u0Uq or by searching for “Daniel Ocampo Daza” on figshare.com. The data used for Papers IV and V will be shared before the manuscripts are submitted for publication. The figshare website provides stable DOIs (Digital Object Identifiers), making the data easily shareable, searchable and citable. The phylogenetic tree files included in the shared datasets are in Phylip/Newick-format and alignment data files are shared in both sequential and interleaved formats. These file formats can be opened by most freely available bioinformatics programs for sequence analysis and phylogenetic tree viewing, such as eBioX (ebioinformatics.org), Jalview (jalview.org) and FigTree (http://tree.bio.ed.ac.uk).
Corrections

Paper I incorrectly states that “the V₂ receptor inhibits adenylyl cyclase, thereby reducing the production of cAMP” on page 136. In fact the V₂-type vasopressin receptors stimulate adenylyl cyclase and increase the cytosolic cyclic AMP release, see for instance (Schöneberg et al. 1998). This mistake was reported in the proof-reading phase of pre-publication, but the correction was not carried to the final version of the article.

Also in Paper I, the first paragraph of the discussion on pages 137-139 suggests that all five identified OT/VP receptor subtypes are present in the anole lizard, frog and teleost fishes. In fact, teleost fishes lack one of the receptor subtypes, V₁B. This does not affect the conclusions drawn in the paper.
References


Garrison JL, Macosko EZ, Bernstein S, Pokala N, Albrecht DR, Bargmann CI. 2012. Oxytocin/vasopressin-related peptides have an ancient role in reproductive behavior. Science (New York, N.Y.) 338:540–3.


Ocampo Daza D, Lewicka M, Larhammar D. 2012. The oxytocin/vasopressin receptor family has at least five members in the gnathostome lineage, including two distinct V2 subtypes. General and comparative endocrinology 175:135–43.


Stein RA, Staros JV. 2006. Insights into the evolution of the ErbB receptor family and their ligands from sequence analysis. BMC evolutionary biology 6:79.


A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine.