Nuclear Receptor Genes - Regulation and Evolution

Yogita Sharma

Dissertation for the degree of philosophiae doctor (PhD)
at the University of Bergen

2016

Dissertation date: 2016-02-26

Scientific Environment

This work has been performed in

Computational Biology Unit (CBU)

UniComputing

Yogita Sharma is affiliated with

Department of Biomedicine

Acknowledgements

I would like to thank all people who have helped and motivated me to walk through this journey and make this thesis possible. In particular:

My supervisor, Professor Boris Lenhard: I would like to express my special thanks and appreciation to you for encouraging my research. Your immense knowledge and guidance helped me in all the time of research.

My co-supervisor, Professor Marit Bakke: thank you very much for the insightful comments to my thesis. It helped me a lot during writing.

I would also like to thank all my group members for their co-operation and making our workplace motivating as well as enjoyable. Gemma Danks: thank you for spending time in discussions, reading manuscripts and introducing tea break. Christopher Previti, Xianjun Dong: thank you for interesting discussions; it helped me getting my first paper.

All my colleagues in CBU – thank you for offering friendly environment. I loved our new CBU seminar series with evaluation and feedback. I am also very thankful to all the administration staff at CBU, Uni Computing and department of Biomedicine. Monika Voit: You have always been very helpful. Amra Grudic: thank you very much for helping me during the submission process.

I am also thankful to Norwegian Research Council for funding my studies.

And last but not the least I am extremely grateful to my family for their love and support. Papa it was your dream. You were the first one who encouraged me to go for a PhD. Words cannot express how grateful I am to my husband. Vinay: it would not have been possible without you. I am also thankful to my sister for her love and support. I cannot forget to thank my little bundle of joy

with whom life is so much fun. I am very thankful to all my friends, especially, Reena and Astha.

Finally, I would like to thank the one who loved me the most, my mother. Mamma this is for YOU. Your prayer for me was what sustained me this far. I wish I could have thanked you in person. You were and will always be my inspiration. Love You!

Abstract

Nuclear receptors are transcription factors that typically bind ligands in order to regulate the expression level of their target genes. Members of this family work with their co-regulators and repressors to maintain a variety of biological and physiological processes such as metabolism, development and reproduction. Nuclear receptors are promising drug targets and have therefore attracted immense attention in recent decades in the field of pharmacology. Irregular expression of nuclear receptor genes is linked to various metabolic and proliferative diseases such as cancer, diabetes and obesity.

Despite extensive study in this area, how nuclear receptor genes are regulated is still poorly understood. As regulators of other genes, nuclear receptors and their activites are tightly regulated themselves. We propose that diversity of their biological and biochemical roles will be reflected in fundamental differences of their transcription regulation mechanism. We aimed to study the impact of regulatory content and evolutionary history of nuclear receptors genes on their expression and current function. To facilitate this work we used the Genomic Regulatory Block (GRB) for studying regulation of nuclear receptor genes in connection with their known function.

In this thesis, I present a new classification of nuclear receptor genes on the basis of their *cis*-regulatory environment. We identified the nuclear receptor genes that are putative targets of long-range gene regulation. These genes are involved with developmental related functions and are characterized by the presence of highly conserved non-coding elements, CpG islands, bivalent promoter marks and specific combinatorial patterns of histone modifications. We also explored the evolutionary history of nculear receptor genes in context to our proposed classificiation. We found that nuclear receptors genes that are under long-range gene regulation exhibit negative selection pressure in comparison to GRB non-target genes. This is suggestive of an evolutionary constraint and shows that the functions of nuclear receptors have been recruited since the ancestral time.

List of Publications included in the thesis

I. Yogita Sharma, Chandra Sekhar Reddy Chilamakuri, Marit Bakke and Boris Lenhard (2014): ”**Computational characterization of modes of transcriptional regulation of nuclear receptor genes**”. *PLoS ONE,* 10.1371/journal.pone.0088880

II. Xianjun Dong, Altuna Akalin, Yogita Sharma and Boris Lenhard (2010): ”**Translog, a web browser for studying the expression divergence of homologous genes**”. *BMC Bioinformatics,* 10.1186/1471-2105-11-S1-S5

III. Yogita Sharma**,** Marit Bakke and Boris Lenhard: ”**Evolution of nuclear receptor genes**”. Manuscript under submission.

List of Publications I contributed to:

I. Madhumohan R. Katika, Siv Gilfillan, Jens Henrik Norum, Elisa Fiorito, Shixiong Wang, Venkata Somisety, Siri Nordhagen, Helga Bergholtz, Helene Zell Thime, Baoyan Bai, Yogita Sharma, Silje Nord, Kristine Kleivi, Meritxell Bellet, Anne-Lise Børresen-Dale, Therese Sørlie and Antoni Hurtado (2015): “**FOXA1 is a central mediator of tumorigenesis in HER2/3-driven breast cancers via Estrogen Receptor independent mechanism”.** Submitted.

II. Elisa Fiorito*, Yogita Sharma*, Siv Gilfillan, Baoyan Bai, Alfonso Urbanucci, Ian Mills and Antoni Hurtado (2015): “**Characterization of the role of CTCF in Estrogen Receptor dependent gene regulation**”. Manuscript.

Table of Contents

Nomenclature

Abbreviation	Description
AF-1	Activation Factor 1
AF-2	Activation Factor 2
DBD	DNA-binding domain
GRB	Genomic regulatory Block
HCNE	Highly conserved non-coding element
HRE	Hormone Response Element
LBD	Ligand-binding domain
NTD	N-terminus domain
TBP	TATA-box binding protein
TSS	Transcription start site

List of Figures

List of Tables

1. General Introduction

Dr. Jensen, you certainly have filled a tremendous gap in the information that we have wanted for a long time; that is, the state of hormones in the tissue during response to hormone.

- Gerald Mueller, Discussion of (Jensen and Jacobson, 1960), 1960

So far, 48 different nuclear receptors have been found in the human body. For many of these we have not yet found the signaling molecule. Just imagine the great strides in medicine when the correct signaling molecules are discovered.

- Michael Brown, Presentation of Albert Lasker Basic Medical Award, 2004

In 1960, Elwood V. Jensen (Jensen and Jacobson, 1960, Jensen and Jacobson, 1962) identified the first nuclear receptor proteins, namely, estrogen receptors. Nuclear receptors such as estrogen receptors facilitate intra-cellular signalling by binding to a specific ligand and subsequently, regulating the expression of specific genes. For example, upon fertilization, the ovary secretes estrogen, which binds to the estrogen receptors present in the cells of the uterus lining. The hormone-activated estrogen receptors translocate to the cell nucleus and regulate the expression level of specific genes. The altered gene expression levels result in increased cell growth in the uterine lining – a prerequisite for survival of the offspring.

Several nuclear receptors have since been identified, and, it is now known that nuclear receptor-based signalling is a key component in several biological processes including sexual maturation, metabolism, mineral absorption, vitamin signalling, drug and hormone detoxification (Evans, 2004). Moreover, nuclear receptor signalling is highly tissue specific; for example, estrogen-based activation of the estrogen receptor regulates different sets of genes in brain, uterus and that are in turn responsible for the different functions of those organs (Sever and Glass, 2013). The ligand-specific and tissue-specific nature of nuclear receptors make them very promising drug targets (Burris et al., 2012).

Given the importance of nuclear receptors in signalling pathways and their viability as potential drug targets, considerable effort has gone into understanding the structure, function and mechanism of cell-specific transcriptional regulation by nuclear receptors as transcription factors.

What is less understood is the *transcriptional regulation of nuclear receptors by other transcription factors including other nuclear receptors.* Understanding how nuclear receptors undergo regulation (including possible co-regulation by other nuclear receptors) provides crucial information regarding their origins and their functionality. For example, long-range transciptional regulation is often associated with developmental genes, which perform several roles in different tissues and therefore, require precise spatial and temporal control. In contrast, tissue-specific genes have a simpler regulatory mechanism. Knowing whether a nuclear receptor gene is the target of long-range transcriptional regulation provides insight into whether the gene is a developmental gene, which can perform distinct functions at different times and in different tissues? Further, for a nuclear orphan receptor, i.e., a nuclear receptor with no known ligand, long-range transcriptional regulation could be indicative that gene activation occurs by transcriptional regulation and histone modification rather than ligand binding.

1.1 Aim of Study

This thesis explores regulation *of* nuclear receptors by examining the cis-regulatory environment of each member of the nuclear receptor family. I attempt to address two main questions in this thesis, namely,

- How does regulation of nuclear receptor genes relate to their functions? More specifically, we investigate whether a nuclear receptor gene is target of long-range gene regulation or not.

- Is there a connection between regulation of nuclear receptor genes and their evolutionary history?

1.2 Key Findings

We used the Genomic Regulatory Block (GRB) model (Kikuta et al., 2007) for studying regulation of nuclear receptor genes in connection with their known function and evolutionary history. The GRB model consists of the target gene, which usually is a developmental gene requiring complex spatio-temporal regulatory architecture for its expression, bystander genes and highly conserved non-coding elements (HCNE) which regulate the expression level of the target gene(s).

- We present a new classification of nuclear receptor genes based on their transcriptional regulatory mechanisms (Paper I). More specifically, we identify nuclear receptor genes that are putative targets of long-range gene regulation in the GRB model. These genes are characterized by presence of highly conserved non-coding elements, CpG islands, bivalent promoter marks, and distinguished combinatorial patterns of histone modifications. This is strongly indicative of these nuclear receptors being involved in developmental processes.

- We explore the duplication mechanism of nuclear receptor genes in the context of our proposed classification using synteny and map-based analysis (Paper II, Paper III). We find that most nuclear receptor genes show purifying selection pressure with targets of GRB model exhibiting more negative selection in comparison to non-targets. Since many nuclear receptor genes were derived from a common ancestor in the course of metazoan evolution, long-range regulation is suggestive of an evolutionary constraint and that it is the ancestral and not the current gene loci that have been recruited into developmental or tissue-specific roles.

2. Background

To take a computer science analogy, DNA is a stored program, which is "executed" by transcription to RNA and expression to protein.

- William W. Cohen, A Computer Scientist's Guide to Cell Biology, 2007

Much of biology is devoted to the study of genes – how they are regulated, their transcription into RNA, and how genetic and epigenetic factors govern the evolution of organisms and species. In this chapter, I review some of the relevant concepts related to regulation of genes and their evolution.

2.1 Gene Regulation

There is no precise consensus definition of a gene but broadly it refers to any part of genome that is being transcribed. The control of gene expression at the stage of synthesis of RNA from DNA (transcription) is referred to as "gene regulation". This is a complex process crucial mechanisms for maintaining the complexity of eukaryotic organisms, and is controlled at many stages including transcription (pre) initiation, elongation and termination (Wasserman and Sandelin, 2004). In general, the process of transcription starts with the recruitment of core transcription machinery consisting of general transcription factors, activators and co-activators (Coulon et al., 2013). Transcription factors (TFs) are a specialized class of proteins that bind to specific DNA sequences located in the regulatory regions of their target gene, and, either inhibit or stimulate the rate of transcription of the target gene (Spitz and Furlong, 2012). They form pre-initiation complexes at core promoters close to the transcription start sites (TSS) of genes. This in turn helps to recruit RNA polymerase to the TSS.

2.2 Long-range transcriptional regulation

Given the complexity of multi-cellular eukaryotes, a very precise temporal and spatial regulation of gene expression is required. Transcription initiation controlled by complex arrangement of cis-regulatory regions consisting of multiple clustered enhancer modules interspersed with silencers and insulators, is the key important feature for regulation of gene expression. (Maston et al., 2006) (Figure1).

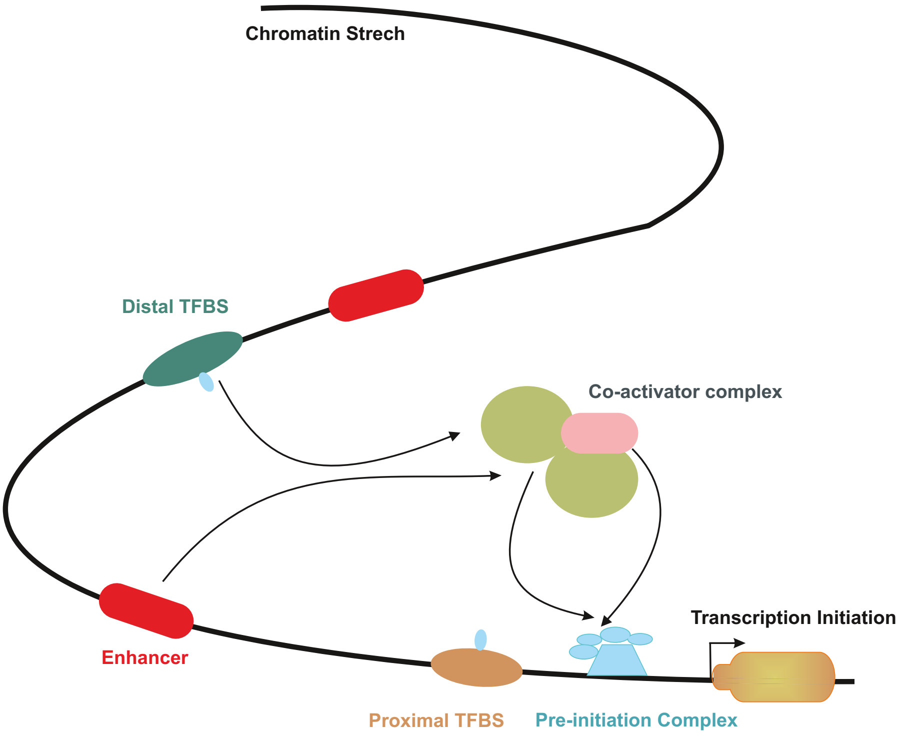


Figure 1. The long-range gene regulatory elements in eukaryotes. The promoter consists of a core promoter and proximal promoter elements. Long-range regulatory elements include enhancers, silencers, insulators and locus control regions. These regulatory elements, can be located far away (1MB) from the promoter region. All these regulatory elements interact with each other to carry out the function of single transcription unit. The figure is adapted from (Wasserman and Sandelin, 2004).

The regulatory elements modulating gene transcription can be broadly categorized into two parts, namely, a promoter composed of core and proximal regulatory elements, and, secondary distal regulatory elements such as enhancers, silencers, insulators and locus control regions. I describe each of these in greater detail below.

The promoter region of eukaryotic genes is complex and composed of core and proximal promoters. The core promoters contain DNA response element for the binding of transcription factors and serve as the site for the basic transcriptional machinery and pre-initiation complexes. It defines the position and direction of transcription. The most well-studied core promoter element is the so-called "TATA box", which serve as binding site for TBP (TATA-box-binding protein) (Sandelin et al., 2007). The core promoter also includes other nearby regulatory regions such as the initiator, downstream core and downstream promoter elements in addition to the TATA-box.

The proximal promoter is defined as the region upstream (some hundred base pairs) of the core promoter. This region contains multiple binding sites (Sandelin et al., 2007). Mutational analyses typically reveal that the proximal promoter needs to be intact for full transcriptional activity (Taylor et al., 2006).

2.2.1 Enhancers

Enhancers are regulatory elements (50-2000 bp) typically located far from the core promoter (e.g., 1Mb away from TSS) that up-regulate or "enhance" the rate of transcription of the target gene. They work in a distance and orientation independent manner, and hence, can be located upstream, downstream or within the coding region of the target gene. A typical enhancer has binding sites for multiple sequence specific transcription factors. (Pennacchio et al., 2013).

The identification of various enhancers over past decades has shown that they often function in a spatial or temporal manner (Shlyueva et al., 2014). Several enhancers located at distinct positions along the chromosome can regulate a single promoter at different time-points, in a tissue-specific manner, in response to different stimuli (Andersson, 2015).

2.2.2 Silencers

Silencers are short regulatory elements (20-2000bp), which down-regulate or "silence" the rate of gene expression either by binding to transcription factors that act as repressors or by recruiting co-repressors. Silencers share many similarities with enhancers - they are highly sequence-specific, can function in a distance and orientation independent manner, and, are located far from the core promoter (Riethoven, 2010). There are several proposed models for the functioning of repressors. For instance, they can either block the binding of the activator or may compete for binding to the same site (Ogbourne and Antalis, 1998).

2.2.3 Insulators

Insulators, also known as “boundary elements/domain boundaries”, are regulatory regions that prevent inappropriate regulation of neighboring genes and their associated enhancers. Insulators act either as barriers to the spread of repressive chromatin or can block enhancer-promoter interactions (Gaszner and Felsenfeld, 2006) . In vertebrates, CTCF (CCCTC binding transcription factor) has been identified to mediate insulator activity (Phillips and Corces, 2009).

2.2.4 Locus control regions

A locus control region is a group of regulatory elements required for the regulation of a set of linked genes. It is composed of multiple cis-regulatory elements (such as enhancers, insulators and silencers) (Li et al., 2002). All these elements serve as target sites for the binding of transcription factors, co-factors (activators/repressors) and chromatin modifiers, and the collective effect of all these regulatory elements accounts for the functioning of the locus control region (reference needed). A locus control region can also function in position independent manner (similar to enhancers and silencers). The most well studied locus control region in the human genome is the beta globulin locus control region, which is composed of five genes (Levings and

Bungert, 2002). This region has shows the evidence of “DNA looping model” to control the transcriptional activity (as in case of enhancers) (Levings and Bungert, 2002). The looping model proposes a direct interaction between LCR and individual genes. This is further supported by another study, which shows that patterns of histone acetylation across the globin locus vary during the development (Forsberg et al., 2000).

Together, all these regulatory DNAs described above work as activators, repressors or they might serve as “tethering elements” to recruit distant enhancers to the core promoter. The circuitous interplay between transcription factors and cofactors along with cis-regulatory module (CRM) stabilize the transcription-initiation machinery that controls the activity of single transcription unit (Figure 1).

Thus, long-range gene regulation comprising of multiple regulatory elements directs the complex patterns of expression in many different cell types during development. Identifying the *cis*-regulatory elements is critical for understanding of the biochemistry underlying transcriptional regulation. One indication of a region being a possible regulatory element is given by the fact that the *cis*-regulatory elements are under purifying selection pressure in order to maintain the function of the target gene. This has been used to identify the long-range regulatory elements in the Genome Regulatory Block model (Kikuta et al., 2007), which I describe next.

2.3 Genome Regulatory Blocks

The term genomic regulatory block (GRB) associated with a target gene refers to locus/block in the genome that has all the long-range regulatory inputs required for normal expression of the target gene. Usually, the target gene is a developmental regulated gene that requires specific spatial-temporal control of its expression through long-range regulatory input.

In addition to the target gene, the GRB model includes bystander genes and highly conserved non-coding elements (HCNEs) (Figure 2) (Kikuta et al., 2007). The bystander genes refer to the neighboring genes of a target gene within the GRB locus.

Bystander genes do not interfere with the (long-range) regulation of the target gene; and, consequently, could be shifted to outside the GRB locus during the process of evolution.

The complete GRB locus needs to be present for the normal expression of the target gene. Any mutation/deletion in the GRB causes improper regulation of gene, and might contribute to diseases. Moreover, it has been reported that many developmental regulated genes are targets of GRB model. Within a putative GRB locus, the target gene is identified based on its known functionality as well as the presence of HCNEs and long CpG islands, which I describe below.

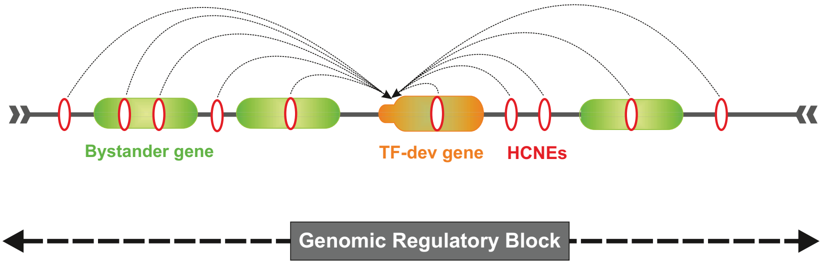


Figure 2. The GRB model. A GRB has developmental and/or transcription factor genes (target gene shown in orange colors in figure) surrounded by highly conserved non-coding elements (HCNEs) (shown in red oval in the figure), which regulates the target gene expression by acting as enhancers/insulators and other un-related neighboring genes (bystander genes, green in figure). This figure is adapted from (Akalin et al., 2009).

2.3.1 Highly conserved non-coding elements

Highly conserved non-coding elements (HCNEs) genomic regions that are highly conserved between two or more species, with a minimum specific length and threshold conservation level. These elements have been identified by many as ultraconserved regions (UCR) and conserved non-coding elements (CNE) (Bejerano et al., 2004, Meisler, 2001). There are many tools for HCNEs detection and visualization (Engstrom et al., 2008). It has been shown by previous studies that these elements tend to cluster around genes encoding transcription factors and developmental genes. These clusters could span far away the gene loci (2 Mb) (Akalin et al., 2009). Several HCNEs act as enhancers and regulate the expression of target gene. The function of these elements is detected using the enhancer-trapping technique (Engstrom et al., 2007). In this technique a transgenic construct containing a mini promoter, which is unable to transcribe in absence of enhancer and a reporter gene is randomly inserted into the genomic location. The activity of the enhancer is confirmed by the increased expression of the reporter gene.

2.3.2 CpG islands

CpG islands (CGIs) refers to regions with high occurence of *CpG*s i.e., a "*CG*" dinucleotide with a phosphodiester "p" bond connected to the cytosine and guanine base pairs (Deaton and Bird, 2011). Typically, a CpG island is defined as a region in the genome that has GC content more than 50%, length greater than 200 base pairs and observed-to-expected ratio of *CpG*s greater than 0.6. The observed-to-expected CpG ratio is given by

$$\frac{observed\ CpG\ counts}{expected\ CpG\ counts} = \frac{\#CpG * N}{\#C * \#G}$$

were $\#CpG, \#C, \#G$ denote the numbers of *CpG* dinucleotides, cytosine (*C*) and guanine (*G*) nucleotides in a DNA sequence of length N, respectively.

The length of CpG islands varies between 300-3000 base pairs in mammalian genomes, and these regions are frequently found near TSSs and are known to influence gene transcription (Antequera, 2003). More than 50% of the promoter regions in the human genome have high CpG content compared to the rest of the genome (Ioshikhes and Zhang, 2000). It is also known that CpG sites are associated with histone modifications (Esteller, 2006).

2.4 Histone modifications

Eukaryotic chromatin consists of DNA coiled with histone proteins, and the accessibility of genes to regulatory proteins in the chromatine structure is a crucial regulatory factor in gene expression. The histone proteins belong to specialized class of proteins that help in the compaction of DNA and manage the cell compartmentalization. There are two main types of histones, namely, core histones (H2A, H2B, H3 and H4) and linker histones (H1 and H5) (Bartova et al., 2008). Two of each of the core histones forms an octamer around which the DNA loops (nucleosome). A nucleosome is a 147 base pair long segment of DNA wrapped around the histone octamer linked by 80 base pair linker DNA to the next nucleosome (Figure 3). Linker histones (H1 or H5) sit at the base of the nucleosome near the binding to the linker DNA. The nucleosome is the most basic unit of chromatin, which further condensed to 30 nm long fiber and finally results in to a tight-coiled 250 nm long chromatid (Marino-Ramirez et al., 2005).

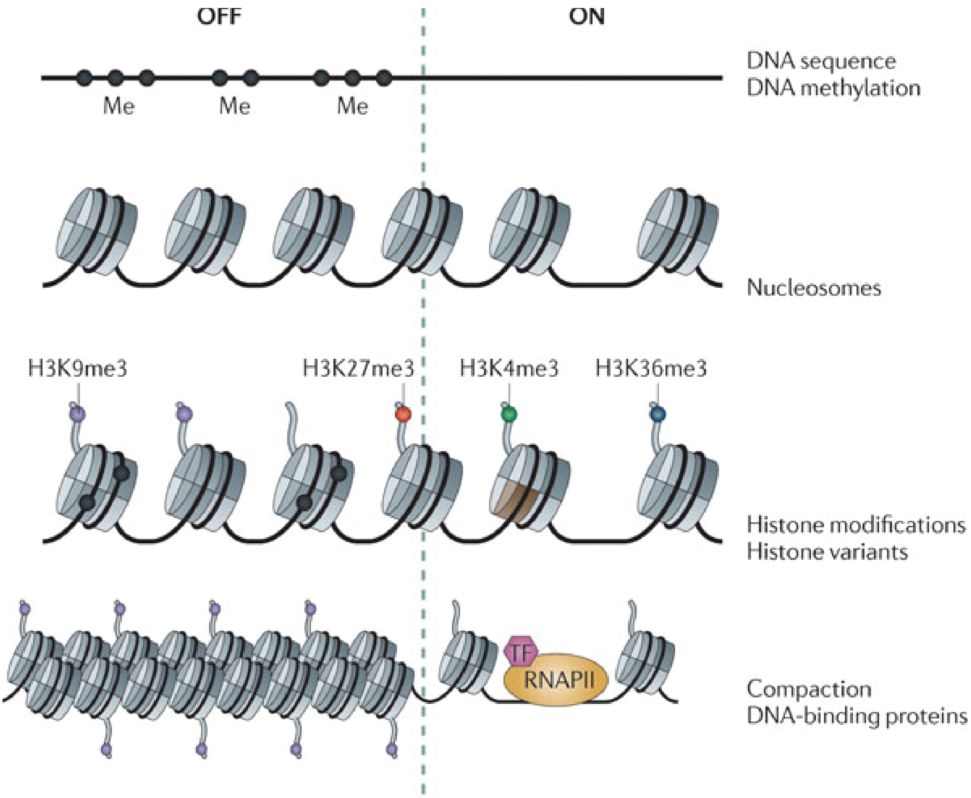


Figure 3. Levels of chromatin organization. DNA is methylated (on cytosine bases) in a context specific manner and packaged into nucleosomes with distinctive histone composition and modifications. This figure is adapted from (Zhou et al., 2011).

Along with DNA compaction, histone proteins also play a very important role in gene regulation through different types of post-translational modifications widely known as "histone modifications" (Bartova et al., 2008). The long protruding N terminal amino acid tail and globular domains of histone proteins serve as sites for covalent modifications. There are more than 60 different residues for modifications including methylation, acetylation, ubiquitination, ADP-ribosylation and simulation of lysines, phosphorylation of serine, methylation of arginine, etc., (Kebede et al., 2015).

Histone modifications have been classified into two classes, namely, euchromatin and heterochromatin modifications, depending on the packing form of chromatin where the modification occurs. Euchromatin is the lightly packed form of chromatin that is usually rich in genes and euchromatin modifications (e.g., acetylation of histone 3 and 4) are typically related to active transcription.. Heterochromatin is the tightly packed form of chromatin and its modifications (e.g., methylation of lysine 9 and 27 of histone 3) are typically associated with repressed transcription (Kim et al., 2003). The combinatorial patterns of different modifications result in the formation of the so-called "histone code", which is essential for maintaining the higher-order chromatin structure.

Recent advances in assays based on high throughput sequencing technologies (like ChIP-Seq) have revolutionized the study of histone modifications by allowing access to genome wide maps of various modifications under specific cell conditions (Wang et al., 2008). It has been shown that several combinatorial patterns of histone modifications exhibit high enrichment near specific functional domains (Ernst et al., 2011). Several histone modifications have been characterized in terms of their influence on regulatory regions in the genome (Table 1).

Histone Modification	Associated Functional region in genome
H3K4me1, 2, 3	Active / poised Enhancers
H3K27ac	Active Enhancers
H3K4me2, 3	Active Transcription
H3K27me3	Repressive Transcription
H3K36me3, H4K20me1	Transcription Elongation

Table 1. Examples of histone modifications and functions associated.

Many studies have focused on analysis of combinatorial patterns of histone modifications and chromatin remodeling to annotate the different functional domains of the genome (Ernst and Kellis, 2010). A recent study (Ernst et al., 2011) exploited chromatin profiling of nine histone modifications across nine cell lines in order to characterize the human genome into fifteen different functional states, namely, strong and weak enhancers; active, poised and repressed promoters; putative insulators; transcribed regions; and large-scale repressed and inactive domains. This state map represents the highly dynamic chromatin landscape responsible for silencing/activating specific functional domains across respective cell lines.

2.5 Evolution of genes and their mechanism of duplication

In this section, I discuss the mechanisms by which genes duplicate and their role in evolution. By necessity, I shall be brief and only touch upon concepts of immediate relevance to this thesis.

2.5.1 Gene duplication and retention

Gene duplication also known as chromosomal duplication or gene amplification refers to the duplication of a DNA region that contains genes. It is a very important mechanism in evolution and is responsible for generating new genetic material (Taylor and Raes, 2004, Harris and Hofmann, 2015). There are two basic type of gene duplication: small scale/tandem duplication (SSD) and whole genome duplication (WGD) (Leister, 2004).

Small scale/tandem duplication (SSD) involves doubling of a section of chromosome, which may result in a functional replicate. There are several mechanisms that cause SSD such as retro-transposition or unequal crossing over during meisos (Walker et al., 1995, Ohta, 1990). These duplications vary in size and arrangement. The duplication can range from few base pairs to several megabases. Further, the duplicate segment can be adjacent to the original segment (tandem duplications) or scattered within the same chromosome (intra-chromosomal) or even, across the genome (inter-chromosomal) (Trask et al., 1998, Ji et al., 1999).

Whole genome duplication (WGD) occurs when the entire genome of the organism is duplicated resulting in polyploidy (more than two paired sets of chromosomes). This is an important phenomenon responsible for complexity of organism during evolution. It has been associated with adaptive radiations of species. It is believed that two rounds of whole genome duplication have occurred in vertebrates (Dehal and Boore, 2005, Berthelot et al., 2014). A duplicated gene (as well as the original gene) can have one of several possible fates. Either of the copies of the gene might become inactive because of accumulation of mutations (nonfunctionalization), or, both copies can acquire different mutations which leads them to perform different roles from their ancestral gene (subfunctionalization) (Hurles, 2004). In some cases, one copy will acquire a completely new function (neofunctionalization) while the other retains its original function (Teshima and Innan, 2008).

Detecting gene duplication is not straightforward because of complications like existence of isoforms, domain shuffling and errors in annonated databases (Li et al., 2003a). Further, it has been shown that different functional classes of genes are enriched by different mechanisms of duplication (Woods et al., 2013). Broadly speaking, there are two classes of methods for detecting gene duplication, and if possible, the mechanism of duplication (Durand and Hoberman, 2006). The first class of methods relies on sequence homology to identify gene homologs and calculate the selection pressure on the gene by comparison to its orthologs. In contrast, map-based methods leverage the ordering of genes on sequenced genomes in order to find instances of microsynteny, i.e., physical co-localization of the genes indicating conserved gene ordering across species.

A key quantity of interest in evolutionary biology is the ratio of non-synonymous mutation rate (K_a) to the synonymous mutation rate (K_s) for a protein-coding gene, which is used to measure the selection pressure on the gene (Kryazhimskiy and Plotkin, 2008). A synonymous (resp., non-synonymous) mutation is a change in the sequence of the gene that does not change (resp., changes) the protein produced by that gene. The underlying hypothesis is that genes under negative (purifying) selection pressure would have fewer non-synonymous mutations compared to synonymous mutations (and hence K_a/K_s ratio less than one) since they possibly code for an important protein, and vice versa. In general, K_a/K_s ratio greater (resp., less) than one implies positive (resp., negative) selection pressure while K_a/K_s ratio close to one is indicative of neutral selection on the gene. There are several tools available to calculate K_a/K_s ratios (Zhang et al., 2006, Yang, 2007).

Though K_a/K_s ratios have been extensively used for inferring selection pressure (and, in turn, conservation) on genes, there are some limitations. For example, this method cannot be used for investigating the selection pressure on non-coding regions of the genome which as previously discussed include several regulatory elements exhibiting high degree of conservation. Also given the heterogeneity of selection pressure within a gene, it is sometimes hard to interpret the result. For example, the K_a/K_s ratio can be close to one due to the positive and negative selection occurring at different loci within the gene body.

2.6 Nuclear Receptors

Nuclear receptors form one of the largest structural classes of transcription factor proteins (Robinson-Rechavi et al., 2003). Along with co-activators/repressors, they control the expression of multiple target genes and affect nearly all physiological processes (Carlberg and Seuter, 2010, Tachibana et al., 2005). Activation by nuclear receptors depends on presence of small molecules (known as ligands), which bind to the nuclear receptor and in turn change the regulatory behaviour of the nuclear receptor. There are 48 nuclear receptor genes in the human genome. For several of them, no ligand is yet known and these are termed "nuclear orphan receptors". Unsurprisingly, nuclear receptors have attracted immense attention in recent decades in the field of pharmacology (Evans, 2005) and the identification of ligands for orphan receptors often allows discovery of new drugs targetting the newly discovered hormone response system of the receptor (Benoit *et al.*, 2006).

In this chapter, I provide a brief overview of the nuclear receptor family in terms of its structural properties, mechanism of action and known classifications based on sequence homology.

2.6.1 Structural organization of nuclear receptor genes

Nuclear receptors are characterized by the presence of two-zinc fingers in the DNA-binding domain (DBD). Each finger contains four cysteine residues coordinating one zinc ion (Olefsky, 2001). All nuclear receptors share a common modular structure that comprises five or six homologous domains (Figure 4).

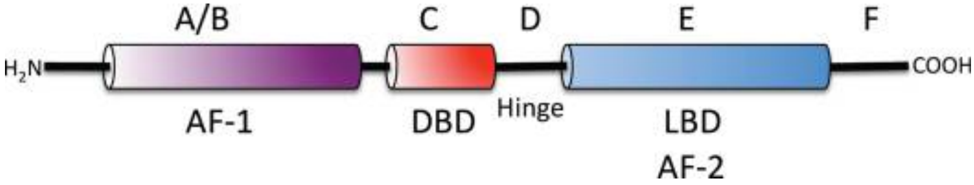


Figure 4. Domain organization of the nuclear receptors. Structural organization of nuclear receptors showing regions (A-F): N-terminal (left), Activation function domain 1 (AF-1), DNA-binding domain (DBD, Ligand binding domain (LBD), Activation function domain 2 (AF-2), C-terminal (right). The figure is adapted from (Olefsky, 2001)

The N terminal domain consists of poorly conserved A/B region that contains activation function domain 1 (AF-1). This region varies highly among nuclear receptors and its structure is yet not well characterized. Next to the A/B region is the highly conserved DNA binding domain in the C region of nuclear receptor. All nuclear receptors except *DAX1* and *SHP* have this domain (Guo et al., 1996). This domain is responsible for binding of nuclear recetor to the hormone response element (HRE) on DNA. The D region (hinge region) contains a poorly conserved domain that acts as hinge between DBD and ligand binding domain (LBD). The ligand-binding domain is the next highly conserved domain other than DBD. This domain is functionally complex. The LBD (E region) consist of ligand binding pocket, dimerization region, co-regulator binding region and activation function 2 domain (AF-2) (Moras and Gronemeyer, 1998). This region mediates ligand dependent trans-activation and co-activator/repressor recruitment. The crystal structure of LBD is well studied for most of the nuclear hormone receptors (Li et al., 2003b, Moras et al., 2015). The last region of nuclear receptor is F region that comprise C terminal domain. The structure of this region is also yet not well understood (Germain et al., 2006).

2.6.2 Mechanism of action

There are many crucial steps in the regulation of expression level of target gene by a nuclear receptor. These include binding of specific ligand to the receptor, involvement/recruitment of several co-regulators and recognition of specific binding sites. Figure 5 presents the classic example describing the mechanism of action of a nuclear hormone receptor, in this case, steroid hormone receptor. Initially, a heat shock protein in the cytoplasm is bound to the nuclear receptor. The binding of the hormone to the ligand binding domain of the nuclear receptor releases the heat shock protein and induces conformational change in the receptor, which further results in homo-dimerization and translocation (active transport) of the receptor into the nucleus. It should be noted, however, that some nuclear receptors are located in the nucleus also in the absence of ligand. Inside the nucleus, the receptor binds to specific sequence on DNA called hormone response element (HRE) and recruits other activator proteins to the complex. The binding of nuclear hormone receptor along with its co-regulatory complex finally results in up/down regulation of target genes.

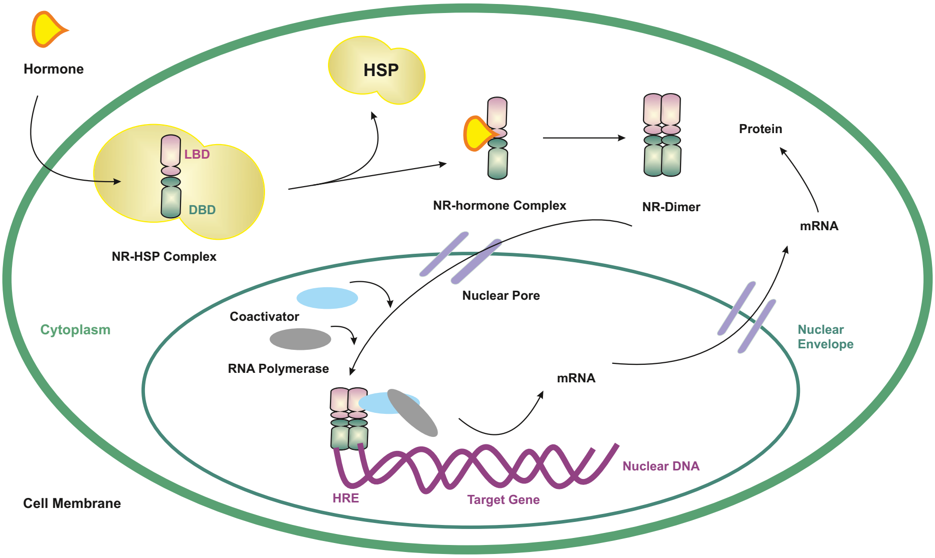


Figure 5. Detailed mechanism of action of nuclear receptors. In the absence of ligand NR is located in the cytosol. Hormone binding (shown in yellow) to the receptor releases the heat shock protein (HSP) and leads to the formation of NR-hormone complex, dimerization and translocation to the nucleus. In the nucleus NR binds to specific sequence in the DNA called Hormone response element (HRE, shown in purple).

2.6.3 Nuclear orphan receptors

Nuclear orphan receptors do not bind to any known ligand, or, in other words, their endogenous ligand is missing. The cognate ligands for nuclear orphan receptors are either missing or physiological functions are not well characterized (Baek and Kim, 2014, Zhao and Bruemmer, 2010, Pols et al., 2007). In the human genome, around half of the nuclear receptor genes code for nuclear orphan receptors. The nomenclature of this class of nuclear receptors has been a topic of debate, as the term “receptor” implies a physiological ligand. Moreover once a ligand has been identified for an orphan receptor, the receptor is no longer classified as “orphan”. The RXRs and PPARs were initially identified as orphan receptors, but now it is clear that they are ligand-dependent receptors (Wayman et al., 2002). The ligand-binding pockets of some receptors (FXRs, LXRs, CAR, PXR) are larger and they bind to diverse range of compounds with lower affinity. It has also been reported that some compounds were found in the binding pocket of HNF-4, RORs, and SF-1. Due to the lack of physical interaction between receptors and compounds, these receptors are still classified as nuclear orphan receptors. The apparent lack of ligand induced regulation of orphan receptors points to the presence of alternative mechanism of gene regulation or an undiscovered ligand – both of which have interesting pharmacological implications. For example, studies have shown that though the ligand binding domain of the nuclear orphan receptor NURR1 folds in the same manner as hormone receptors, it lacks the cavity for ligand binding (Wayman et al., 2002). Hydrophobic amino acid side chains fill the ligand-binding domain of NURR1. This orphan nuclear receptor thus follows an alternative mechanism for controlling the expression of its target genes rather than ligand binding. We note that the PPAR and NR4A groups of nuclear receptors (including NURR1) have emerged as potential pharmacological targets for obesity, diabetes and Parkinson's disease respectively (Eells et al., 2012).

2.6.4 Existing classifications

Nuclear receptor genes have been classified on the basis of sequence similarity as well as their functional roles.

Homology based classification

Based on their sequence similarity, nuclear receptor genes have been classified into seven subfamilies.

Subfamily	**Description**	**Number of Genes**
Subfamily 1	Thyroid Hormone Receptor-like	19
Subfamily 2	Retinoid X Receptor-	12

	like	
Subfamily 3	Estrogen Receptor like	9
Subfamily 4	Nerve Growth Factor IB-like	3
Subfamily 5	Streroidogeneic Factor-1 like	2
Subfamily 6	Germ Cell Nuclear Factor-like	1
Subfamily 0	Miscellaneous	2

Table 2. The homology based classification of the nuclear receptors. Each column represents subfamily type, description and number of genes respectively.

Each family is represented by presence of their duplicated paralogs (Robinson-Rechavi et al., 2001). The last subfamily comprises of *DAX-1* and *SHP,* which are different from other nuclear receptors in terms of both structure and function. They lack the characteristic DBD (Zanaria et al., 1994, Seol et al., 1996). The genes NR0B1 and NR0B2 encode the *DAX-1* and *SHP* proteins respectively. *DAX-1* controls the activity of genes that form hormone-producing tissues. The function of SHP is to repress other nuclear receptors.

Function-based classification

Nuclear hormone receptors have been further classified into three different subtypes on the basis of their cellular location, dimerization and DNA binding properties (see Figure 6). For type I nuclear hormone receptors, the ligand binds to the receptor in cytosol resulting in dissociation of heat shock protein and homo-dimerization of receptor molecules. Then, the complex consisting of the dimerized receptor and the ligand translocates into the cell nucleus where it binds to its cognate hormone response element (HRE; inverted repeat) of the target promoter and modulates the expression level of the corresponding gene (Figure 6). The members of Subfamily 3 (Estrogen-receptor like) are classical examples of this subtype (Li and Al-Azzawi, 2009, Bjornstrom and Sjoberg, 2005).

Type II nuclear hormone receptors are retained in the nucleus regardless of their ligand binding status. In the absence of the ligand, they are often present in the complex with co-repressors (Figure 6). Upon ligand binding the co-repressors dissociate and the type II nuclear receptor form a hetero-dimer complex with retinoic X receptor (RXR), which also acts as a co-activator. The members of Subfamily 2 (Thyroid, retinoid receptors) are the examples of this mechanism (Berrodin et al., 1992, Minucci et al., 1997). Type III Nuclear hormone receptors exert a similar mode of action as type I nuclear receptors except that they bind to direct repeats in HREs instead of inverted repeats (Figure 6).

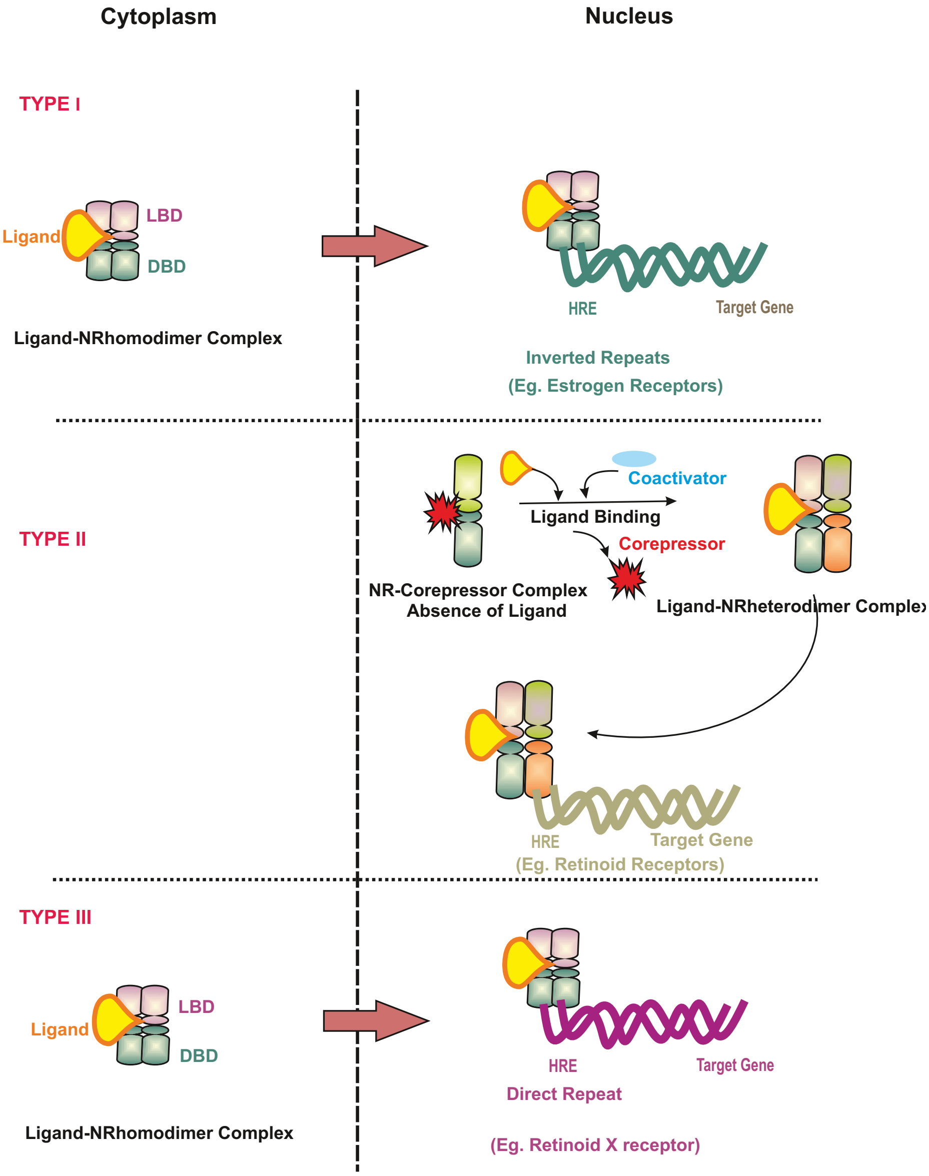


Figure 6. Different modes (Type I, Type II and Type III) of action of the nuclear hormone receptor. The central line reprents the cell localization (left side is cytoplasm and right is nucleus). Arrow (shown in red color inbetween center vertical line) shows the translocation of nuclear hormone receptor complex.

2.6.5 Evolution of nuclear receptors

Nuclear receptors are found throughout the metazoan lineage. Both nuclear hormone and orphan receptors were present at the diversification of protostomes from the deuterostomes. Their evolutionary history is complex: there are two nuclear receptors in sponges and twenty-one in the phylum *Cnidaria*. Some members of the Cnidarian receptors have ligand binding and DNA binding domains of different nuclear receptors in one. This clearly shows that nuclear receptors originate from a common ancestor and that they diversified during the course of metazoan evolution (Laudet et al., 1992). How nuclear receptors evolved has been a matter of debate since long time, specifically it was debated which one of the two (nuclear hormone or orphan) receptors that came first. Bridgham and coworkers have argued that orphan nuclear receptors originate from ligand-dependent receptors (Bridgham et al., 2010). They suggested that evolutionary tinkering has created the diversity in modern receptors. Subtle modification in the internal cavity and various mutations have stabilized the evolution of ligand independent receptor across evolution. Nuclear receptors are still evolving, as there are other human pseudogenes (ERR-r) reported along with their paralogs members. These might be future functional nuclear receptor members waiting for selection pressure to incorporate them into complex regulatory networks (Zhang et al., 2004).

3 Results and Discussion

We can't be sure about this, but we've analyzed genes on several of your chromosomes, and it's hard to avoid the conclusion:

At some point, your parents had sex.

- Randall Munroe, "Genetic Analysis", XKCD[1]

"I am a good statistician. I can get you any p-value you want!"

- Susan Holmes[2], Ascona (2015)

In this chapter I first describe the work leading to a new classification of nuclear receptor genes characterized by the *cis*-regulatory environment of the genes using the GRB model (Paper I). Following this, I elucidate the evolutionary mechanisms of nuclear receptor genes within the context of our proposed classification (Papers II-III).

3.1 Transcriptional regulatory features of nuclear receptor genes

Our main hypothesis was that the diverse functional roles of nuclear receptors are linked to the differences in the transcriptional regulation mechanisms that control the expression of the nuclear receptor genes. In working towards this hypothesis, we found that nuclear receptors can be divided into two main clusters on the basis of their *cis*-regulatory enviornment (Paper I). The work is based on the GRB model (Kikuta et al., 2007) and *cis*-regulatory elements of the nuclear receptor genes.

The first question to answer was whether nuclear receptor genes have properties (HCNEs, CpG islands) of targets of GRB-mediated regulation. Though the GRB model is well established, it has some limitations. First the length and boundaries of

[1] https://xkcd.com/830/

[2] http://www.statweb.stanford.edu/~susan/

GRBs are not known precisely. HCNEs are known to act as distant enhancers and there are studies demonstrating that enhancers can be located as far as 1 or 2 Mb away from their target gene loci (Naranjo et al., 2010, Lettice et al., 2003, Symmons and Spitz, 2013). Therefore, we chose to analyze 2Mb region around each nuclear receptor gene locus and found that the genes naturally separate into two main clusters (Figure 2, Paper I). Cluster 1 (25 genes) comprised of genes that are targets of the GRB model. These genes have high densities of HCNEs around their gene loci and also contain large and often multiple CpG islands. On the other hand, the genes in Cluster 2 (23 genes) have neither HCNEs nor long CpG islands.

Moreover, we established that there are three cases in which more than one nuclear receptor gene is present within one GRB locus (multi-gene GRB). We identified a total of 3 multi-gene GRBs comprising 9 genes. The genes in the same GRB are in close proximity to each other on the same chromosome, hence fall in the one respective GRB. The next challenge was to find which of the nuclear receptor genes was the target of the GRB. We checked the proximity of HCNEs to the closest nuclear receptor gene loci. However, as enhancers can act from a long-range distance, additional considerations (promoter architecture) were included. With the help of publically available histone modifications data we tried to explore the promoter architecture of each gene. For example, *THRA, RARA, EAR1* share the same GRB locus. In this case, all of these genes share proximity of HCNEs around their gene loci, but *RARA* and *EAR1* have bivalent promoter mark while *THRA* does not. Therefore we assigned *RARA* and *EAR1* as putative targets of the GRB locus.

To gain an in-depth understanding of transcription regulatory features of nuclear receptor genes, we explored their *cis*-regulatory environment in the context of the obtained clustering. We used publicly available datasets (ENCODE) of gene expression (RNAseq) and histone modifications (ChIPseq) that are known to be associated to the promoter and enhancer regions of the gene. We found that most of the genes in Cluster 1 are expressed in an embryonic stem cell line (H1hesc) while the genes in Cluster 2 are either not expressed in any of the five cell lines or highly expressed in all for which datasets were available from ENCODE project (Table S3, Paper I). This indicates that GRB target nuclear receptor genes play role in developmental functions while GRB non-taregts are more likely to have either ubiquitous or tissue-specific roles.

We further validated our expression analysis by using histone modifications known to be associated with transcriptional activity (H3K4me3) and elongation (H3K36me3) and found that genes that are highly expressed have higher enrichment of both these histone modifications in comparison to low-expressed genes in their respective cell lines. Moreover, we observed that the genes in Cluster 1 have a significantly higher enrichment of enhancer-associated histone modification (H3K4me1) in comparison to the genes in Cluster 2 (Figure 4, Paper I). This is in accord with the fact that HCNEs have been shown to be functioning as distal enhancers in several studies (Navratilova et al., 2009, Engstrom et al., 2007, Pennacchio et al., 2007). Moreover, this finding provided additional evidence that the genes in Cluster 1 are putative targets of long-range gene regulation.

Developmentally regulated genes are known to have bivalent domains in their promoter regions, showing enrichment of both transcription activating (H3K4me3) and repressing (H3K27me3) histone modifications at the same loci (Sachs et al., 2013). It is believed that bivalent domain enrichment allows genes to be turned on and off during different stages of development (Voigt et al., 2013). We found that Cluster 1 genes have bivalent domains in their promoter region while the genes in Cluster 2 do not (Figure 5, Paper I). Moreover, we noticed that GRB target nuclear receptor genes have higher enrichment of H3K27me3 around their promoter regions, especially when not expressed in the H1hesc cell line. However the most interesting observation was in the case of 5 genes (*NUR77, LRH-1, EAR1, RORB* and *ESRRG*) that retained repression mark while being actively transcribed (in the H1hesc cell line). This finidng might indicate that these genes are in transition state from high to low expressed with no repressive to high repressive mark respectively or vice versa. Detailed analyses of the promoter region of *EAR1* and *RORB* demonstrated that enrichment of H3K27me3 starts slightly downstream of TSS and extend in to the first intron of the gene. The functional implication such an arrangement is not known at present and further time-series experiments are required to understand the complete dynamics. On the other hand, genes in Cluster 2 do not have this mark (H3K27me3) regardless of their expression state. This leads credence to our hypothesis that genes in Cluster 1 are developmentally regulated genes that require long-range regulatory elements for correct expression, while genes in Cluster 2 are house-keeping or tissue-specific genes.

To have a clearer picture of landscape of transcription regulation in nuclear receptor genes, we analyzed enrichment of combinatorial patterns of histone modifications in the H1hesc cell line. We recovered our original clustering, providing independent evidence that our analysis correctly identifies the nuclear receptors that are targets of long-range regulation.

In summary, we identified nuclear receptor genes that are targets of long-range gene regulation using the GRB model. The *cis*-regulatory environment of these genes is characterized by a high span of HCNEs, long CpG islands, enhancer and bivalent promoter marks. Functionally, these genes are developmentally regulated genes which require above-mentioned *cis*-regulatory elements in order to be turned on and off at different time points during development. In contrast, nuclear receptors that we annotate as GRB non-targets are either housekeeping genes or tissue-specific genes, which do not have (or need) a complex *cis*-regulatory environment.

Our work leads to a number of follow-up questions about the mechanism of action, evolutionary history and functional roles of nuclear receptors genes. For example, several nuclear receptors are still classified as "orphan", i.e., there is no known endogenous ligand. One can ask whether our functional classification of nuclear receptors depending on their *cis*-regulatory environment provides any insight into the mechanism of action of these (as yet) orphan receptors. From an evolutionary perspective, several authors have reconstructed the phylogenetic tree of this gene family (Zhao et al., 2015, Bertrand et al., 2004). However, understanding the mechanism of duplication can help us to know what role evolution has played in order to maintain current function and expression of nuclear receptor genes.

3.2 Evolution of nuclear receptor genes

In Paper I, we showed that nuclear receptor genes have different transcription regulatory environments, which help explain their functional divergence. In follow-up work, our aim was to identify how evolution of nuclear receptor genes has affected their functionality, and possibly finds some links between our proposed classification of nuclear receptor genes and their evolutionary history.

A key challenge when studying evolution is the absence of data from ancestral species, so most of evolutionary genomics is necessarily built upon phylogenetic analysis of known species. During the course of evolutionary history, a genome undergoes several large-scale and more localized events caused by different mechanisms and leading to different outcomes and ultimately speciation. Even disregarding the changes in the non-coding regions and large-scale chromosomal re-arrangements of the genome – gene loss, neo(/sub)-functionalization of duplicated genes, etc., are some of the possibilities. This is further complicated by presence of isoforms and incomplete sequencing information. Nonetheless, modern evolutionary biology is a mature field with well-founded methods that broadly speaking, focus on one of these two approaches: inference of phylogenetic trees based on sequence homology and identification of localized changes based on conservation of gene order across species (microsynteny).

The main hypothesis in the GRB model is that the target gene needs its *cis*-regulatory environment in order to achieve the expression patterns needed for its developmental functions. Consequently, the target gene *as well as its cis-regulatory elements* will be under purifying (negative) selection pressure in order to maintain its functions (Kikuta et al., 2007).

If a developmental gene has already acquired its function (and the *cis*-regulatory elements needed for it) in the ancestral species, i.e., it is already a GRB target in the ancestral species, it is less likely to exhibit tandem duplicationsm, and microsynteny is more likely to be preserved across species due to strong purifying selection on the gene as well as its regulatory environment.

In constrast, whole genome duplication provides a mechanism by which the entire GRB locus present in the ancestral species may be duplicated. During this process, the bystander genes may migrate outside the duplicate GRB locus, especially if there is no selection pressure on their intronic region due to a regulatory element (e.g., an enhancer) of the target gene being present there (Kikuta et al., 2007). Subsequent to duplication – the newly formed GRB locus may become inactive, acquire new function (neofunctionalization) or the original target gene and its duplicated copy (along with their respective regulatory environments) might perform different functions, which were originally performed by the ancestral target gene (subfunctionalization). Therefore, taking into account the long-range regulatory environment provides crucial insight when inferring the mechanism of duplication and consequently, understanding the evolutionary history of the developmental gene in question.

In Papers II-III, we use the GRB model and our proposed classification in Paper I to study the evolutionary history of nuclear receptor genes. Our main hypothesis was that many nuclear receptors have already been recruited in the ancestral genome, and accordingly, have maintained the *cis*-regulatory environment needed for their function through evolution. We used microsynteny and sequence homology to investigate the mechanism of duplication of nuclear receptor genes in the context of their known functions and transcriptional regulatory mechanisms.

Following standard methodology, we used the ratio of non-synonymous to synonymous mutation rates (K_a/K_s) to study the selection pressure on nuclear receptor genes. We found that *all* the genes (except *DAX1*) proposed as putative targets of the GRB model in Paper I show strong purifying selection (Wilcoxon signed-rank test, median=1.0, one-tail, p=2.7181e-05). On the other hand, several nuclear receptors genes proposed as non-targets of the GRB model in Paper I nuclear receptors are under positive selection pressure which makes them available for new functions across evolution (see Paper III, Figure 3). This lends credence to our hypothesis that genes in Cluster I (GRB targets) had their function determined in the ancestral species, and accordingly, have been under strong negative selection pressure during evolution.

We next investigated the degree of conservation of the genomic neighborhood of each nuclear receptor gene by considering the homologous genes in mouse and zebrafish genomes. Using methods developed in Paper II, we calculated CGN score (conservation of genomic neighbourhood) for each nuclear receptor gene which captures the proportion of conserved orthologs between human and mouse (respectively zebrafish) in a 2Mb region centered around the gene (De et al., 2009). It has previously been shown that the genomic neighborhood of genes is correlated to their functional diversion (De et al., 2009) – genes with high degree of conserved genomic neighbourhood ($CGN > 0.5$) are more likely to perform similar function in the two species in contrast to genes with low conservation.

We found that all of the nuclear receptor genes (except *HNF4G, TFCOUP2* and *TFCOUP1*) have high conserved genomic neighborhood between human and mouse (Paper III, Figure 4). This observation is not surprising given the high similarity between human and mouse genomes (Guenet, 2005). Indeed, a closer look reveals that the neighbourhoods of *HNF4G*, *TFCOUP1* and *TFCOUP2* are gene deserts (i.e., these have few genes in their close neighbourhood in the human genome), which explains the low CGN scores.

We repeated the same analysis for human:zebrafish knowing that zebrafish is at much greater evolutionary distance to human (~450Myr) than mouse. Moreover, the zebrafish genome has undergone one additional round of whole genome duplication (3R) compared to human (2R), and as a result, several nuclear receptor genes have two orthologs in zebrafish. The conservation of genomic neighbourhood is strikingly different – as expected, most of the nuclear genes do not have a high conservation of their genomic neighbourhood in zebrafish. Nonetheless, we find seven nuclear receptor genes (five targets and two non-targets) for which one of the orthologs in Zebrafish shows high conservation of neighbourhood ($CGN > 0.5$). Further, the genes in Clusters 1 (GRB targets) show much higher conservation compared to the genes in Cluster 2 (non-targets) (Paper III, Supplementary Figure).

The original study (De et al., 2009) used a cut-off score of 0.5 when comparing conserved neighbourhoods between human and chimpanzee genome. This cut-off is not well suited for identifying conserved gene neighbourhoods in our case given the much greater evolutionary distance between the human and zebrafish genomes. In order to address this, we proposed a simple (and admittedly simplistic) model that assumes each neighbouring gene within the 2Mb region is conserved or lost independently of other genes in the neighbourhood with probability p_c, which is identical (the second assumption) for all the nuclear receptor genes. Under these assumptions, the maximum likelihood estimator (MLE) for the conservation probability is given by:

$$p_{\{c\}} = \frac{\sum_i n_c^{(i)}}{\sum_i n_t^{(i)}}$$

where $n_t^{(i)}$ and $n_c^{(i)}$ denote for the $i^{\{th\}}$ nuclear receptor gene, the total number of neighbouring genes (in the human genome) and the number of neighbouring genes having conserved orthologs (in a 2Mb region in the zebrafish genome), respectively. For the nuclear receptor genes, the probability distribution over the number of conserved neighbours is given by a Binomial distribution, and, one can find if a nuclear receptor gene has conserved genomic neighbourhood at a chosen significance level (Paper III, Figure 3).

It is clear that the above analysis makes strong simplifying assumptions, namely, (1) conservation of each neighbouring gene is independent of other genes, and, (2) identical conservation probabilities for the neighbouring genes of *all* nuclear receptors. One can relax the independence assumption by taking into account the length and proximity of the neighbouring genes, presence of regulatory elements in their intronic regions and known function. It is also possible to correct the assumption of identical conservation probability for the neighbourhood of each nuclear receptor gene using a Bayesian approach with suitable prior and model, e.g., similar conservation probability for nuclear receptors belonging to the same structural class. Further, the conservation of gene neighbourhoods in other gene families can also provide valuable information.

Nonetheless, a more sophisticated model would have to accommodate the impact of known biological disruptions such as large-scale chromosomal rearrangements, additional round of whole genome duplication, etc. We believe this would constitute a thesis in its own right and defer this to future work.

4 Future Perspective

"Science never solves a problem without creating ten more"

- George Bernard Shaw

This thesis focuses mainly on understanding the impact of regulation and evolution of nuclear receptors on their current function. We propose a new classification to this gene family on the basis of their different transcriptional regulation. We come up with many interesting observations that lead to further investigations and experimental confirmations. For instance, precise function (enhancer/repressor) of HCNEs can be detected by using enhancer-trapping experiments.

We found that genes in cluster 1 have significantly higher enrichment of H3K4me1 in comparison to cluster 2. This further support our HCNEs based clusters. In order to capture functionally active regulatory regions it would be interesting to study the time point based correlation between HCNEs and the histone modification known to be associated with active enhancer (H3k27ac) regions.

5 References

AKALIN, A., FREDMAN, D., ARNER, E., DONG, X., BRYNE, J. C., SUZUKI, H., DAUB, C. O., HAYASHIZAKI, Y. & LENHARD, B. 2009. Transcriptional features of genomic regulatory blocks. *Genome Biol,* 10**,** R38.

ANDERSSON, R. 2015. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays,* 37**,** 314-23.

ANTEQUERA, F. 2003. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci,* 60**,** 1647-58.

BAEK, S. H. & KIM, K. I. 2014. Emerging roles of orphan nuclear receptors in cancer. *Annu Rev Physiol,* 76**,** 177-95.

BARTOVA, E., KREJCI, J., HARNICAROVA, A., GALIOVA, G. & KOZUBEK, S. 2008. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem,* 56**,** 711-21.

BEJERANO, G., PHEASANT, M., MAKUNIN, I., STEPHEN, S., KENT, W. J., MATTICK, J. S. & HAUSSLER, D. 2004. Ultraconserved elements in the human genome. *Science,* 304**,** 1321-5.

BERRODIN, T. J., MARKS, M. S., OZATO, K., LINNEY, E. & LAZAR, M. A. 1992. Heterodimerization among thyroid hormone receptor, retinoic acid receptor, retinoid X receptor, chicken ovalbumin upstream promoter transcription factor, and an endogenous liver protein. *Mol Endocrinol,* 6**,** 1468-78.

BERTHELOT, C., BRUNET, F., CHALOPIN, D., JUANCHICH, A., BERNARD, M., NOEL, B., BENTO, P., DA SILVA, C., LABADIE, K., ALBERTI, A., AURY, J. M., LOUIS, A., DEHAIS, P., BARDOU, P., MONTFORT, J., KLOPP, C., CABAU, C., GASPIN, C., THORGAARD, G. H., BOUSSAHA, M., QUILLET, E., GUYOMARD, R., GALIANA, D., BOBE, J., VOLFF, J. N., GENET, C., WINCKER, P., JAILLON, O., ROEST CROLLIUS, H. & GUIGUEN, Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun,* 5**,** 3657.

BERTRAND, S., BRUNET, F. G., ESCRIVA, H., PARMENTIER, G., LAUDET, V. & ROBINSON-RECHAVI, M. 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. *Mol Biol Evol,* 21**,** 1923-37.

BJORNSTROM, L. & SJOBERG, M. 2005. Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes. *Mol Endocrinol,* 19**,** 833-42.

BRIDGHAM, J. T., EICK, G. N., LARROUX, C., DESHPANDE, K., HARMS, M. J., GAUTHIER, M. E., ORTLUND, E. A., DEGNAN, B. M. & THORNTON, J. W. 2010. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol,* 8.

BURRIS, T. P., BUSBY, S. A. & GRIFFIN, P. R. 2012. Targeting orphan nuclear receptors for treatment of metabolic diseases and autoimmunity. *Chemistry & biology,* 19**,** 51-59.

CARLBERG, C. & SEUTER, S. 2010. Dynamics of nuclear receptor target gene regulation. *Chromosoma,* 119**,** 479-84.

COULON, A., CHOW, C. C., SINGER, R. H. & LARSON, D. R. 2013. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat Rev Genet,* 14**,** 572-84.

DE, S., TEICHMANN, S. A. & BABU, M. M. 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Res,* 19**,** 785-94.

DEATON, A. M. & BIRD, A. 2011. CpG islands and the regulation of transcription. *Genes Dev,* 25**,** 1010-22.

DEHAL, P. & BOORE, J. L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol,* 3**,** e314.

DURAND, D. & HOBERMAN, R. 2006. Diagnosing duplications--can it be done? *Trends Genet,* 22**,** 156-64.

EELLS, J. B., WILCOTS, J., SISK, S. & GUO-ROSS, S. X. 2012. NR4A gene expression is dynamically regulated in the ventral tegmental area dopamine neurons and is related to expression of dopamine neurotransmission genes. *J Mol Neurosci,* 46**,** 545-53.

ENGSTROM, P. G., FREDMAN, D. & LENHARD, B. 2008. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol,* 9**,** R34.

ENGSTROM, P. G., HO SUI, S. J., DRIVENES, O., BECKER, T. S. & LENHARD, B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res,* 17**,** 1898-908.

ERNST, J. & KELLIS, M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol,* 28**,** 817-25.

ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R., COYNE, M., KU, M., DURHAM, T., KELLIS, M. & BERNSTEIN, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

ESTELLER, M. 2006. CpG island methylation and histone modifications: biology and clinical significance. *Ernst Schering Res Found Workshop***,** 115-26.

EVANS, R. 2004. A transcriptional basis for physiology. *Nature medicine,* 10**,** 1022-1026.

EVANS, R. M. 2005. The nuclear receptor superfamily: a rosetta stone for physiology. *Mol Endocrinol,* 19**,** 1429-38.

FORSBERG, E. C., DOWNS, K. M., CHRISTENSEN, H. M., IM, H., NUZZI, P. A. & BRESNICK, E. H. 2000. Developmentally dynamic histone acetylation pattern of a tissue-specific chromatin domain. *Proc Natl Acad Sci U S A,* 97**,** 14494-9.

GASZNER, M. & FELSENFELD, G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet,* 7**,** 703-13.

GERMAIN, P., STAELS, B., DACQUET, C., SPEDDING, M. & LAUDET, V. 2006. Overview of nomenclature of nuclear receptors. *Pharmacol Rev,* 58**,** 685-704.

GUENET, J. L. 2005. The mouse genome. *Genome Res,* 15**,** 1729-40.

GUO, W., BURRIS, T. P., ZHANG, Y. H., HUANG, B. L., MASON, J., COPELAND, K. C., KUPFER, S. R., PAGON, R. A. & MCCABE, E. R. 1996. Genomic sequence of the DAX1 gene: an orphan nuclear receptor responsible for X-linked adrenal hypoplasia congenita and hypogonadotropic hypogonadism. *J Clin Endocrinol Metab,* 81**,** 2481-6.

HARRIS, R. M. & HOFMANN, H. A. 2015. Seeing is believing: Dynamic evolution of gene families. *Proc Natl Acad Sci U S A,* 112**,** 1252-3.

HURLES, M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol,* 2**,** E206.

IOSHIKHES, I. P. & ZHANG, M. Q. 2000. Large-scale human promoter mapping using CpG islands. *Nat Genet,* 26**,** 61-3.

JENSEN, E. V. & JACOBSON, H. I. 1960. Fate of steroid estrogens in target tissues. Academic Press, New York.

JENSEN, E. V. & JACOBSON, H. I. 1962. Basic guides to the mechanism of estrogen action. *Recent Prog Horm Res,* 18**,** 387.

JI, Y., WALKOWICZ, M. J., BUITING, K., JOHNSON, D. K., TARVIN, R. E., RINCHIK, E. M., HORSTHEMKE, B., STUBBS, L. & NICHOLLS, R. D. 1999. The ancestral gene for transcribed, low-copy repeats in the Prader-Willi/Angelman region encodes a large protein implicated in protein trafficking, which is deficient in mice with neuromuscular and spermiogenic abnormalities. *Hum Mol Genet,* 8**,** 533-42.

KEBEDE, A. F., SCHNEIDER, R. & DAUJAT, S. 2015. Novel types and sites of histone modifications emerge as players in the transcriptional regulation contest. *FEBS J,* 282**,** 1658-74.

KIKUTA, H., LAPLANTE, M., NAVRATILOVA, P., KOMISARCZUK, A. Z., ENGSTROM, P. G., FREDMAN, D., AKALIN, A., CACCAMO, M., SEALY, I., HOWE, K., GHISLAIN, J., PEZERON, G., MOURRAIN, P., ELLINGSEN, S., OATES, A. C., THISSE, C., THISSE, B., FOUCHER, I., ADOLF, B., GELING, A., LENHARD, B. & BECKER, T. S. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res,* 17**,** 545-55.

KIM, S. M., DUBEY, D. D. & HUBERMAN, J. A. 2003. Early-replicating heterochromatin. *Genes Dev,* 17**,** 330-5.

KRYAZHIMSKIY, S. & PLOTKIN, J. B. 2008. The population genetics of dN/dS. *PLoS Genet,* 4**,** e1000304.

LAUDET, V., HANNI, C., COLL, J., CATZEFLIS, F. & STEHELIN, D. 1992. Evolution of the nuclear receptor gene superfamily. *EMBO J,* 11**,** 1003-13.

LEISTER, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet,* 20**,** 116-22.

LETTICE, L. A., HEANEY, S. J., PURDIE, L. A., LI, L., DE BEER, P., OOSTRA, B. A., GOODE, D., ELGAR, G., HILL, R. E. & DE GRAAFF, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet,* 12**,** 1725-35.

LEVINGS, P. P. & BUNGERT, J. 2002. The human beta-globin locus control region. *Eur J Biochem,* 269**,** 1589-99.

LI, J. & AL-AZZAWI, F. 2009. Mechanism of androgen receptor action. *Maturitas,* 63**,** 142-8.

LI, Q., PETERSON, K. R., FANG, X. & STAMATOYANNOPOULOS, G. 2002. Locus control regions. *Blood,* 100**,** 3077-86.

LI, W. H., GU, Z., CAVALCANTI, A. R. & NEKRUTENKO, A. 2003a. Detection of gene duplications and block duplications in eukaryotic genomes. *J Struct Funct Genomics,* 3**,** 27-34.

LI, Y., LAMBERT, M. H. & XU, H. E. 2003b. Activation of nuclear receptors: a perspective from structural genomics. *Structure,* 11**,** 741-6.

MARINO-RAMIREZ, L., KANN, M. G., SHOEMAKER, B. A. & LANDSMAN, D. 2005. Histone structure and nucleosome stability. *Expert Rev Proteomics,* 2**,** 719-29.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

MEISLER, M. H. 2001. Evolutionarily conserved noncoding DNA in the human genome: how much and what for? *Genome Res,* 11**,** 1617-8.

MINUCCI, S., LEID, M., TOYAMA, R., SAINT-JEANNET, J. P., PETERSON, V. J., HORN, V., ISHMAEL, J. E., BHATTACHARYYA, N., DEY, A., DAWID, I. B. & OZATO, K. 1997. Retinoid X receptor (RXR) within the RXR-retinoic acid receptor heterodimer binds its ligand and enhances retinoid-dependent gene expression. *Mol Cell Biol,* 17**,** 644-55.

MORAS, D., BILLAS, I. M., ROCHEL, N. & KLAHOLZ, B. P. 2015. Structure-function relationships in nuclear receptors: the facts. *Trends Biochem Sci,* 40**,** 287-90.

MORAS, D. & GRONEMEYER, H. 1998. The nuclear receptor ligand-binding domain: structure and function. *Curr Opin Cell Biol,* 10**,** 384-91.

NARANJO, S., VOESENEK, K., DE LA CALLE-MUSTIENES, E., ROBERT-MORENO, A., KOKOTAS, H., GRIGORIADOU, M., ECONOMIDES, J., VAN CAMP, G., HILGERT, N., MORENO, F., ALSINA, B., PETERSEN, M. B., KREMER, H. & GOMEZ-SKARMETA, J. L. 2010. Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression during inner ear development and may be required for hearing. *Hum Genet,* 128**,** 411-9.

NAVRATILOVA, P., FREDMAN, D., HAWKINS, T. A., TURNER, K., LENHARD, B. & BECKER, T. S. 2009. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev Biol,* 327**,** 526-40.

OGBOURNE, S. & ANTALIS, T. M. 1998. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J,* 331 (Pt 1)**,** 1-14.

OHTA, T. 1990. How gene families evolve. *Theor Popul Biol,* 37**,** 213-9.

OLEFSKY, J. M. 2001. Nuclear receptor minireview series. *J Biol Chem,* 276**,** 36863-4.

PENNACCHIO, L. A., BICKMORE, W., DEAN, A., NOBREGA, M. A. & BEJERANO, G. 2013. Enhancers: five essential questions. *Nat Rev Genet,* 14**,** 288-95.

PENNACCHIO, L. A., LOOTS, G. G., NOBREGA, M. A. & OVCHARENKO, I. 2007. Predicting tissue-specific enhancers in the human genome. *Genome Res,* 17**,** 201-11.

PHILLIPS, J. E. & CORCES, V. G. 2009. CTCF: master weaver of the genome. *Cell,* 137**,** 1194-211.

POLS, T. W., BONTA, P. I. & DE VRIES, C. J. 2007. NR4A nuclear orphan receptors: protective in vascular disease? *Curr Opin Lipidol,* 18**,** 515-20.

RIETHOVEN, J. J. 2010. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol,* 674**,** 33-42.

ROBINSON-RECHAVI, M., CARPENTIER, A. S., DUFFRAISSE, M. & LAUDET, V. 2001. How many nuclear hormone receptors are there in the human genome? *Trends Genet,* 17**,** 554-6.

ROBINSON-RECHAVI, M., ESCRIVA GARCIA, H. & LAUDET, V. 2003. The nuclear receptor superfamily. *J Cell Sci,* 116**,** 585-6.

SACHS, M., ONODERA, C., BLASCHKE, K., EBATA, K. T., SONG, J. S. & RAMALHO-SANTOS, M. 2013. Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. *Cell Rep,* 3**,** 1777-84.

SANDELIN, A., CARNINCI, P., LENHARD, B., PONJAVIC, J., HAYASHIZAKI, Y. & HUME, D. A. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet,* 8**,** 424-36.

SEOL, W., CHOI, H. S. & MOORE, D. D. 1996. An orphan nuclear hormone receptor that lacks a DNA binding domain and heterodimerizes with other receptors. *Science,* 272**,** 1336-9.

SEVER, R. & GLASS, C. K. 2013. Signaling by Nuclear Receptors. *Cold Spring Harbor Perspectives in Biology,* 5.

SHLYUEVA, D., STAMPFEL, G. & STARK, A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet,* 15**,** 272-86.

SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet,* 13**,** 613-26.

SYMMONS, O. & SPITZ, F. 2013. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120358.

TACHIBANA, K., KOBAYASHI, Y., TANAKA, T., TAGAMI, M., SUGIYAMA, A., KATAYAMA, T., UEDA, C., YAMASAKI, D., ISHIMOTO, K., SUMITOMO, M., UCHIYAMA, Y., KOHRO, T., SAKAI, J., HAMAKUBO, T., KODAMA, T. & DOI, T. 2005. Gene expression profiling of potential peroxisome proliferator-activated receptor (PPAR) target genes in human hepatoblastoma cell lines inducibly expressing different PPAR isoforms. *Nucl Recept,* 3**,** 3.

TAYLOR, J. S. & RAES, J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet,* 38**,** 615-43.

TAYLOR, M. S., KAI, C., KAWAI, J., CARNINCI, P., HAYASHIZAKI, Y. & SEMPLE, C. A. 2006. Heterotachy in mammalian promoter evolution. *PLoS Genet,* 2**,** e30.

TESHIMA, K. M. & INNAN, H. 2008. Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics,* 178**,** 1385-98.

TRASK, B. J., FRIEDMAN, C., MARTIN-GALLARDO, A., ROWEN, L., AKINBAMI, C., BLANKENSHIP, J., COLLINS, C., GIORGI, D., IADONATO, S., JOHNSON, F., KUO, W. L., MASSA, H., MORRISH, T., NAYLOR, S., NGUYEN, O. T., ROUQUIER, S., SMITH, T., WONG, D. J., YOUNGBLOM, J. & VAN DEN ENGH, G. 1998. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum Mol Genet,* 7**,** 13-26.

VOIGT, P., TEE, W. W. & REINBERG, D. 2013. A double take on bivalent promoters. *Genes Dev,* 27**,** 1318-38.

WALKER, E. L., ROBBINS, T. P., BUREAU, T. E., KERMICLE, J. & DELLAPORTA, S. L. 1995. Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex. *EMBO J,* 14**,** 2350-63.

WANG, Z., ZANG, C., ROSENFELD, J. A., SCHONES, D. E., BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., PENG, W., ZHANG, M. Q. & ZHAO, K. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet,* 40**,** 897-903.

WASSERMAN, W. W. & SANDELIN, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet,* 5**,** 276-87.

WAYMAN, N. S., HATTORI, Y., MCDONALD, M. C., MOTA-FILIPE, H., CUZZOCREA, S., PISANO, B., CHATTERJEE, P. K. & THIEMERMANN, C. 2002. Ligands of the peroxisome proliferator-activated receptors (PPAR-gamma and PPAR-alpha) reduce myocardial infarct size. *FASEB J,* 16**,** 1027-40.

WOODS, S., COGHLAN, A., RIVERS, D., WARNECKE, T., JEFFRIES, S. J., KWON, T., ROGERS, A., HURST, L. D. & AHRINGER, J. 2013. Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet,* 9**,** e1003330.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

ZANARIA, E., MUSCATELLI, F., BARDONI, B., STROM, T. M., GUIOLI, S., GUO, W., LALLI, E., MOSER, C., WALKER, A. P., MCCABE, E. R. & ET AL. 1994. An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita. *Nature,* 372**,** 635-41.

ZHANG, Z., BURCH, P. E., COONEY, A. J., LANZ, R. B., PEREIRA, F. A., WU, J., GIBBS, R. A., WEINSTOCK, G. & WHEELER, D. A. 2004. Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res,* 14**,** 580-90.

ZHANG, Z., LI, J., ZHAO, X. Q., WANG, J., WONG, G. K. & YU, J. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics,* 4**,** 259-63.

ZHAO, Y. & BRUEMMER, D. 2010. NR4A orphan nuclear receptors: transcriptional regulators of gene expression in metabolism and vascular biology. *Arterioscler Thromb Vasc Biol,* 30**,** 1535-41.

ZHAO, Y., ZHANG, K., GIESY, J. P. & HU, J. 2015. Families of nuclear receptors in vertebrate models: characteristic and comparative toxicological perspective. *Sci Rep,* 5**,** 8554.

ZHOU, V. W., GOREN, A. & BERNSTEIN, B. E. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet,* 12**,** 7-18.

I

OPEN ACCESS Freely available online

Computational Characterization of Modes of Transcriptional Regulation of Nuclear Receptor Genes

Yogita Sharma[1], Chandra Sekhar Reddy Chilamakuri[2¤], Marit Bakke[1], Boris Lenhard[3,4*]

1 Department of Biomedicine, University of Bergen, Bergen, Norway, **2** Department of Clinical Medicine, University of Bergen, Bergen, Norway, **3** Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, and MRC Clinical Sciences Centre, London, United Kingdom, **4** Department of Informatics, University of Bergen, Bergen, Norway

Abstract

Background: Nuclear receptors are a large structural class of transcription factors that act with their co-regulators and repressors to maintain a variety of biological and physiological processes such as metabolism, development and reproduction. They are activated through the binding of small ligands, which can be replaced by drug molecules, making nuclear receptors promising drug targets. Transcriptional regulation of the genes that encode them is central to gaining a deeper understanding of the diversity of their biochemical and biophysical roles and their role in disease and therapy. Even though they share evolutionary history, nuclear receptor genes have fundamentally different expression patterns, ranging from ubiquitously expressed to tissue-specific and spatiotemporally complex. However, current understanding of regulation in nuclear receptor gene family is still nascent.

Methodology/Principal Findings: In this study, we investigate the relationship between long-range regulation of nuclear receptor family and their known functionality. Towards this goal, we identify the nuclear receptor genes that are potential targets based on counts of highly conserved non-coding elements. We validate our results using publicly available expression (RNA-seq) and histone modification (ChIP-seq) data from the ENCODE project. We find that nuclear receptor genes involved in developmental roles show strong evidence of long-range mechanism of transcription regulation with distinct *cis*-regulatory content they feature clusters of highly conserved non-coding elements distributed in regions spanning several Megabases, long and multiple CpG islands, bivalent promoter marks and statistically significant higher enrichment of enhancer mark around their gene loci. On the other hand nuclear receptor genes that are involved in tissue-specific roles lack these features, having simple transcriptional controls and a greater variety of mechanisms for producing paralogs. We further examine the combinatorial patterns of histone maps associated with dynamic functional elements in order to explore the regulatory landscape of the gene family. The results show that our proposed classification capturing long-range regulation is strongly indicative of the functional roles of the nuclear receptors compared to existing classifications.

Conclusions/Significanc: We present a new classification for nuclear receptor gene family capturing whether a nuclear receptor is a possible target of long-range regulation or not. We compare our classification to existing structural (mechanism of action) and homology-based classifications. Our results show that understanding long-range regulation of nuclear receptors can provide key insight into their functional roles as well as evolutionary history; and this strongly merits further study.

Citation: Sharma Y, Chilamakuri CSR, Bakke M, Lenhard B (2014) Computational Characterization of Modes of Transcriptional Regulation of Nuclear Receptor Genes. PLoS ONE 9(2): e88880. doi:10.1371/journal.pone.0088880

Editor: Rolf Müller, Philipps University, Germany

Received September 6, 2013; **Accepted** January 15, 2014; **Published** February 13, 2014

Copyright: © 2014 Sharma et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: YS was supported by the National Program for Research in Functional Genomics in Norway (FUGE) in the Research Council of Norway. CSRC was supported by the EU project NeuroXsys (FP7 grant 223262). BL was supported by Norwegian Research Council (YFF scheme), Bergen Research Foundation and Medical Research Council (UK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: b.lenhard@imperial.ac.uk

¤ Current address: Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway

Introduction

Nuclear receptors comprise one of the largest groups of transcription factors that regulate the activity of complex gene networks [1,2,3]. These genes work in concert with co-activators and co-repressors to regulate a wide variety of biological processes such as embryonic development, organogenesis and metabolic homeostasis [4,5]. Improper functioning of nuclear receptors has been implicated in various developmental and physiological disorders [6], and nuclear receptors are known to be promising drug targets [7,8].

Nuclear receptors are broadly classified either based on their sequence similarity [9] or depending on their ligands [10]. Based on sequence homology, nuclear receptors have been categorized into 7 subclasses [9]. Alternatively, nuclear receptors are classified as nuclear hormone receptors (NHR) or nuclear orphan receptors (NOR) based on their mechanism of action. Nuclear hormone receptors are activated via ligand binding, but ligand binding by nuclear orphan receptors has not been demonstrated [11] and their mechanism of action is poorly understood. Some studies have reported that they are activated by post-translational modification or direct transcriptional activation [12,13]. Furthermore, some nuclear receptors have been categorized into tissue-specific and developmental regulatory based on their known functional roles [14,15,16].

Early research explored the structural properties of nuclear receptors [17], while recent work has focused on understanding how individual nuclear receptors control the transcription of their target genes [18,19,20,21]. However, how nuclear receptors are themselves regulated (rather than how they regulate their target genes) is not well understood [22,23]. This leads to the following question: Does regulation of nuclear receptor genes exhibit characteristic behavior in terms of their sequence similarity, mechanism of action or functional roles? Understanding regulation of nuclear receptors promises fresh insight into the functional roles of these genes, and possibly, accounting for at least a subset of disease-associated variation found in their vicinity.

In this paper, we hypothesize that the diversity of the biological and biochemical roles of nuclear receptors is reflected in fundamental differences in their transcriptional regulation e.g. whether the nuclear receptor in question is a target of long-range regulation or not. Like many other genes specific for one tissue, tissue-specific ligand-modulated nuclear receptors are expected to have relatively simple transcriptional control: they will be turned on in their target tissue only, and consequently, may not be targets of long-range regulation. On the other hand, nuclear receptors involved in developmental processes should exhibit properties that have been established for developmentally regulated genes [24]. These properties include long-range control of gene regulation by highly conserved non-coding elements and multiple long CpG islands. The highly conserved non-coding elements form clusters in a large region around their target gene loci and can function as enhancers [25].

It has been proposed that nuclear receptors first appeared as a single gene that has duplicated and diversified into current seven subfamilies during evolution [26]. We hypothesize that in many cases, it is the ancestral and not the currently extant gene loci that have been recruited into the developmental or the tissue-specific roles. Those functions were then passed to their duplicate offspring loci, which then sub-functionalized or acquired entirely new functions with different mode of regulation.

In this study of the nuclear receptor gene family, our aim was to establish whether or not they possess properties that would classify them as targets of long-range developmental regulation, and analyzed the relationship between their *cis*-regulatory content and their known functions. To facilitate this work we used an established genomic regulatory block (GRB) model [27,28]. A GRB is a locus on a chromosome that carries all the regulatory input required for the expression of a 'target' gene. This block comprises a target gene, its enhancers including highly conserved non-coding elements (HCNEs) and often bystander genes. Target genes receive regulatory input from HCNEs, which can be present either in inter- or intra-genic regions (Figure 1). Bystander genes contain HCNEs in their introns or beyond, but do not respond to their regulatory input; these HCNEs also control the target gene resulting in conservation of synteny between the two genes as a by-product of maintaining the organization of GRBs, which needs to be conserved for the normal functioning of the target gene [29,30].

Our first aim was to establish which genes among the nuclear receptors are potential GRB target genes. We then investigated the impact of the *cis*-regulatory content of each gene in order to gain a deeper understanding of its transcriptional regulation. Using publicly available datasets from the ENCODE project [31], we considered histone modifications known to be associated with promoters, enhancers, transcriptional repression and transcription elongation. Finally, to understand the complete regulatory landscape of nuclear receptors, we used chromatin states map data obtained by ChromHMM segmentation on ENCODE cell lines [32], consisting of the genome-wide combinatorial patterns of various histone marks, which are known to be associated with distinct biological functions [33]. We studied the enrichment pattern of all the defined chromatin states in nuclear receptors in the H1 human embryonic stem cell line (H1hesc). We define a new classification of nuclear receptor genes on the basis of their transcriptional regulation, and show that nuclear receptors naturally fall into two clusters: one comprising GRB target genes, i.e. developmental regulators that maintain a complex pattern of expression; and one comprising non-target genes that require simpler transcriptional control. The evolutionary history of nuclear receptor genes shows the differential use of whole-genome versus gene duplications between the two groups. This study will aid in better understanding of the regulatory mechanism of nuclear receptor genes and their functional diversity.

Results

Classification of Nuclear Receptors with Respect to GRB Model

Our first aim was to determine which nuclear receptor genes possess the properties of GRB target genes. To facilitate this, we analyzed the HCNE regions around each nuclear receptor gene locus across five vertebrate genomes. Since it has been shown that most HCNEs act as long-range enhancers of their target genes [34], we analyzed HCNEs in 1 Mb or 2 Mb span upstream and downstream of gene loci, using custom levels of conservation for different species. To maximize the information from the set of elements for each of the selected vertebrate species, the conservation threshold for different species was chosen between 70 to 100 percent, depending on the evolutionary distance from human (see Table S1 for details). We calculated HCNE counts around 2 Mb region of each nuclear receptor gene loci.

Detection of HCNE regions was the first step towards identifying which genes in the nuclear receptor family have the features of GRB target genes. We computed dissimilarity matrix of HCNEs between human and five selected vertebrate genomes and performed the hierarchical clustering (see Methods section on "HCNE and CpG islands detection"). We found that whole gene family can be broadly divided into two main clusters containing 25 and 23 genes respectively (Figure 2).

Table 1 shows the list of genes in the two clusters as well as their functional and structural classification. The genes in cluster 1 have a higher span of HCNEs around their gene loci, whereas cluster 2 genes have few or no HCNEs (Table 1). Interestingly, the first cluster comprises of many genes that are known targets of long-range gene regulation (e.g. *NR2F2*, *PPARG* [24]). Thus, cluster 1 corresponding to high HCNE counts in the GRB model is indicative of possible targets of long-range gene regulation. In the sequel, we explore this hypothesis further by considering other promoters and *cis*-regulatory elements.

We observe that the genes are dispersed among the two clusters irrespective of their homology-based classification (Table 1), indicating that following duplication events in evolutionary history, one of the genes acquired a different mode of regulation. However, we observe that most recent paralog pairs of genes (e.g. *NR2F2* and *NR2F1*; *NR5A2* and *NR5A1*) reside in the same cluster, with few exceptions (e.g. *PPARG* and *PPARA; NR2E1* and *NR2E3*).

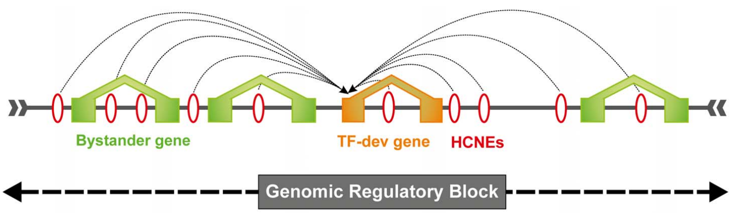


Figure 1. The GRB Model. GRB has developmental and/or transcription factor gene (target gene, orange) spanned by a cluster of highly conserved non-coding elements (red ovals), which regulates the target gene expression by acting as enhancers/insulators and other un-related neighboring genes (bystander genes, green).
doi:10.1371/journal.pone.0088880.g001

Indeed, close paralogs belonging to the first cluster can be traced back to one of the two rounds of whole-genome duplication that happened at the root of vertebrates. This is naturally indicative that the genes in the first cluster having high HCNE counts have possibly evolved through whole-genome duplication rather than tandem duplication. Due to the megabase span of their regulatory regions, it is practically impossible for GRB target genes to undergo tandem duplication without disrupting the array of associated regulatory elements.

The above analysis is based on the genomes of five species. To understand the variation within species, we perform subsequent analysis by comparing HCNE counts among each species to human. We visualized HCNEs of each gene loci across 2 Mb region using 1 kb windows in the two clusters (Figure 3). We observe that the genes in cluster 1 (shown in red) have a higher number as well as a wider span of HCNEs around their gene loci in comparison to the genes in cluster 2 (shown in blue). Both the number and the maximum span of HCNEs decreased with increasing evolutionary distance from human, e.g. human-mouse compared to human -zebrafish. However, the number of HCNEs decreases with increasing evolutionary distance but still does not completely disappear in cluster 1 even at the highest investigated distance i.e. human-zebrafish.

It has been shown earlier that GRB target genes often have higher ratios between CpG island length and transcript length [25]. In contrast to most other genes, CpG islands in GRB target genes not only cover the promoter region but also extend into the body of the gene, in some cases, spanning the entire target gene. Therefore we checked the CpG islands around gene loci in cluster 1 and 2 and found that most of the genes in cluster 1 have longer CpG islands in comparison to cluster 2 (Wilcoxon test, p-value $<$ 0.0001), confirming that the high HCNE counts and multiple long CpG islands are correlated features of the genes present in cluster 1. Since we are analyzing the length of CpG islands among genes; we excluded the genes that do not overlap with any CpG island in both clusters. We also checked the CpG length of putative GRB target nuclear receptors (cluster 1) with randomly selected transcription factor genes, and with the set of all genes overlapping CpG islands. From the cumulative distribution plots (Figure S1), it is clear that GRB target nuclear receptors have longer CpG islands than the other sets.

Extended Validation based on other Transcription Factors

To further validate the two classes, we compared the HCNE counts of the nuclear receptor gene family with other transcription factors. Specifically, we created a random dataset of 48 transcription factor genes and computed the HCNEs across the five vertebrate genomes (see Methods for details). We repeated previous experiment using the extended set of 96 genes (48 nuclear receptors and 48 randomly selected transcription factors) with the same distance and conservation threshold as before. We found that the extended set was divided into two major clusters (Figure S2 and Table S2). The first cluster comprised of 31 genes in total, out of which 25 are nuclear receptors and 6 are other transcription factors (Cluster A in Table S2). The second cluster has 65 genes, 23 of which are nuclear receptors and 42 are other transcription factors (Cluster B). The resulting clustering agrees with previous results i.e. the genes that clustered together in previous HCNE analysis (cluster 1 in Table 1) are part of the same cluster here (cluster 1 in Table S2). Interestingly, we also found other transcription factors (*PAX2, SOX2, MEIS2*) in this cluster that are known targets of long-range gene regulation [36,37,38]. This shows that the previous clustering is robust and functionally significant, and more generally, that this method can be used to study other developmental regulated genes as well.

Identification of Target Nuclear Receptor in GRB Loci having Several Genes

In the previous analysis (Table 1), we found three cases of GRB loci with several target genes appearing in cluster 1, namely (*THRB, RARB, NR1D2*), (*THRA, RARA, NR1D1*), and (*NR6A1, NR5A1*) wherein the genes in each case share a common locus w.r.t. HCNEs within a ±2 Mb region. In such a scenario, it is not immediately clear which of the gene (or genes) is the target in the corresponding GRB locus. Investigating further, we found that in each of the cases above, the genes are present in synteny in human and mouse (see Figure S3) – lending further credence to the idea that these genes were part of whole-genome duplication.

However, the problem of identifying target genes in a GRB locus remains. While proximity of each gene to HCNE peaks offers some indication, it is not sufficient. In the sequel, we report experiments based on expression and histone-modification data in the H1hesc embryonic stem cell line. The results (which are described in more detail later in the manuscript) address the afore-mentioned problem based on presence of bivalent domain in the promoter region of the gene.

In the first case, *RARB* was located most closely to the peaks of highest HCNEs and also it has bivalent promoter (though very weak) in H1hesc cell line. On the other hand, the genes *NR1D2* and *THRB* have neither a proximal HCNEs peak (in comparison to other common gene in GRB locus) nor a bivalent promoter. Therefore, we annotate *RARB* to be the putative target of this GRB locus. In the second case, all the three genes (*THRA, RARA, NR1D1*) shares the same proximity of HCNEs around each other but only two (*RARA* and *NR1D1*) have bivalent promoters; therefore we annotated these two as targets of the same GRB locus. (Both of these follow same expression pattern in rest of the

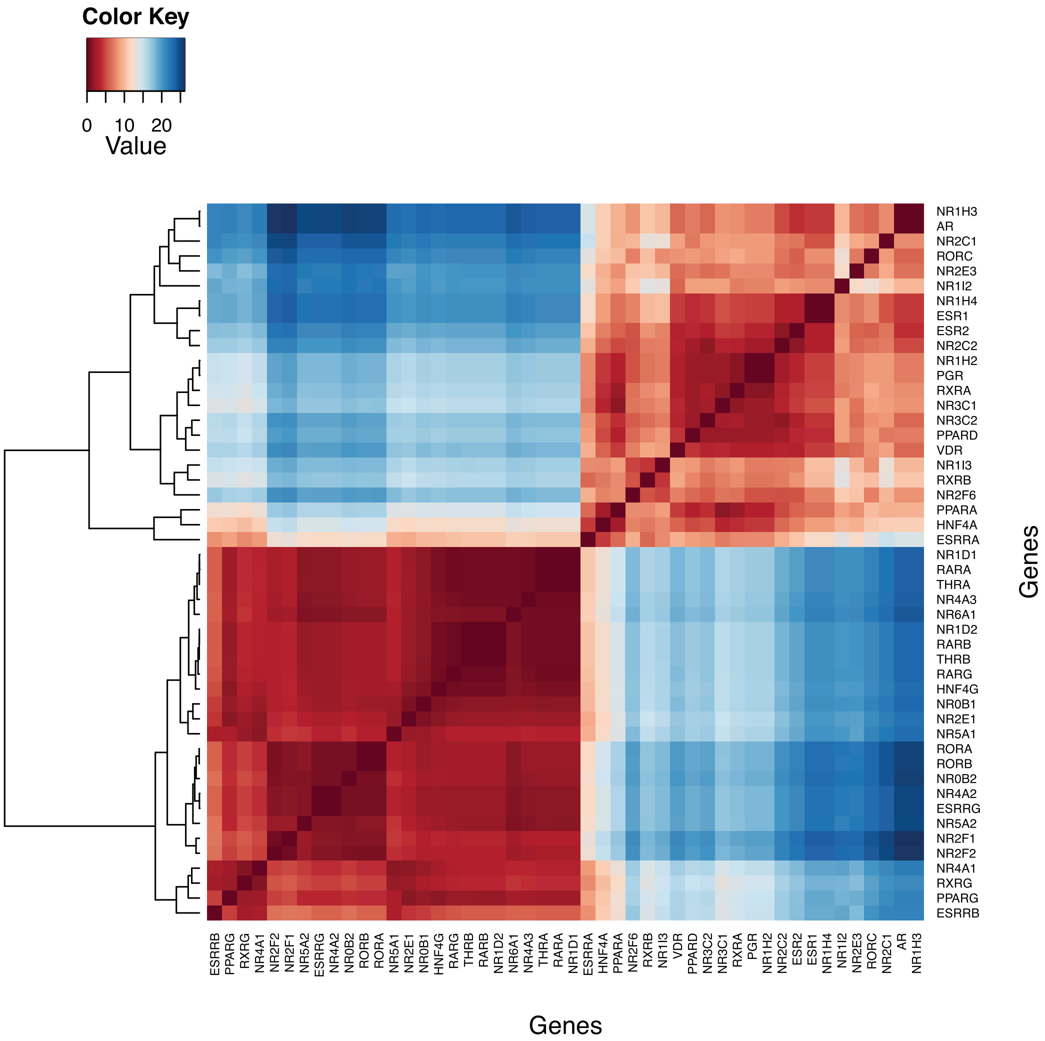


Figure 2. The dissimilarity matrix of HCNE content among nuclear receptors and its clustering. Nuclear receptor genes broadly divided in to two clusters on the basis of higher and lower enrichment of HCNEs around 2 Mb region of their gene loci in 5 vertebrate genomes. The first cluster (shown below) consists of 25 genes having higher enrichment of HCNE, while cluster 2 consists of the remaining 23 genes.
doi:10.1371/journal.pone.0088880.g002

cell lines). In the third case, both *NR6A1* and *NR5A1* exhibit similar proximity of HCNEs but neither have a bivalent domain. In this case, the *NR6A1* gene is already highly expressed in H1hesc cell line in comparison to other expressed genes, while gene *NR5A1* is completely shut down. Therefore we annotated both of these genes as putative targets of the GRB.

Distinct Expression Profiles of Cluster 1 and Cluster 2 Genes

To investigate the expression properties of cluster 1 and cluster 2 genes, we used read per kilobase per million (RPKM) values for each gene from RNA-seq data across 5 ENCODE cell lines (Table S3). Based on this, we categorized each gene set on the basis of expression significantly above the background (RPKM = 0.3) in respective cell lines, following approach in [39]. The total number of genes expressed across different cell lines was highest in the H1hesc and HepG2 cells. For each cell line, we considered four sets of genes obtained on the basis of their expression significantly above and below the background across both the clusters.

We observe that most genes belonging to cluster 1 are expressed in H1hesc (Table S3) and had relatively lower RPKM with few exceptions. On the other hand, the genes in cluster 2 had either

Table 1. The list of genes in clusters obtained using HCNE based analysis in the GRB model.

Gene Name	Cluster ID	Homology-based subfamily	Mechanism of action
NR1D1	1	I	NHR
RARA	1	I	NHR
THRA	1	I	NHR
NR4A3	1	IV	NOR
NR6A1	1	VI	NOR
NR1D2	1	I	NHR
RARB	1	I	NHR
THRB	1	I	NHR
RARG	1	I	NHR
HNF4G	1	II	NHR
NR0B1	1	0	NOR
NR2E1	1	II	NOR
NR5A1	1	V	NHR
RORA	1	I	NHR
RORB	1	I	NHR
NR0B2	1	0	NOR
NR4A2	1	IV	NOR
ESRRG	1	III	NOR
NR5A2	1	V	NHR
NR2F1	1	II	NOR
NR2F2	1	II	NOR
NR4A1	1	IV	NOR
RXRG	1	II	NHR
PPARG	1	I	NHR
ESRRB	1	III	NOR
NR1H3	2	I	NHR
AR	2	III	NHR
NR2C1	2	II	NOR
RORC	2	I	NHR
NR2E3	2	II	NOR
NR1I2	2	I	NHR
NR1H4	2	I	NHR
ESR1	2	III	NHR
ESR2	2	III	NHR
NR2C2	2	II	NOR
NR1H2	2	I	NHR
PGR	2	III	NHR
RXRA	2	II	NHR
NR3C1	2	III	NHR
NR3C2	2	III	NHR
PPARD	2	I	NHR
VDR	2	I	NHR
NR1I3	2	I	NHR
RXRB	2	II	NHR
NR2F6	2	II	NOR
PPARA	2	I	NHR
HNF4A	2	II	NHR
ESRRA	2	III	NOR

The homology-based classification is into seven categories: (I) Thyroid Hormone Receptor-like, (II) Retinoid X Receptor-like, (III) Estrogen Receptor-like, (IV) Nerve Growth Factor IB-like, (V) Steroidogenic Factor-like, (VI) Germ Cell Nuclear Factor-like, and (0) Miscellaneous. The functional classification is into nuclear hormone receptors (NHR) and nuclear orphan receptors (NOR).

doi:10.1371/journal.pone.0088880.t001

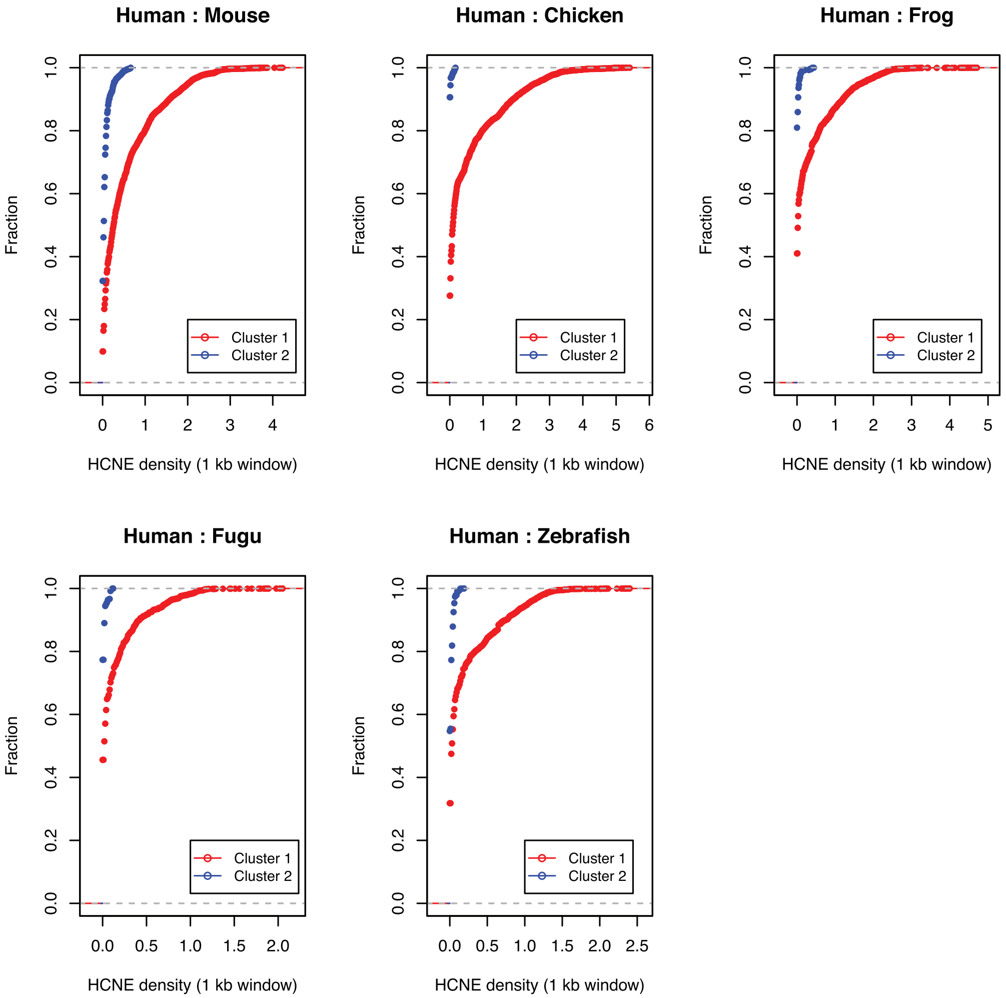


Figure 3. Cumulative distribution plots of HCNE content for human versus 5 vertebrate genomes in 2 Mb region from gene loci across different clusters. Cluster 1 (putative GRB target genes) is shown in red and cluster 2 (GRB non-target genes) is shown in blue. The x-axis shows HCNE distribution in 1 kb window and y-axis show the fraction of HCNE in selected window. This figure shows that Cluster 1 has higher fraction of HCNEs in comparison to cluster 2.

doi:10.1371/journal.pone.0088880.g003

expression in one cell line (e.g. *HNF4A*, and *NR1H4* were specific for HepG2 cell line) or they had very high expression values across all the cell lines (e.g. *NR2C2* and *NR2C1*). This shows that the clustering likely separates developmentally regulated genes from all other genes (ubiquitous and tissue specific) in line with the ability of their promoters to respond to long-range regulation [40].

H3K4me3 and H3K36me3 Enrichment Confirms Expression-based Analysis

To check the expression status of genes, it was crucial to check if the selected RPKM threshold of 0.3 actually correlates with the histone marks of expressed genes. To confirm this, in both clusters we studied the enrichment profiles of histone modification that relates to active promoter (H3K4me3) in respective cell lines (see section on "ChIP-seq data" in Methods for details). We selected

±10 kb region around transcription start sites for the analysis and plotted the coverage. We found the enrichment of active promoter mark peaks in promoter region of genes expressed significantly above the background across both the cluster 1 and cluster 2 gene sets. No enrichment was observed when the genes are in low expression state (Figures S4 and S5).

We also analyzed the enrichment of transcription elongation mark (H3K36me3) across genes in both the clusters (see section on "ChIP-seq data" in Methods for details). To be able to handle the difference in gene coordinates, we used ±20 kb genomic ranges around the midpoint of each gene where the midpoint is chosen to be the mean of the gene start and end coordinates. The enrichment of transcription elongation mark was observed across the gene body of only those genes that express significantly above the background in both the clusters in their respective cell lines; there was no enrichment when genes are low expressed. Both of these analyses confirm the main objective and showed the accuracy of expression state of gene sets created on the basis of selected threshold value.

Loci of Cluster 1 Genes have Significantly Higher Enrichment of H3K4me1

We are mainly interested in exploring the differences in regulatory content of genes with respect to their functions; those involved in developmental regulation must be under long-range control. Therefore, we analyzed the enrichment profiles of histone modification (H3K4me1) in H1hesc stem cell line (see section "ChIP-seq data" in Methods), a modification associated with active and poised enhancers. For H3k4me1 analysis across the different clusters, we did not consider the expression state of genes in respective cell lines, as its already shown in various studies that this mark is related to active and poised enhancer, and is not predictive of current transcription state.

We plotted the average coverage plots ±50 kb around transcription start site (TSS) for both of the clusters. We chose ±50 kb as a compromise value between establishing the existence of long-range regulation and avoidance of inclusion of regulatory elements of neighboring genes. We found that cluster 1 has higher enrichment of enhancer marks in comparison to cluster 2.

To check whether the observed difference is statistically significant, we created background distribution of H3K4me1 number of reads as well as specific datasets of CpG-overlapping and non-CpG promoters (see Methods for details). We study enhancer mark for each dataset with respect to this background distribution across different genomic ranges (see Methods for details).

Figure 4 shows the distribution of reads for each of the selected genomic ranges (respectively, ±10 kb, ±1 Mb and ±2 Mb). We define the critical region for each of the chosen widths by considering log2 value computed from the 0.95-quantile of the corresponding background distribution. Finally we check the occurrence of each dataset with respect to this critical region by considering log2 value of the average number of reads in each of the four original datasets, namely, nuclear receptors in clusters 1 and 2, as well as background set with and without CpG-islands.

We find that for each genomic range under consideration (respectively, ±10 kb, ±1 Mb and ±2 Mb), cluster 1 consistently falls well outside the critical region of the corresponding background distribution (Figure 4). We also observe that the set of CpG genes falls outside of critical region when we consider a region of ±10 kb around TSS. This concurs with the fact that in general CpG genes tend to have higher enrichment of H3k4me1 around their promoter region in comparison to non-CpG genes. However, when we consider ±1 Mb and ±2 Mb genomic regions; three of the four sets of gene, namely, cluster 2, the set of CpG genes, and the set of non-CpG genes, fall within the critical region of the background distribution. This analysis clearly shows that cluster 1 genes have statistically significant higher enrichment of enhancer mark around ±1 Mb and ±2 Mb of their transcription start site, indicating that they follow long-range mechanism of gene regulation, unlike the genes of cluster 2. To exclude the possibility of bias, we have also repeated the experiment by using genes on chromosome 5 for the background distribution. We found that genes in cluster 1 still have significantly higher enrichment of H3K4me1 across the different genomic ranges (Figure S8).

Cluster 1 Genes have Bivalent Promoters in H1hesc Stem Cell Line

It is known that genes involved in developmental regulation have bivalent promoters in stem cells [41], which means they have both active (H3K4me3) and repressive (H3K27me3) histone mark enrichment on the same locus. The presence of bivalent promoter mark enables these genes to turn on and off rapidly across different time points of development [41]. The bivalent state indicates a repressed state poised for activation. On activation, H3K27me3 is removed and only H3K4me3 remains. We were interested to test this observation across genes of both clusters in human embryonic stem cell line (H1hesc). We found that repression mark was completely absent in cluster 2 irrespective of their expression state in embryonic stem cell line, confirming that this cluster consists of a mixture of ubiquitously expressed genes and genes specifically expressed in later stages of differentiation.

The genes in cluster 1 consistently show evidence of involvement in developmental processes. We observed very high enrichment of repression mark around promoter region across genes in cluster 1 specifically when they are not expressed (Figure S6), showing that they have the type of promoter required to facilitate their complex pattern of expression.

Figure 5 shows the correlation of the two promoter marks across both clusters, we plotted bubble plots for each gene showing H3K27me3 and H3K4me3 marks for each gene at x-axis and y-axis respectively, and the expression level (derived from RNA-seq RPKM values, see Methods for details) represented by the size of the bubble. The genes in cluster 2 (marked in black) do not have read counts for H3K27me3 repression mark even when they are not expressed, while on other hand genes in cluster 1 (marked in red) have very high read counts for repression mark when they are not expressed (appearing in bottom-right quadrant). This is consistent with our hypothesis that genes in cluster 2 do not have long-range regulation, and consequently, do not need a repressive promoter mark. On the other hand, we posit that genes in cluster 1 as targets of long-range regulation; and show high repressive mark pausing transcription and resulting in low expression (bottom-right quadrant in Figure 5).

We further notice a handful of genes in cluster 1 (*ESRRA, NR6A1, RARG, RORA, RARA*) do not have repression mark (appearing in top-left quadrant), while having high expression values (large bubbles in the plot). These genes are likely turned on early enough to be active in H1 hESC cells, but their expression pattern across other cell lines and H3K4me1 mark content at their loci still confirm that they are under developmental regulation.

The most interesting observation we make is that few genes in cluster 1 (*NR4A1, NR5A2, NR1D1, RORB* and *ESRRG*) still retain repression read counts even when they are actively transcribed (shown in top-right quadrant of Figure 5). We believe these genes represent the transition either from expressed and no repressive mark (top-left quadrant) to low expressed and high repressive mark

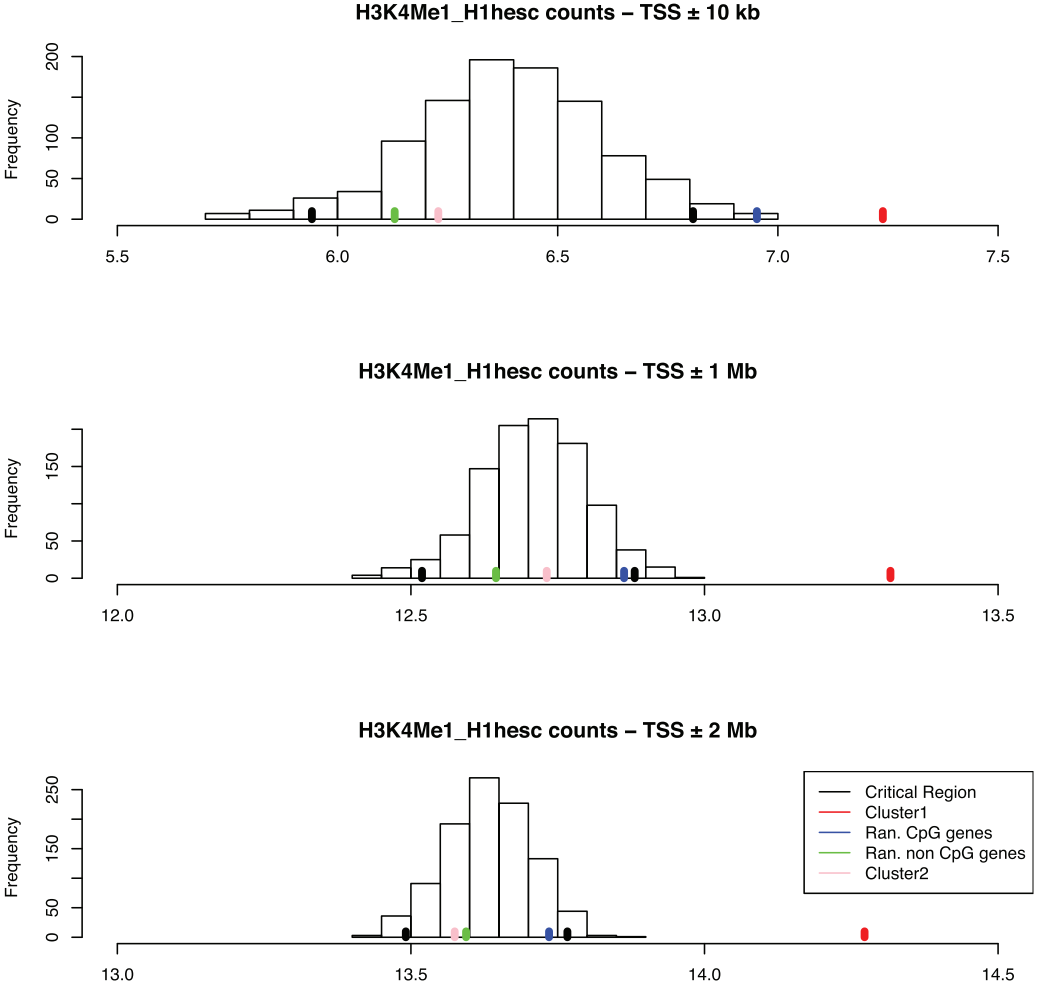


Figure 4. Statistical significance test for H3K4me1 around different genomic distributions. A) H3K4me1 distribution in different clusters across ±10 kb TSS against the random background distribution. B) H3K4me1 distribution in different clusters across ±1 Mb TSS with respect to random background distribution. C) H3k4me1 distribution in different clusters across ±2 Mb TSS with respect to random background distribution. This figure shows that cluster 1 (shown by red bar) has significantly higher distribution of H3K4me1 in comparison to random selected background region (marked by black bars), CpG and non-CpG region (shown by blue and green bar respectively) and cluster 2 genes (shown by pink bar).
doi:10.1371/journal.pone.0088880.g004

(bottom-right quadrant), or vice versa. We further investigated how exactly the promoter region looked in these five cases (Figure S7). A closer look at promoter region reveals that in case of *NR1D1* and *RORB,* it seems like the promoter itself is not covered by the repression mark, which starts slightly downstream and extends into the first intron (Figure S7). The functional significance of this arrangement is unknown, but may represent a configuration conductive to rapid repression. The remaining three genes, namely *NR4A1*, *NR5A2* and *ESRRG*, also retain repression mark but are possibly transcribed from an alternative promoter. This merits further study possibly using time-series experiments in order to capture the dynamic activation and repression during development.

GRB-based Clustering is Recovered from Chromatin State Map Analysis

To have better understanding of regulatory regions of nuclear receptors, we analyzed the chromatin state maps data for each gene in H1hesc cell line. This data represents the genome-wide mapping of different combinatorial patterns of histone marks, each of which is associated with specific biological function. The chromatin state map from [33] consists of 15 states, corresponding to the different functional elements of genome. To distinguish

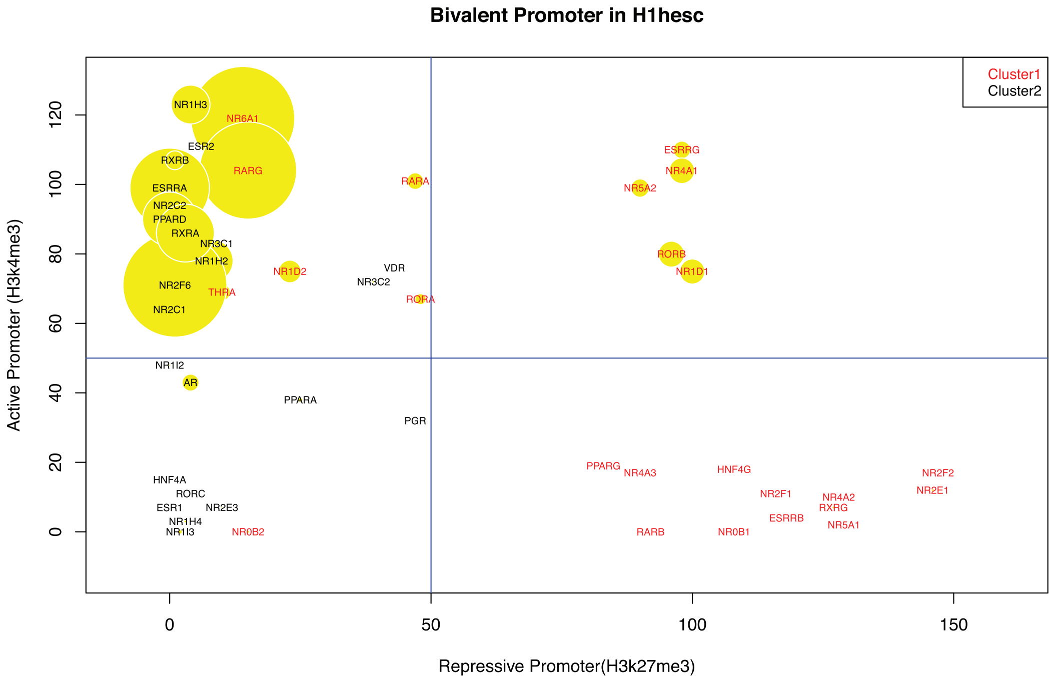


Figure 5. The bubble plots for bivalent promoter mark for each gene in human embryonic stem cell line. The x-axis shows read counts for repression (H3K27me3) mark around ±10 KB TSS. The y-axis shows read counts for active promoter (H3K4me3) mark around ±10 KB TSS. The size of the bubble (yellow) shows RPKM value for respective gene. The left section of the plot comprises all of the genes (black) in cluster 2 (except few cases where cluster 1 gene have very high expression). This shows that cluster 2 genes does not have any enrichment of repression mark around their TSS irrespective of their expression. The top and bottom right sections consist of genes from cluster 1 (red). This shows that when genes in cluster 1 are not expressed they have higher read counts for repression mark while still some of the genes retain repression mark even when they are expressed.

doi:10.1371/journal.pone.0088880.g005

between active and repressed state of a gene, we also included the expression data in this analysis. For each nuclear receptor gene, we studied the correlation of different states with its expression.

Like in the case of previous analyses, we found that nuclear receptor genes separated into two major clusters on the basis of different enrichment of various chromatin states (Figure 6). The obtained clusters were based on the two main criteria: the expression status of the gene, and the difference in *cis*-regulatory functional elements. The column dendrogram shows that state correspond to active promoter correlates well with the expression (RNA-seq) data, which means that when genes are expressed significantly above the background they have higher number of counts for active promoter state and vice versa. The states that correspond to transcribed regions also correlate with the active promoter state, which confirms the presence of active transcription. The states that correspond to poised promoter and Polycomb repression occur together and are in a different column. Similarly the states that correspond to poised and weak enhancer show high correlation to each other, and so do the states that represent heterochromatin and insulator region. This shows that the column dendrogram corresponds well with the active biological functions.

However, in the row dendrogram i.e. at the gene level, nuclear receptors have broadly separated into two clusters, and each cluster is sub-classified in further two clusters depending on the expression level of the genes. The genes have different combinatorial patterns of states with respect to their expression state across the same cluster. We note that the obtained clustering based on HMM state map is consistent with the previous clusters found based on HCNE analysis (Table 1), with three exceptions, namely *THRA*, *THRB* and *RARB*. This is because GRB-based clustering takes into account the fact that these genes are in close proximity to other target genes, while HMM state maps do not take spatial proximity into account.

The genes present in cluster 1 exhibit enrichment of poised promoter state except three genes (*NR6A1, ESRRA, RARG*), because of their very high expression in this cell line. The genes having expression significantly above the background present in cluster 1 show enrichment of state that corresponds to active promoter and transcribed region, as well as higher enrichment of states that relates to weak enhancers. In contrast, the genes that do not have expression significantly above the background in cluster 1 are highly enriched in poised promoter state along with strong Polycomb repression and complete loss of active transcription states and RNA-seq signal.

Cluster 2 can be further sub-divided into two subclusters on the basis of expression level, but the associated states are distinct from those in cluster 1. The main difference lies in the enrichment of poised promoter and poised enhancer states. The genes present in

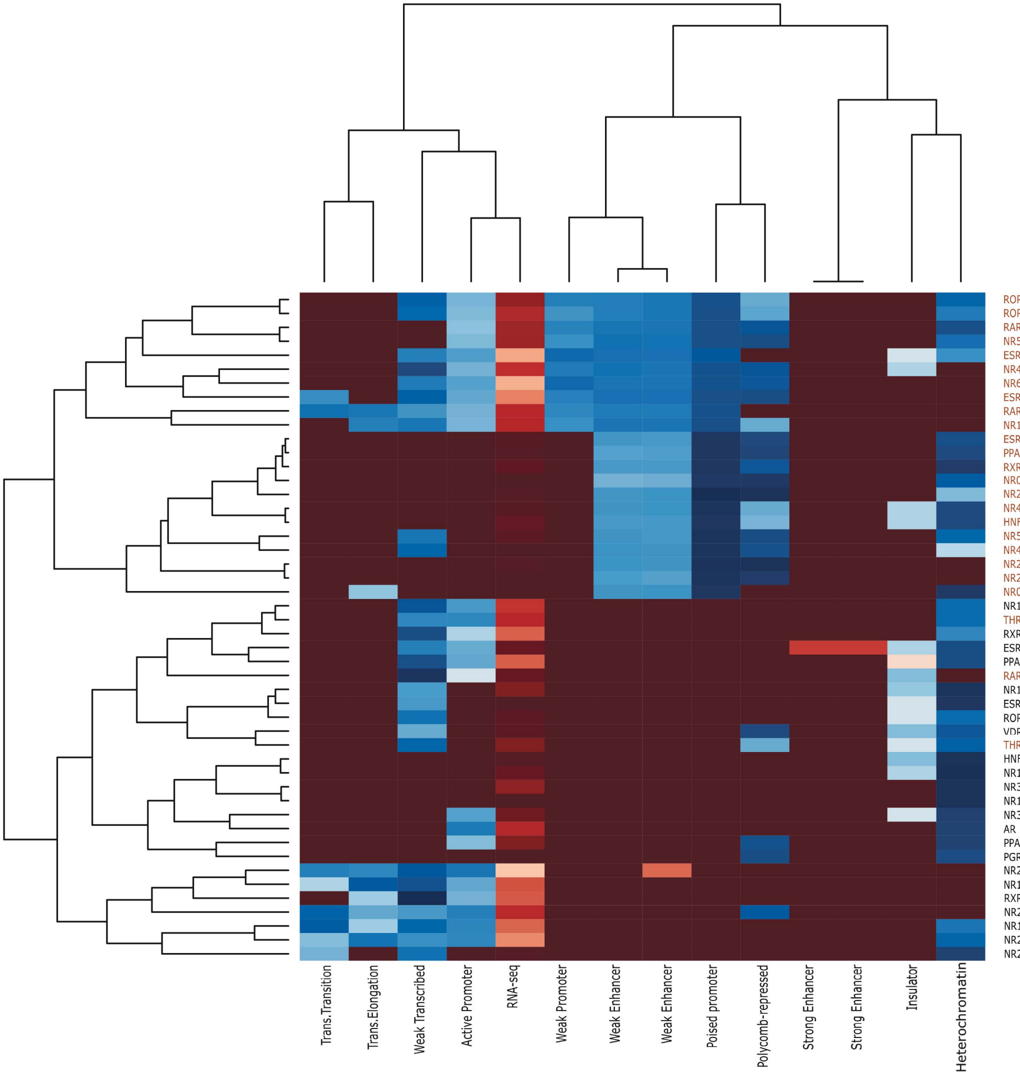


Figure 6. HMM state map analysis recovers the two clusters of nuclear receptor genes obtained using HCNE-based analysis. The columns of the heatmap show 13 different chromatin states alongwith RNA-seq data. The rows correspond to each nuclear receptor gene (Cluster 1 shown in red, Cluster 2 shown in black). The column and row side dendrogram represents the clusters of nuclear receptor genes on the basis of difference in their *cis*-regulatory functional elements and expression state.
doi:10.1371/journal.pone.0088880.g006

cluster 2 are not associated with poised promoter or enhancer-related marks regardless of their expression state. This novel result further confirms the differences in regulatory mechanisms between the genes belonging to two clusters, indicating that cluster 1 (representing genes that are possible targets of long-range regulation) are the only ones that rely on poised configuration for rapid activation of gene expression.

Discussion

Diverse functional roles of nuclear receptors and their direct/indirect involvement in physiological and developmental disorders and their potential as drug targets call for a better understanding of this important gene family. Insight into regulation mechanisms governing the transcription of nuclear receptor genes is central to this task. Further, this can provide clues towards the evolutionary history of nuclear receptors in question, e.g. recent paralogs sharing same mechanism of regulation are likely to have evolved through whole-genome duplication rather than tandem duplication. More fundamentally, analyzing the regulation mechanism for nuclear receptors can help decipher their diverse functional roles, and possibly accounting for genome variants found in their vicinity.

In this study, we investigated the properties of *cis*-regulatory environment of nuclear receptors towards understanding the diversity in their biological roles. The mode of transcription regulation of nuclear receptors is crucial for deciphering their function, which is not sufficiently captured by existing classifications of nuclear receptors based on their sequence homology [9] or mechanism of action.

Towards this goal, we have studied the *cis*-regulatory environment of each member of the gene family. We used the GRB model, which consists of target gene surrounded by highly conserved non-coding elements (HCNEs) and bystander genes, to analyze the neighborhood of each nuclear receptor gene. This allowed us to categorize nuclear receptors into two functional classes –25 nuclear receptors which we hypothesize to be targets of long-range regulation (cluster 1 in Table 1), and remaining 23 nuclear receptors which are not targets (cluster 2). We discuss our key findings below.

A number of developmental genes are present in cluster 1, including some that are known targets of long-range gene regulation. On the other hand, cluster 2 contain several genes which are tissue-specific and consequently do not utilize long-range regulation. Further, genes present in cluster 1 have longer and often multiple CpG islands, a known characteristic of target genes under the GRB model.

We have also identified cases of multiple nuclear receptors present in the same GRB locus (Figure S3). It is not unusual to have GRBs with multiple targets – HOX, IRX and DLX loci are known examples - and at least some GRB targets that occur in separate loci in vertebrates are found next to each other in e.g. *Drosophila* genome [28]. However, this makes it hard to predict which of the genes present in the same locus are being regulated. To address this, we used other promoter-related features, e.g. presence of bivalent domain, which are known to be present in genes having long-range regulation (Figure 5). Our analysis provides strong indication as to which genes are the targets of long-range regulation and therefore, can be used when investigating other GRBs with multiple targets.

To further validate our results, we have investigated the impact of different individual histone modifications. We found that genes present in cluster 1 have significantly higher enrichment of enhancer mark (H3K4me1) around their gene loci compared to genes in cluster 2 (Figure 4), indicating multiple enhancers including those overlapping HCNEs. Subsequent analysis of repressive marks (H3K27me3) reveals that several genes in cluster 1 have bivalent domain in their promoter regions (Figure 5). This provides further indication that these genes require spatio-temporal control of their transcription facilitated by gain/loss of active and repressive promoter marks. Further experimental study using time-series data can elucidate this phenomenon.

We also studied combinatorial patterns of histone modifications, which have been shown to capture functional dynamics associating with specific biological functions of the genome [33]. We note that our original categorization is recovered (except for two genes, see Results for details) using this approach, lending crucial evidence that long-range regulation (captured by our method) is key to the functional roles of more than half of the nuclear receptors.

Figure 7 presents our final classification of nuclear receptors into possible targets of long-range regulation (shown in red) and non-targets (shown in blue) taking into account presence of multiple targets in the same GRB loci. We show sequence-based similarity, highlighting the fact that new paralogs in evolution often acquire a different mode of regulation. Following further with above classification, investigation of evolutionary mechanism whereby the paralogs acquired different regulation is the logical next step. We expect nuclear receptors implicated to be targets of long-range regulation have likely evolved by whole genome duplication events, and therefore, retained their regulatory inputs over a wide region. In contrast, other nuclear receptors possibly evolved through more localized (tandem) duplications.

Materials and Methods

HCNE based Analysis and CpG Islands Detection

We have used the following genome assemblies for this study: human (hg19), mouse (mm10), chicken (galGal4), fugu (fr3) and zebrafish (Zv9). All the gene coordinates were obtained from Ensembl ([42]; http://www.ensembl.org; version 72) using Biomart (http://www.biomart.org). The associated scripts are available at http://www.bitbucket.org/yogita_sharma/nr_classification/.

The genomic coordinates of HCNEs were obtained from the Ancora genome browser ([43]; http://ancora.genereg.net). The selected conservation threshold and length cut offs for each species are specified in Table S1. The CpG island locations were downloaded from the UCSC Genome Table Browser ([44]; http://genome-euro.ucsc.edu/cgi-bin/hgTables?hgsid=194624867*)*. For each pair-wise comparison between human and one of the other genomes, we computed the HCNEs ±2 Mb region of each nuclear receptor gene loci. This is to capture *cis*-regulatory elements, which may occur far from the gene location.

The extension of genomic co-ordinates around each gene loci for HCNE detection might create biasness towards the longer genes. To avoid this we normalized the obtained HCNE counts with respect to the gene length. The log2 values of the HCNE counts were used to compute the dissimilarity matrix for all the genes across different five genomes (Euclidean distance measure). Finally we performed the hierarchical clustering, using complete linkage, to analyze the HCNEs across the gene set. This method is more robust to outliers compared to classification based on a single threshold such as mean etc.

The CpG island locations were downloaded from the UCSC Genome Table Browser [44]. For this analysis, we used three gene sets; nuclear receptors, transcription factors and CpG genes. The ±1 kb flanking region around all the genes were scanned to count the total number of CpG base pairs. Along with the calculation of CpG island number we also calculated the total CpG island lengths for the gene sets. The cumulative distributions of the CpG island length were plotted for all the genes.

We also compared the HCNE counts between nuclear receptors and other random selected transcription factors. We randomly selected 48 genes out of around 900 (Table S4, Sheet 2) using GNU R function sample() with default seed and burn-in of 500. We obtain transcription factor gene coordinates from the Ensembl database (version 72). To be able to compare between the different gene sets we pooled the randomly selected set of genes with the nuclear receptor gene family and repeated previous experiment. The HCNEs were calculated and plotted in the same way as in the previous experiment.

RNA-seq Data

The RPKM files for expression-based analysis (RNA-seq) was downloaded from ENCODE ([31]; http://genome-euro.ucsc.

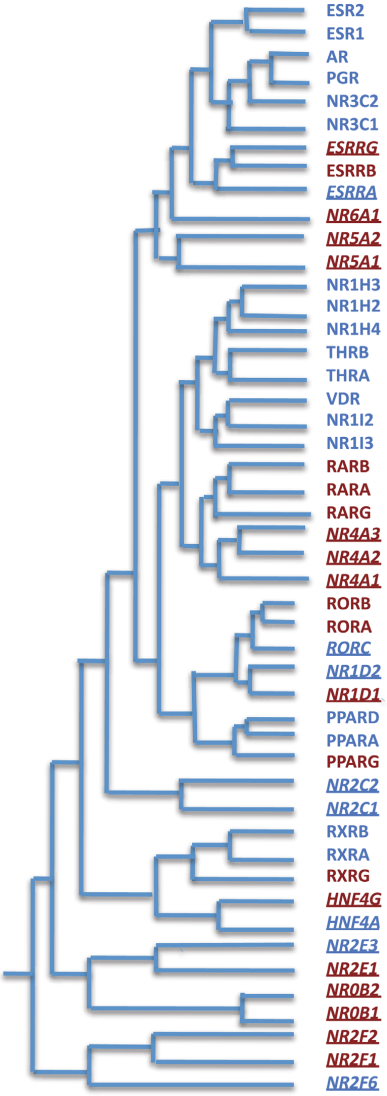


Figure 7. Classification comparison of nuclear receptors gene family with respect to sequence homology and transcriptional mechanism and function based. The GRB target genes (cluster 1 in Table 1) are shown in red, while non-targets are in blue. Nuclear hormone receptors are presented in normal bold text while orphan receptors are underlined and in italics. There are in total 23 nuclear receptor GRB target genes and 25 nuclear receptor non-GRB target nuclear receptor genes. It is clear from the figure that both GRB target and non-target nuclear receptors are dispersed among seven families classified on the basis of sequence homology.
doi:10.1371/journal.pone.0088880.g007

edu/ENCODE/downloads.html) for five cell lines (Gm12878, H1hesc, Huvec, HepG2, k562) for hg18 genome assembly.

ChIP-seq Data

The tag aligned files downloaded for five cell lines (Gm12878, H1hesc, Huvec, HepG2, k562) from hg18 genome assembly of ENCODE [31] project were used for the peak calling. We extracted the significant enriched regions between chip versus control using CCAT package [45]. Standardized settings (fragmentSize = 200, isStrandSensitiveMode = 0, slidingWinSize = 500, movingStep = 50, outputNum = 100000, minCount = 4, minScore = 3.0, bootstrapPass = 50, randSeed = 123456) were implemented for the analysis. Finally top 10,000 peaks (with p-value < 0.001) were used for further downstream analysis. After preprocessing the data set we extracted coverage (vector representing read per million values for each bin) across different genomic ranges of interest. To be able to compare across different cell lines we normalized the coverage across the dataset by dividing obtained coverage w.r.t. library size. Table S5 presents the genomic ranges used for analysis of different histone marks [46,47,48].

Statistical Significance Test for Enhancer Data

To check the significance of the difference obtained in enrichment of H3k4me1 mark across both clusters, we performed statistical testing against background set as follows: We extracted a set of 2054 genes (chromosome X in hg18 genome assembly) from Ensembl database using the R library (biomaRt). Subsequently, we classified this gene set based on presence of CpG island within ±1 kb region of transcription start site of each gene; obtaining a candidate set of 402 genes with CpG islands, and the remaining set of 1652 genes without CpG islands.

We constructed the background set consisting of 2054 genes obtained as described above as well as the set of 48 nuclear receptor genes, resulting in a total size of 2102 genes. We drew 1000 bootstrap samples from this set, and for each sample, we counted the number of reads overlapping regions of different width (±10 kb, ±1 Mb and ±2 Mb) around the transcription start site for each gene. This was used to construct background distribution of the number of reads for each of the different region widths (respectively ±10 kb, ±1 Mb and ±2 Mb).

We have also extracted a set of 1455 genes on chromosome 5 and classified the gene set in to CpG (650) and non-CpG genes (805) on the basis of presence/absence of CpG island. We performed the statistical analysis in the similar way as mentioned above.

Chromatin State Map Analysis

Chromatin state map is a hidden Markov model-based mapping of different chromatin states across the different cell lines [33]. The data was downloaded from UCSC genome browser [44]. Since we were interested to see the difference in regulatory content of developmental related and non-related nuclear receptor genes, we only considered the embryonic stem cell line (H1hesc) data for this analysis. We calculated the total number of state counts for each gene in all the states across selected genomic ranges in H1hesc. We used different random genomic ranges (±10 kb and ±100 kb around TSS) to study the enrichment of chromatin states. To see the combinatorial patterns of histone modifications around all genes we prepared a heatmap using log2 ratio of the number of state counts for each gene using the default parameters (Hierarchical clustering with full/complete linkage using Euclidean distance measure).

Supporting Information

Figure S1 Cumulative distribution plots of total CpG island length across three data sets. The GRB targets nuclear receptors have longer CpG islands than randomly selected CpG and transcription factor genes. The GRB target NR, random selected transcription factors and CpG genes are presented in green, red and black, respectively.
(EPS)

Figure S2 Clustering of genes based on HCNE counts in augmented set of nuclear receptors and randomly selected transcription factors. The nuclear receptors in cluster 1 (Table 1) are present in the same cluster here as well.
(EPS)

Figure S3 Cases of multiple targets present in same GRB locus. A) Block of three genes (*THRB, RARB* and *NR1D2*) in human on chromosome 3 and their 1-to-1 orthologs in mouse in chromosome 14. B) Block of three genes (*THRA, RARA* and *NR1D1*) in human on chromosome 17 and their 1-to-1 orthologs in mouse in chromosome 11. C) Block of two genes in human (*NR6A1, NR5A1*).
(EPS)

Figure S4 H3K4me3 average coverage plot for nuclear receptor genes in cluster 1 (putative targets of long-range regulation). The average H3K4me3 coverage plots around ±10 kb TSS across different cell lines when genes are expressed (left) and not expressed (right). The x-axis shows position around ±10 kb TSS and y-axis represent average coverage. It shows when genes are expressed they have peak of active promoter around their TSS. Different colors represent different cell lines.
(EPS)

Figure S5 H3K4me3 average coverage plots for nuclear receptor genes in cluster 2 (non-targets based on GRB model). The average H3K4me3 coverage plots around ±10 kb TSS across different cell lines when non-GRB target genes are expressed (left) and not expressed (right). The x-axis shows position around ±10 kb TSS and y-axis represent average coverage. Expressed genes have active promoter signal around their TSS. Different colors represent respective cell lines.
(EPS)

Figure S6 UCSC genome browser view of promoter region of selected five cases from Cluster 1 genes. The promoter region of five (*NR4A1, NR5A2, NR1D1, RORB* and *ESRRG*) genes around ±5 KB TSS. The direction of arrow represents transcription direction. The first peak corresponds to active transcription (H3K4me3) followed by the peak of repression mark (H3K27me3) in the track below. CpG islands are shown in green.
(EPS)

Figure S7 Average coverage plots of repression mark (H3k27me3) around different clusters. The x-axis shows position around ±10 kb TSS and y-axis coverage. Cluster 1 (red color) has higher coverage of repression mark in comparison to cluster 2 (green color). The blue line represents TSS.
(EPS)

Figure S8 Statistical significance test for H3K4me1 around different genomic distributions on chromosome 5. A) H3K4me1 distribution in different clusters across ±10 kb TSS against the random background distribution. B) H3K4me1 distribution in different clusters across ±1 Mb TSS with respect to random background distribution. C) H3k4me1 distribution in different clusters across ±2 Mb TSS with respect to random background distribution. This figure shows that cluster 1 (shown by red bar) has significantly higher distribution of H3K4me1 in comparison to random selected background region (marked by black bars), CpG and non-CpG region (shown by blue and green bar respectively) and cluster 2 genes (shown by pink bar).
(EPS)

Table S1 The percentage of conservation and length cut offs for HCNE counts.
(DOC)

Table S2 The list of genes in HCNE based clustering of augmented set consisting of 48 nuclear receptors and 48 randomly selected transcription factors. Known targets of long-range gene regulation are marked with asterisk (*).
(DOC)

Table S3 The RPKM values of each nuclear receptor gene across 5 cell lines.
(XLS)

Table S4 List of HMM states associated with specific functional elements of the genome.
(XLS)

Table S5 The genomic ranges for different histone modifications.
(DOC)

Acknowledgments

YS sincerely thanks Christopher Previti for general discussion, Vinay Jethava for critical review and revision of the manuscript, and Gemma Danks for careful reading and language editing of the manuscript.

Author Contributions

Conceived and designed the experiments: YS BL MB. Performed the experiments: YS. Analyzed the data: YS. Contributed reagents/materials/analysis tools: YS CC. Wrote the paper: YS BL. Processed the data for HMM state map analysis: CSRC. Critically read and approved the final manuscript: YS CSRC MB BL.

References

1. Olefsky JM (2001) Nuclear receptor minireview series. J Biol Chem 276: 36863–36864.
2. Gronemeyer H, Gustafsson JA, Laudet V (2004) Principles for modulation of the nuclear receptor superfamily. Nat Rev Drug Discov 3: 950–964.
3. Robinson-Rechavi M, Escriva Garcia H, Laudet V (2003) The nuclear receptor superfamily. J Cell Sci 116: 585–586.
4. Klinge CM (2000) Estrogen receptor interaction with co-activators and co-repressors. Steroids 65: 227–251.
5. Linja MJ, Porkka KP, Kang Z, Savinainen KJ, Janne OA, et al. (2004) Expression of androgen receptor coregulators in prostate cancer. Clin Cancer Res 10: 1032–1040.
6. Serpente P, Tumpel S, Ghyselinck NB, Niederreither K, Wiedemann LM, et al. (2005) Direct crossregulation between retinoic acid receptor {beta} and Hox genes during hindbrain segmentation. Development 132: 503–513.
7. Tobin JF, Freedman LP (2006) Nuclear receptors as drug targets in metabolic diseases: new approaches to therapy. Trends Endocrinol Metab 17: 284–290.
8. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993–996.
9. Owen GI, Zelent A (2000) Origins and evolutionary diversification of the nuclear receptor superfamily. Cell Mol Life Sci 57: 809–827.
10. Laudet V, Hanni C, Coll J, Catzeflis F, Stehelin D (1992) Evolution of the nuclear receptor gene superfamily. EMBO J 11: 1003–1013.

11. Jetten AM, Kurebayashi S, Ueda E (2001) The ROR nuclear orphan receptor subfamily: critical regulators of multiple biological processes. Prog Nucleic Acid Res Mol Biol 69: 205–247.

12. Huq MD, Wei LN (2005) Post-translational modification of nuclear co-repressor receptor-interacting protein 140 by acetylation. Mol Cell Proteomics 4: 975–983.

13. McMorrow JP, Murphy EP (2011) Inflammation: a role for NR4A orphan nuclear receptors? Biochem Soc Trans 39: 688–693.

14. Park SP, Hong IH, Tsang SH, Lee W, Horowitz J, et al. (2013) Disruption of the human cone photoreceptor mosaic from a defect in NR2E3 transcription factor function in young adults. Graefes Arch Clin Exp Ophthalmol 251: 2299–2309.

15. Takeda Y, Liu X, Sumiyoshi M, Matsushima A, Shimohigashi M, et al. (2009) Placenta expressing the greatest quantity of bisphenol A receptor ERR{gamma} among the human reproductive tissues: Predominant expression of type-1 ERRgamma isoform. J Biochem 146: 113–122.

16. Tomassy GS, De Leonibus E, Jabaudon D, Lodato S, Alfano C, et al. (2010) Area-specific temporal control of corticospinal motor neuron differentiation by COUP-TFI. Proc Natl Acad Sci U S A 107: 3576–3581.

17. Kumar R, Thompson EB (1999) The structure of the nuclear hormone receptors. Steroids 64: 310–319.

18. Lee SK, Jung SY, Kim YS, Na SY, Lee YC, et al. (2001) Two distinct nuclear receptor-interaction domains and CREB-binding protein-dependent transactivation function of activating signal cointegrator-2. Mol Endocrinol 15: 241–254.

19. Wolf IM, Heitzer MD, Grubisha M, DeFranco DB (2008) Coactivators and nuclear receptor transactivation. J Cell Biochem 104: 1580–1586.

20. Pascual G, Glass CK (2006) Nuclear receptors versus inflammation: mechanisms of transrepression. Trends Endocrinol Metab 17: 321–327.

21. Zhang Z, Burch PE, Cooney AJ, Lanz RB, Pereira FA, et al. (2004) Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. Genome Res 14: 580–590.

22. Abu-Hayyeh S, Papacleovoulou G, Williamson C (2013) Nuclear receptors, bile acids and cholesterol homeostasis series - bile acids and pregnancy. Mol Cell Endocrinol 368: 120–128.

23. Biddie SC (2011) Chromatin architecture and the regulation of nuclear receptor inducible transcription. J Neuroendocrinol 23: 94–106.

24. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, et al. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5: 99.

25. Akalin A, Fredman D, Arner E, Dong X, Bryne JC, et al. (2009) Transcriptional features of genomic regulatory blocks. Genome Biol 10: R38.

26. Sáez PJ, Lange S, Pérez-Acle T, Owen GI (2010) Nuclear Receptor Genes: Evolution. eLS.

27. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, et al. (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 17: 545–555.

28. Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17: 1898–1908.

29. Ragvin A, Moro E, Fredman D, Navratilova P, Drivenes O, et al. (2010) Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. Proc Natl Acad Sci U S A 107: 775–780.

30. Navratilova P, Becker TS (2009) Genomic regulatory blocks in vertebrates and implications in human disease. Brief Funct Genomic Proteomic 8: 333–342.

31. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

32. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817–825.

33. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49.

34. Kikuta H, Fredman D, Rinkwitz S, Lenhard B, Becker TS (2007) Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. Genome Biol 8 Suppl 1: S4.

35. Hokamp K, McLysaght A, Wolfe KH (2003) The 2R hypothesis and the human genome sequence. J Struct Funct Genomics 3: 95–110.

36. Gao F, Wei Z, An W, Wang K, Lu W (2013) The interactomes of POU5F1 and SOX2 enhancers in human embryonic stem cells. Sci Rep 3: 1588.

37. Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, et al. (2010) Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. Nucleic Acids Res 38: 1071–1085.

38. Pfeffer PL, Payer B, Reim G, di Magliano MP, Busslinger M (2002) The activation and maintenance of Pax2 expression at the mid-hindbrain boundary is controlled by separate enhancers. Development 129: 307–318.

39. Ramskold D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol 5: e1000598.

40. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13: 233–245.

41. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315–326.

42. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, et al. (2010) Ensembl's 10th year. Nucleic Acids Res 38: D557–562.

43. Engstrom PG, Fredman D, Lenhard B (2008) Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. Genome Biol 9: R34.

44. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37: D755–761.

45. Xu H, Handoko L, Wei X, Ye C, Sheng J, et al. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics 26: 1199–1204.

46. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12: 7–18.

47. Zentner GE, Tesar PJ, Scacheri PC (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res 21: 1273–1283.

48. King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, et al. (2007) Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res 17: 775–786.

II

BMC Bioinformatics

Research

Open Access

Translog, a web browser for studying the expression divergence of homologous genes

Xianjun Dong*[1,2], Altuna Akalin[1,2], Yogita Sharma[1] and Boris Lenhard[1,2]

Addresses: [1]Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway and [2]Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway

E-mail: Xianjun Dong* - xianjun.dong@bccs.uib.no; Altuna Akalin - altuna.akalin@bccs.uib.no;
Yogita Sharma - yogita.sharma@biomed.uib.no; Boris Lenhard - boris.lenhard@bccs.uib.no
*Corresponding author

from The Eighth Asia Pacific Bioinformatics Conference (APBC 2010)
Bangalore, India 18-21 January 2010

Published: 18 January 2010

BMC Bioinformatics 2010, **11**(Suppl 1):S59 doi: 10.1186/1471-2105-11-S1-S59

This article is available from: http://www.biomedcentral.com/1471-2105/11/S1/S59

© 2010 Dong et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Increasing amount of data from comparative genomics, and newly developed technologies producing accurate gene expression data facilitate the study of the expression divergence of homologous genes. Previous studies have individually highlighted factors that contribute to the expression divergence of duplicate genes, e.g. promoter changes, exon structure heterogeneity, asymmetric histone modifications and genomic neighborhood conservation. However, there is a lack of a tool to integrate multiple factors and visualize their variety among homologous genes in a straightforward way.

Results: We introduce Translog (a web-based tool for Transcriptome comparison of homologous genes) that assists in the comparison of homologous genes by displaying the loci in three different views: promoter view for studying the sharing/turnover of transcription initiations, exon structure for displaying the exon-intron structure changes, and genomic neighborhood to show the macro-synteny conservation in a larger scale. CAGE data for transcription initiation are mapped for each transcript and can be used to study transcription turnover and expression changes. Alignment anchors between homologous loci can be used to define the precise homologous transcripts. We demonstrate how these views can be used to visualize the changes of homologous genes during evolution, particularly after the 2R and 3R whole genome duplication.

Conclusion: We have developed a web-based tool for assisting in the transcriptome comparison of homologous genes, facilitating the study of expression divergence.

Background

One of the challenges in the post-genomic era is to understand the mechanisms which drive the divergence of gene expression, and how this causes phenotypic changes, ultimately leading to the evolution of new species [1-6]. This is important both at the level of orthologs (genes separated by a speciation even) and paralogs (genes separated by a duplication event). For example, *PAX6*, the most studied Pax gene, is a "master control" gene for the development of eyes and sensory organs, and other homologous structures, usually derived from ectodermal tissues [7]. Its protein function is highly conserved across bilaterian species: mouse *PAX6* can trigger eye development in *D. melanogaster* [8]. However, genomic organization of genes sharing the ancestry with the human *PAX6* and its immediate neighborhood varies considerably among species, with differences in the number and distribution of exons, *cis*-regulatory elements and transcription start sites. For paralogous genes, derived from gene duplication or whole genome duplication, it has been shown that duplicate genes increase expression divergence and enable tissue or developmental specialization to evolve, as shown in mammals [9], fish [10], worm[10], yeast[11], and plants[12]. By comparing the transcription patterns of duplicate genes, we can often trace the factors that influence the expression pattern changes in evolution.

At the genomic level, previous studies have focused on examining the relationship between the divergence of gene expression and type of the promoter[13], exon structure[14], TSS turnover[14], genomic neighborhood [15], cis-regulatory inputs [16], histone modifications [17], and recently, the DNA-encoded nucleosome organization of promoters , possibly further complicated by external environmental factors are involved [18].

The increasing volume of available transcriptome data such as CAGE[19] and RNA-seq [20] for different developmental stages and tissues for different species can be harnessed to understand the mechanisms of spatiotemporal expression changes of genes that share a (not so ancient) common ancestor. The investigation should start with the integrated analysis of the available data. A suitable tool for this type of analysis should enable the comparison of homologous genes on different scales, from the position and activity of their proximal promoters to the corresponding information on their long-range regulatory inputs. Similar tools, like the comparative genome viewer in DBTSS[21], also contribute to compare the promoter and transcripts for homolog genes, but they don't use high-throughput sequencing like CAGE and their visualization methods are not so enhanced. In this paper, we describe Translog (the tool for **Trans**criptome comparison of homo**logs**) [22], a web-based application providing 1) a *promoter* view where a region containing all proximal promoters of a gene's transcript(s) is aligned to its homolog and cross-mapped between the two loci using alignment anchors, 2) a *gene structure* view where a gene's exon-intron structure is compared to that of its homolog, alongside its transcriptional features, and 3) a *genomic neighborhood* view which displays the neighbors of a gene in a large flanking region, and show their conservation in the homologous loci. CAGE data is displayed along with the genomic features to indicate the expression of transcripts. We demonstrate how Translog can be used to discover and visualize homologous relationships, expression pattern changes after duplication or speciation, and to explore the divergences of promoter usage, gene structure and genomic neighborhood between two homologs. We anticipate that Translog will be useful in looking for the factors of impacting expression divergence between two homologous genes, and finally contribute to understanding the mechanism of evolution of gene expression.

Methods

Gen(om)e annotations

To define the promoter region and gene structure, we use the gene name and genomic locations of all Ensembl genes and transcripts from Ensembl v52 [23]. Currently, these include three genomes (human, mouse and zebrafish): 25233 genes in *D. rerio* (assembly version 7), 37436 in *H. sapiens* (assembly NCBI Build 36.1), and 31805 in *M. musculus* (assembly NCBI Build 37). We use these three species because i) there is CAGE data available for them, and ii) comparison of human: human, zebrafish:zebrafish paralogs can reveal the expression changes along with 1R/2R, 3R whole genome duplication, respectively. The orthologs (human:mouse, human:zebrafish, human:tetraodon) and paralogs (human:human) were downloaded from Ensembl Compara v52 [23], using BioMart[24]. For zebrafish:zebrafish paralogs, instead of taking all paralogs from Ensembl, we are primarily interested in those duplicates arisen in the event of fish-specific WGD. For the latter, we used human:zebrafish orthologs as a bridge (approximating ancestral genome before WGD) to extract those zebrafish gene pairs which have the same human ortholog genes. RefSeq genes were downloaded from refGene table in UCSC Table Browser (on 2009-08-01) for each genome.

Defining TSS using CAGE tag clusters

In order to define CAGE TSSes and clusters, we used all publicly available CAGE tags (from http://fantom.gsc.riken.jp/4/download/, [25]) for human (hg18) and mouse (mm9). We used only uniquely mapping tags and clustered CAGE tags into tag clusters (TCs) if the member tags map to the same chromosome strand and overlap by at least 1 bp. For each TC we defined a representative location (as that supported by the highest number of tags). Afterwards, we grouped TCs into Sharp or Broad promoters using previous classification algorithm [26]. TCs are mapped to Ensembl genes on the [-500 bp, +500 bp] region around Ensembl TSS. If multiple TCs map to a given region, the one with the highest number of tags per million (tpm) is selected as representative TC for the gene.

Alignments of homologous loci

To align two homologous loci, we used UCSC chain and net alignment data [27], which is a whole genome alignment by blastz [28]. Any alignment block in the UCSC chain database is taken as an anchor to link two loci. If a region in the reference species aligns to only one locus in the target species, we denoted it as a 1-to-1 anchor; otherwise, we extracted the overlapping parts of *M* (two or more) anchors and defined as 1-to-*M* anchor. For those having many ($M > 2$) aligned loci (e.g. genes by tandem duplication or from a large protein family), we only took the two highest scoring ones and display them as 1-to-2 anchors. The 1-to-2 mammal:zebrafish orthologs originating from teleost whole-genome duplication are expected to have 1-to-2 anchors. To distinguish the anchors from different scenarios, we marked them in different colors (by default, 1-to-1 anchors in gray, and 1-to-2 in blue: see Figure 1).

For human:human homolog comparison, we used the UCSC selfChain alignment to generate the anchors for paralogous loci. If no selfChain anchors are found, a link to Ensembl clustalW alignment is given. For zebrafish: zebrafish, we used human:zebrafish 1-to-2 chain alignment as a bridge to get zebrafish:zebrafish alignment due to the absence of zebrafish selfChain data at present. This method cannot detect the region only conserved between two zebrafish loci, but not conserved in human, for example those fast evolving regions specific in human lineage [29]; on the other hand, it can provide insight into the probable ancestral state of the locus[6].

Results and discussion

Identification of homologous genes

We extracted an initial homolog set from Ensembl Compara [30,31]. Out of 21416 human protein-coding genes, 79% and 51% have orthologs in mouse and zebrafish respectively. There are 29721 human:human paralog pair combinations altogether. To investigate how many of them are duplicates from 1R/2R WGD we grouped the paralogs by their last common ancestor. As shown in Figure S1, the largest category, which includes ~8400 human duplicates, falls in the time span before the split between bony fish (e.g. zebrafish) and tetrapods, and after the split between lancelets and jawless fish. This corresponds well with the proposed 1R/2R WGD timing (see Figure S1 in Additional file 1).

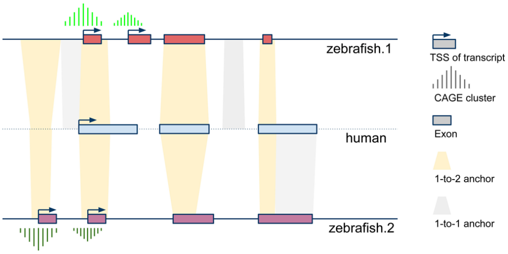


Figure 1
The principle of comparing homologous genes using alignment anchors.

Out of all human:zebrafish 1-to-2 orthologous genes, we wanted to determine how many date from teleost-specific WGD (3R WGD in Figure S1 in Additional file 1). To exclude the cases which have arisen by zebrafish-specific tandem duplications, ideally we should infer it from phylogenetic tree. A recent study [32], which identified gene duplicates retained from the last, teleost-specific WGD, found 615 human:zebrafish orthologs from the teleost WGD with high or medium confidence; most (94%) of them are included in the 1-to-2 orthologs we have defined here.

To study the expression divergence and differential promoter use of orthologous genes, we mapped CAGE tag clusters (TCs) to the human and mouse Ensembl genes. Most of the CAGE tags have a corresponding tissue in mouse and human in which they were detected. Only 7 out of 55 of those tissues in mouse do not have corresponding human tissue, whereas all the human tissues have corresponding mouse tissues (see Table S1 in Additional file 1). If multiple TCs map to one gene, the TC with the highest expression is chosen as representative TC. ~90% of the 1-to-1 orthologous gene pairs (13895 pairs in total) have at least one TC associated with them in both species.

Comparing transcriptional initiation in homologous genes using Translog

Users can compare homologous genes and their CAGE data in three different views (Promoter, Gene structure, and Genomic neighborhood, see Figure 2A) through the links in the top-left corner of the Translog start page (see Figure 2B). The 'Promoter' view shows a region covering all transcription start sites (from both Ensembl transcripts and RefSeq genes) and extends 500 bp upstream and downstream (Figure 2B). The 'Gene structure' view shows the exon-intron structures of the pair of homologs (Figure 2C). The 'Genomic neighborhood' view shows the conservation of the query gene and its neighborhoods using the anchor of gene homology (Figure 2D). Translog currently supports comparison of human: human, human:mouse, human:zebrafish and zebrafish:

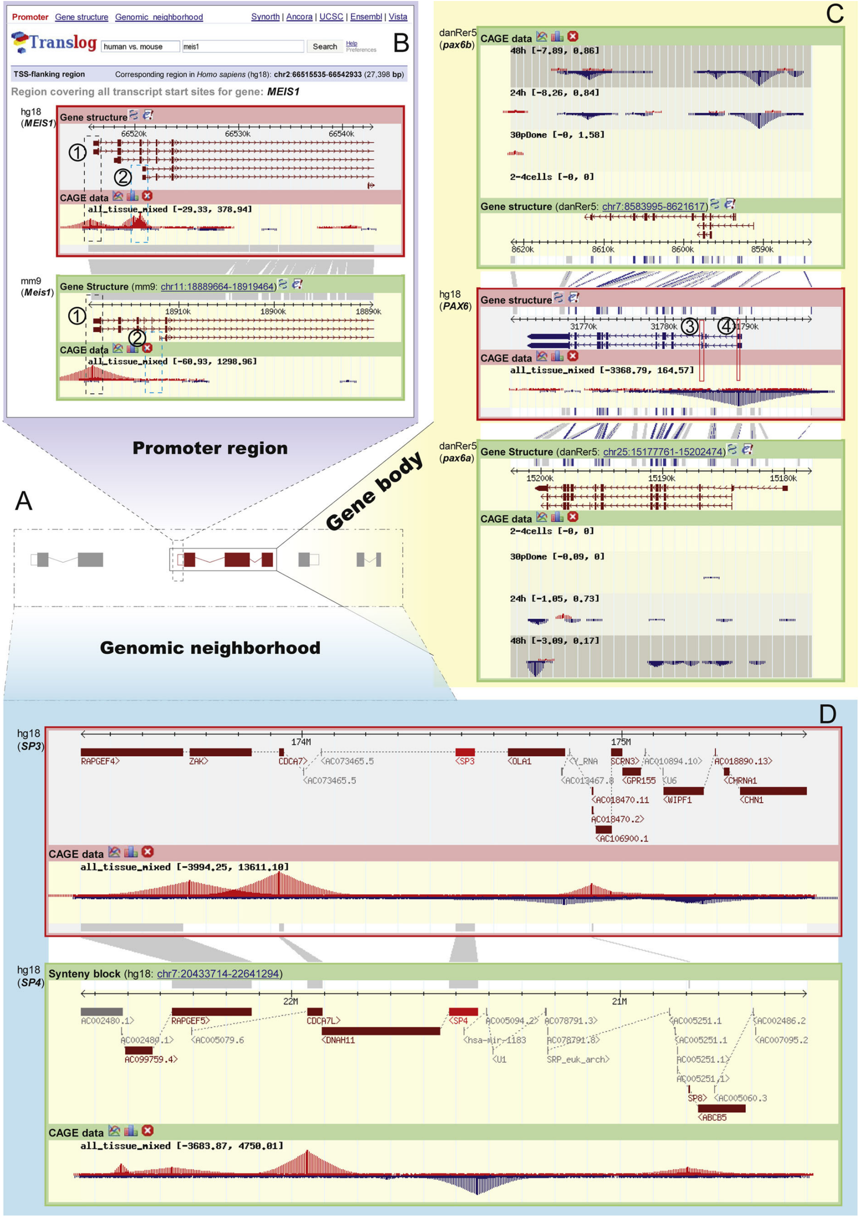


Figure 2

Translog structure and three views modes. A) Definition of three views in Translog. B) Promoter view of the *MEIS1* (human) vs. *Meis1* (mouse). C) Gene structure view, with *PAX6* (human) vs. *pax6a/b* (zebrafish) comparison. D) Genomic neighborhood view, with *SP3* (human) vs. *SP4* (human) comparison.

zebrafish (not for zebrafish CAGE data right now). We aim to expand this list in the future to provide other perspectives or enable the study of other instances of whole-genome duplications, after the suitable genome assemblies and expression data become available. The first on the list are the new Zv8 zebrafish genome alignment data (whose annotation is still incomplete at present) and the upcoming zebrafish CAGE data, followed by the lamprey genome for studying the 2R whole genome duplication directly.

Basic usage

For any supported query identifier (Ensembl gene ID, HGNC gene symbol or gene synonyms), the browser shows CAGE data and corresponding genomic features relevant to the input query. In the right corner of each page there are links to the external resources (e.g. Synorth [33], Ancora [34], UCSC Browser [27], Ensembl [30] and Vista[35]) for each displayed region.

Promoter view

We define a (proximal) promoter region as a 1000 bp region centering on the TSS of a transcript. A genomic region spanned by the union of all promoter regions for the query gene is displayed in the reference genome, same for the target genome. Alignment anchors (if any) are also displayed linking the two loci, which can assist the user in mapping the homology of transcription start sites. This is particularly useful if a gene has several transcripts with different TSSes. For example, in Figure 2B, the human gene *MEIS1* has 6 transcription isoforms with four different TSSes while its mouse ortholog *Meis1* has 3 transcripts with two different TSSes. Most of the TSSes are covered by CAGE tag cluster, with different peak heights (corresponding to tissue-weighted expression level).

After we align the transcripts of the two genes by the alignment anchors (the gray bar between the red frame and green frame in Figure 2B), we can inspect the sharing and turnover of TSSes between the homologous transcripts. For example, the leftmost transcripts (the black dotted frame ① in Figure 2B) of the two *Meis1* genes apparently share a CAGE cluster, indicating a shared ancestry of this particular promoter. On the other hand, the transcript with strongest expression (② in Figure 2B) in human *MEIS1* does not have a CAGE cluster in the same position in its orthologous transcript in mouse. Looking at the difference in peak heights of each CAGE cluster, we can spot cases in which the most highly expressed transcript in one gene is not always the most expressed one in its ortholog. Compared to the methods used by previous studies (e.g. [14]), Translog can be used to define pairs of homologous transcripts more precisely.

Gene structure view

In this view, we define a transcript region as a region containing a transcript and a 500 bp flanking region both upstream and downstream of it. Analogously to the Promoter view, a region spanning the union of all transcripts for the query gene will be displayed for both loci, along with the anchors connecting them. By linking two homologous genes with alignment anchors, this view can be used to distinguish the structural heterogeneity of the coding region and pinpoint major differences in intron-exon structure and splice form usage between related genes. Figure 2C shows human *PAX6* locus along with its two zebrafish not all human *PAX6* exons are conserved in zebrafish; the 4th exon of *ENST00000379123* is not conserved at all, while its second exon is only conserved in zebrafish *pax6a*.

After assigning the CAGE cluster to its transcript, the user can also investigate the relationship of expression changes and exon structure heterogeneity between homologous genes. Park et al.[14] classified each pair of duplicate genes into one of two structural categories: completely similar and incompletely similar. The latter were further classified in one of the three non-overlapping groups: 5′ similar, 3′ similar, and neither 5′ nor 3′ similar, with different extent of expression correlation. Using the 'Gene structure' view in Translog, the study of these kinds of correlations can be enhanced by quantifying the exon structure similarity only for those transcript pairs with shared TSS, instead of classifying them into a limited number of categories.

Genomic neighborhood view

This view displays the gene contents and CAGE data in a wider region around the query gene (see Methods). For human:zebrafish, we used the synteny blocks from[36]. For comparisons whose split events are too close (e.g. human:mouse ortholog from ~80 Myr ago) or too far (e. g. human:human paralogs from 1R/2R WGD ~550 Myr ago), we used a 2 Mb region centering on the query gene and its homologs. This view can be used to detect the synteny blocks dating from ancient segmental or whole genome duplications. For example, three genes in the human *SP3* gene locus (*RAPGEF4*, *CDCA7* and *AC018470.2* [synonym of *SP9*]) also have paralogs next to *SP4* (*RAPGEF5*, *CDCA7L* and *SP8*, respectively; see Figure 2D), with conserved gene order and orientation. This indicates that *SP3* and its paralogous gene *SP4* are not a consequence of the SP gene family expansion, but rather from duplications of whole loci, most likely whole genome duplications. Moreover, we found that the neighborhood synteny for *SP3* is also conserved in another paralogous locus in which the arrangement of ZAK, *SP3* and *SP9* is mirrored by their paralogs *ZPK*, *SP1* and *SP7*, respectively (Figure S2 in Additional file 1). This suggests that *SP3* and *SP1* neighborhoods arose from another duplication event, distinct of that that separated the ancestral neighborhoods of SP3 and SP4. But which duplication event occurred first? A previous study [37] found that *SP1* and *SP3* are more closely related to each other, and that their common ancestor was split from the ancestral form of SP4 by an earlier duplications. However, according to the picture (Figure S2 in Additional file 1) that we get from the synteny data, a more parsimonious explanation is that the gene content of *SP1* and *SP4* neighborhoods are from the result of a complementary gene loss after a recent duplication, while *SP3* locus is an out-group to them.

Conclusion

Translog is designed for studying the gene expression divergence of homologous genes across vertebrate genomes or paralogous loci within a genome. Based on the homology and CAGE expression data available for human, mouse and zebrafish, it provides a genome browser for visualizing and assessing the difference between homologous genes, on three different levels: promoter usage, gene structure changes, and genomic neighborhood conservation. One of the novel features of Translog is the possibility to display the comparison of two genomic loci in one browser by using alignment anchors. CAGE data is used to identify the true transcription start sites, measure the expression strength, and define the turnover or shift of promoter usage between homologous features. We anticipate that Translog will be highly useful for examining the factors that influence expression divergence between homologous genes.

List of abbreviations used

WGD: whole genome duplication; GRB: genomic regulatory block; HCNE: highly conserved non-coding element; TSS: transcription start site; CAGE: cap analysis gene expression; bp: base pair; Myr: million years.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

XD and BL designed the study. XD analyzed the data set, designed the Translog web resource and the underlying database, and generated examples and figures for the manuscript. AA prepared the CAGE data set and analyzed the data set. XD, AA, YS and BL wrote the manuscript.

Additional material

Additional file 1

Supplementary Table S1 and Figures S1, S2*. A document file contains supplementary Table S1 and Figures S1, S2.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-S1-S59-S1.pdf]

Acknowledgements

The work was supported by grants from the Young Future Research Leaders (YFF) program of the Norwegian Research Council (NFR) and Bergen Research Foundation (BFS) awarded to BL. We thank Gemma Danks for comments and proofreading, Ying Sheng for helps in preparing for the CAGE library. We thank Jan Christian Bryne, Supat Thongjuea and Christopher Previti for comments on the browser.

This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 1, 2010: Selected articles from the Eighth Asia-Pacific Bioinformatics Conference (APBC 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S1.

References

1. King MC and Wilson AC: **Evolution at two levels in humans and chimpanzees.** *Science* 1975, **188:**107–116.
2. White KP: **Functional genomics and the study of development, variation and evolution.** *Nat Rev Genet* 2001, **2:**528–537.
3. Carroll SB: **Genetics and the making of Homo sapiens.** *Nature* 2003, **422:**849–857.
4. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV and Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20:**1377–1419.
5. West-Eberhard MJ: **Developmental plasticity and the origin of species differences.** *Proc Natl Acad Sci USA* 2005, **102(Suppl 1):** 6543–6549.
6. Khaitovich P, Enard W, Lachmann M and Paabo S: **Evolution of primate gene expression.** *Nat Rev Genet* 2006, **7:**693–702.
7. Callaerts P, Halder G and Gehring WJ: **PAX-6 in development and evolution.** *Annu Rev Neurosci* 1997, **20:**483–532.
8. Gehring WJ: **New perspectives on eye development and the evolution of eyes and photoreceptors.** *J Hered* 2005, **96:**171–184.
9. Makova KD and Li WH: **Divergence in the spatial pattern of gene expression between human duplicate genes.** *Genome Res* 2003, **13:**1638–1645.
10. Levy SF and Siegal ML: **Network hubs buffer environmental variation in Saccharomyces cerevisiae.** *PLoS Biol* 2008, **6:**e264.
11. Gu X, Zhang Z and Huang W: **Rapid evolution of expression and regulatory divergences after yeast gene duplication.** *Proc Natl Acad Sci USA* 2005, **102:**707–712.
12. Ha M, Kim ED and Chen ZJ: **Duplicate genes increase expression diversity in closely related species and allopolyploids.** *Proc Natl Acad Sci USA* 2009, **106:**2295–2300.
13. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y and Semple CA: **Heterotachy in mammalian promoter evolution.** *PLoS Genet* 2006, **2:**e30.
14. Park C and Makova KD: **Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes.** *Genome Biol* 2009, **10:**R10.
15. De S, Teichmann SA and Babu MM: **The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome.** *Genome Res* 2009, **19:**785–794.
16. Papp B, Pal C and Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast.** *Trends Genet* 2003, **19:**417–422.
17. Zheng D: **Asymmetric histone modifications between the original and derived loci of human segmental duplications.** *Genome Biol* 2008, **9:**R105.

18. Ha M, Li WH and Chen ZJ: **External factors accelerate expression divergence between duplicate genes.** *Trends Genet* 2007, **23:**162–166.
19. de Hoon M and Hayashizaki Y: **Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference.** *Biotechniques* 2008, **44:**627–628, 630, 632.
20. Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621–628.
21. Wakaguri H, Yamashita R, Suzuki Y, Sugano S and Nakai K: **DBTSS: database of transcription start sites, progress report 2008.** *Nucleic Acids Res* 2008, **36:**D97–101.
22. **Translog.** http://translog.genereg.net.
23. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F and Cutts T, *et al*: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36:**D707–714.
24. Haider S, Ballester B, Smedley D, Zhang J, Rice P and Kasprzyk A: **BioMart Central Portal–unified access to biological data.** *Nucleic Acids Res* 2009, **37:**W23–27.
25. Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y and de Hoon MJ, *et al*: **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nat Genet* 2009, **41:**553–562.
26. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y and Hume DA: **Mammalian RNA polymerase II core promoters: insights from genome-wide studies.** *Nat Rev Genet* 2007, **8:**424–436.
27. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A and Pheasant M, *et al*: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, **37:**D755–761.
28. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D and Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13:**103–107.
29. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G and Baertsch R, *et al*: **Forces shaping the fastest evolving regions in the human genome.** *PLoS Genet* 2006, **2:**e168.
30. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P and Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37:**D690–697.
31. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R and Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res* 2009, **19:**327–335.
32. Kassahn KS, Dang VT, Wilkins SJ, Perkins AC and Ragan MA: **Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates.** *Genome Res* 2009, **19:**1404–1418.
33. Dong X, Fredman D and Lenhard B: **Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes.** *Genome Biol* 2009, **10:**R86.
34. Engstrom PG, Fredman D and Lenhard B: **Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes.** *Genome Biol* 2008, **9:**R34.
35. Visel A, Minovitsky S, Dubchak I and Pennacchio LA: **VISTA Enhancer Browser–a database of tissue-specific human enhancers.** *Nucleic Acids Res* 2007, **35:**D88–92.
36. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I and Howe K, *et al*: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.** *Genome Res* 2007, **17:**545–555.
37. Kolell KJ and Crawford DL: **Evolution of Sp transcription factors.** *Mol Biol Evol* 2002, **19:**216–222.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

References

Akalin,A. *et al.* (2009) Transcriptional features of genomic regulatory blocks. *Genome biology*, **10**, R38.

Anderson,R.P. and Roth,J.R. (1977) Tandem genetic duplications in phage and bacteria. *Annu. Rev. Microbiol.*, **31**, 473–505.

Arisue,N. *et al.* (2007) Phylogeny and evolution of the SERA multigene family in the genus Plasmodium. *J. Mol. Evol.*, **65**, 82–91.

Becker, TS and Lenhard, B (2007) The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Molecular Genetics and Genomics*.

Bookout,A. *et al.* (2006) Anatomical Profiling of Nuclear Receptor Expression Reveals a Hierarchical Transcriptional Network. *Cell*, **126**, 789–799.

Bridgham,J. *et al.* (2010) Protein Evolution by Molecular Tinkering: Diversification of the Nuclear Receptor Superfamily from a Ligand-Dependent Ancestor. *PLoS Biology*.

Brown, CJ *et al.* (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution*.

Davis,J.C. and Petrov,D.A. (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.*, **21**, 548–51.

De,S. *et al.* (2009) The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome research*, **19**, 785–794.

Dong,X. *et al.* (2010) Translog, a web browser for studying the expression divergence of homologous genes. *BMC bioinformatics*, **11**, S59.

Durand,D. and Hoberman,R. (2006) Diagnosing duplications – can it be done? *Trends in Genetics*, **22**, 156164.

Eichler,E.E. and Sankoff,D. (2003) Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**, 793–7.

Emes, RD *et al.* (2003) Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Human molecular …*.

Engström,P. *et al.* (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Research*, **17**, 1898–1908.

Escriva,H. *et al.* (2004) The evolution of the nuclear receptor superfamily. *Essays Biochem*, **40**, 11–26.

Fagerberg, L *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular ….*

Francis, GA *et al.* (2003) Nuclear receptors and the control of metabolism. *Annual review of ….*

Germain,P. *et al.* (2006) Overview of Nomenclature of Nuclear Receptors. *Pharmacological Reviews*, **58**, 685–704.

Guénet, JL (2005) The mouse genome. *Genome Research*.

Harris,R. and Hofmann,H. (2015) Seeing is believing: Dynamic evolution of gene families. *Proceedings of the National Academy of Sciences*, **112**, 1252–1253.

Hurles,M. (2004) Gene Duplication: The Genomic Trade in Spare Parts. *PLoS Biology*.

Irimia, M *et al.* (2008) Convergent evolution of clustering of Iroquois homeobox genes across metazoans. *Molecular biology and evolution*.

Irimia,M. *et al.* (2012) Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.*, **22**, 2356–2367.

Jordan, IK *et al.* (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology*.

Kikuta,H. *et al.* (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome &*.

Kimura,M. (1985) The neutral theory of molecular evolution.

Kitambi,S. and Hauptmann,G. (2006) The zebrafish orphan nuclear receptor genes nr2e1 and nr2e3 are expressed in developing eye and forebrain. *Gene expression patterns : GEP*.

Kojetin,D.J. and Burris,T.P. (2013) Small molecule modulation of nuclear receptor conformational dynamics: implications for function and drug discovery. *Molecular pharmacology*, **83**, 1–8.

Kuo, M *et al.* (2005) Gene duplication, gene loss and evolution of expression domains in the vertebrate nuclear receptor NR5A (Ftz-F1) family. *Biochem. J.*

Laudet, V *et al.* (1992) Evolution of the nuclear receptor gene superfamily. *The EMBO journal*.

Leister, D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics*.

Li, X *et al.* (2006) Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. *Plant* ….

Li,W. *et al.* (2007) Nuclear receptor TLX regulates cell cycle progression in neural stem cells of the developing brain. *Molecular endocrinology (Baltimore, Md.)*.

Liu,H.-K. *et al.* (2008) The nuclear receptor tailless is required for neurogenesis in the adult subventricular zone. *Genes & development*, **22**, 2473–2478.

Maeso, I *et al.* (2012) An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome* ….

Meyer,A. *et al.* (2003) Genome Evolution. *Journal of structural and functional genomics*.

Novac, N and Heinzel, T (2004) Nuclear receptors: overview and classification. *Current Drug Targets-Inflammation &* ….

Ohno, S (1970) Evolution by gene duplication.

Ponce,R. and Hartl,D.L. (2006) The evolution of the novel Sdic gene cluster in Drosophila melanogaster. *Gene*, **376**, 174–83.

Robinson-Rechavi,M. *et al.* (2003) The nuclear receptor superfamily. *Journal of Cell Science*.

Roth,C. *et al.* (2007) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zool. B Mol. Dev. Evol.*, **308**, 58–73.

Sharma,Y. *et al.* (2014) Computational Characterization of Modes of Transcriptional Regulation of Nuclear Receptor Genes. *PLoS ONE*.

Stark,G.R. (1993) Regulation and mechanisms of mammalian gene amplification. *Adv. Cancer Res.*, **61**, 87–113.

Vishnoi, A *et al.* (2010) Young proteins experience more variable selection pressures than old proteins. *Genome* ….

Wang, W *et al.* (2007) Comparison of Pax1/9 locus reveals 500-Myr-old syntenic block and evolutionary conserved noncoding regions. *Molecular biology and evolution*.

Wang, W *et al.* (2002) Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proceedings of the* ….

Woods, S *et al.* (2013) Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS genetics*.

Xie,C.-Q. *et al.* (2009) Expression Profiling of Nuclear Receptors in Human and

Mouse Embryonic Stem Cells. *Molecular Endocrinology*, **23**, 724–733.

Yang,X. *et al.* (2006) Nuclear Receptor Expression Links the Circadian Clock to Metabolism. *Cell*, **126**, 801–810.

Zhang,Z. *et al.* (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, proteomics & bioinformatics*, **4**, 259–263.

Table 1. Expression classification of nuclear receptor genes.

Gene	CLASS
NR6A1	TISSUE-ENRICHED
NR5A2	EXPRESSED IN ALL LOW
NR5A1	GROUP ENRICHED
NR4A3	EXPRESSED IN ALL LOW
NR4A2	EXPRESSED IN ALL LOW
NR4A1	EXPRESSED IN ALL LOW
PGR	TISSUE-ENRICHED
NR3C2	EXPRESSED IN ALL LOW
NR3C1	EXPRESSED IN ALL LOW
ESRRG	MIXED LOW
ESRRB	MIXED LOW
ESRRA	EXPRESSED IN ALL LOW
ESR2	EXPRESSED IN ALL LOW
ESR1	EXPRESSED IN ALL LOW
AR	EXPRESSED IN ALL LOW
RXRG	MIXED LOW
RXRB	EXPRESSED IN ALL HIGH
RXRA	EXPRESSED IN ALL LOW
NR2F6	EXPRESSED IN ALL LOW
NR2F2	EXPRESSED IN ALL LOW
NR2F1	EXPRESSED IN ALL LOW
NR2E1	TISSUE-ENRICHED
NR2C2	EXPRESSED IN ALL LOW
NR2C1	EXPRESSED IN ALL LOW
HNF4G	MIXED LOW
HNF4A	MIXED LOW
VDR	EXPRESSED IN ALL LOW
THRB	EXPRESSED IN ALL LOW
THRA	EXPRESSED IN ALL LOW
RORC	EXPRESSED IN ALL LOW
RORB	EXPRESSED IN ALL LOW
RORA	EXPRESSED IN ALL LOW
RARG	EXPRESSED IN ALL LOW
RARB	EXPRESSED IN ALL LOW
RARA	EXPRESSED IN ALL LOW
PPARG	EXPRESSED IN ALL LOW
PPARD	EXPRESSED IN ALL LOW
PPARA	EXPRESSED IN ALL LOW
NR1I3	TISSUE-ENRICHED
NR1I2	MIXED LOW
NR1H4	MIXED LOW

NR1H3	EXPRESSED IN ALL LOW
NR1H2	EXPRESSED IN ALL HIGH
NR1D2	EXPRESSED IN ALL LOW
NR1D1	EXPRESSED IN ALL LOW
NR0B2	MIXED LOW
NR0B1	MIXED LOW

Figure 1. This figure shows ratio of non-synonymous to synonymous substitution rate (K_a/K_s) for nuclear receptor genes. Genes that are targets of the GRB model are shown in red. The K_a/K_s ratios are calculated using the model-averaging algorithm of (Zhang *et al.*, 2006) .

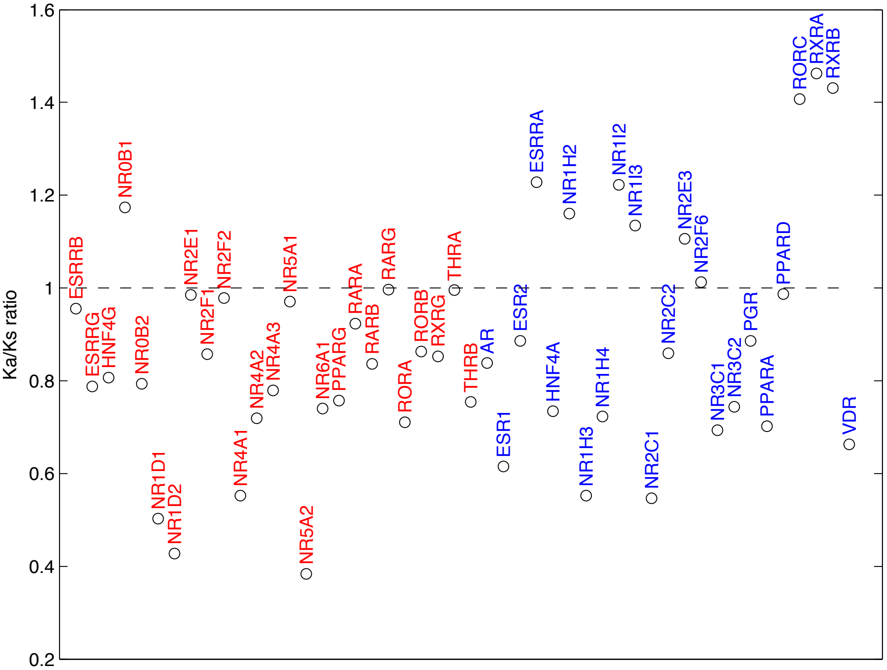


Figure 2. Boxplots of CGN scores between human to zebrafish. The x and y axis represents GRB non-target and GRB target genes and CGN scores respectively.

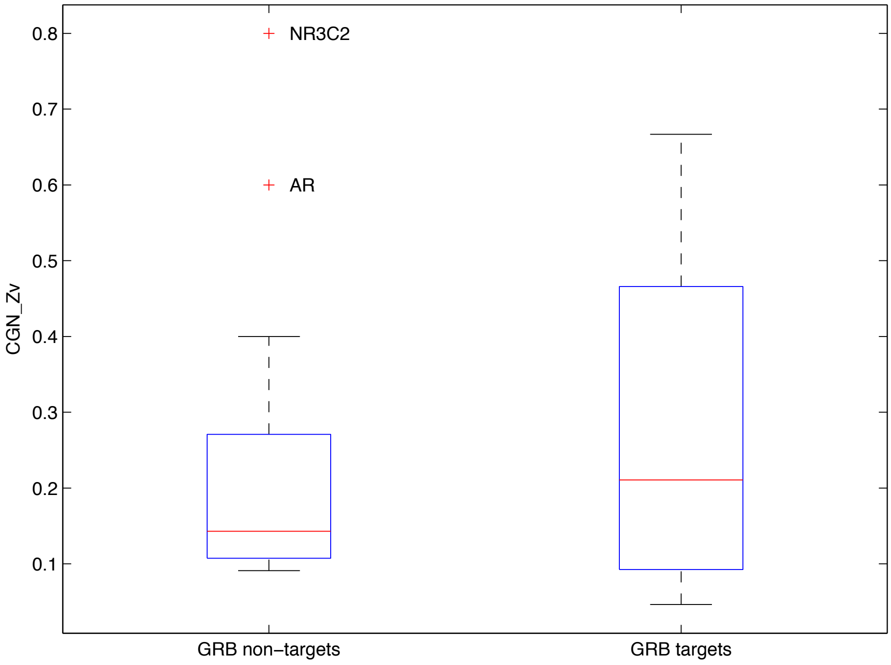


Figure 3. Significance of CGN scores between human and zebrafish. The x and y-axis represents number of neighboring genes and CGN scores respectively.

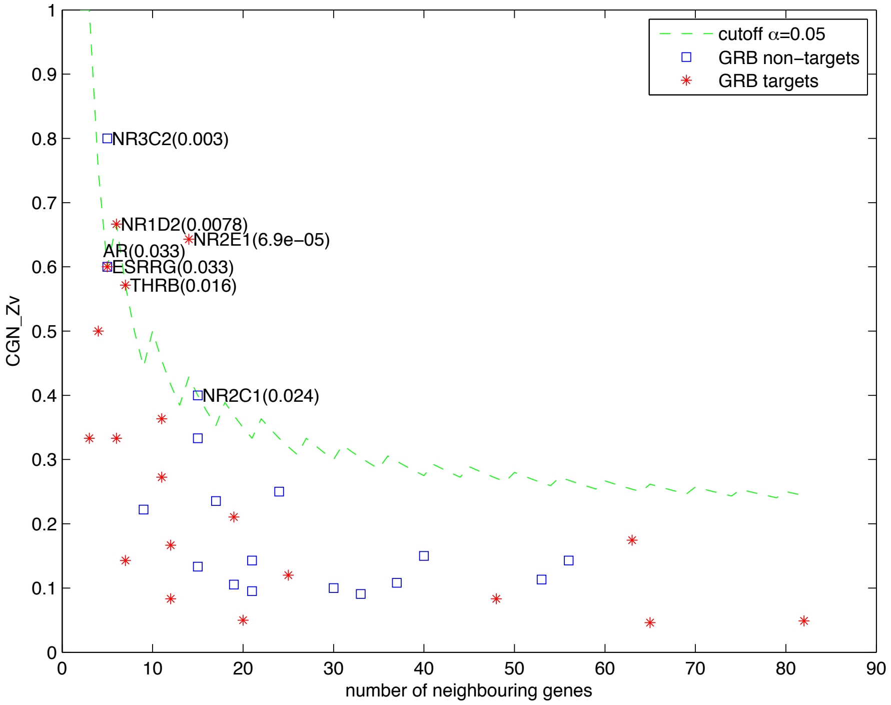


Figure 4. Expression profiles of the nuclear receptor genes. The x and y axis represents tissue and gene names respectively. Each gene is classified on the basis of subfamily (0 to 6), mechanism of action (Nuclear Hormone (H), Nuclear Orphan (O) receptor) and GRB Target (T) and GRB non-target (N).

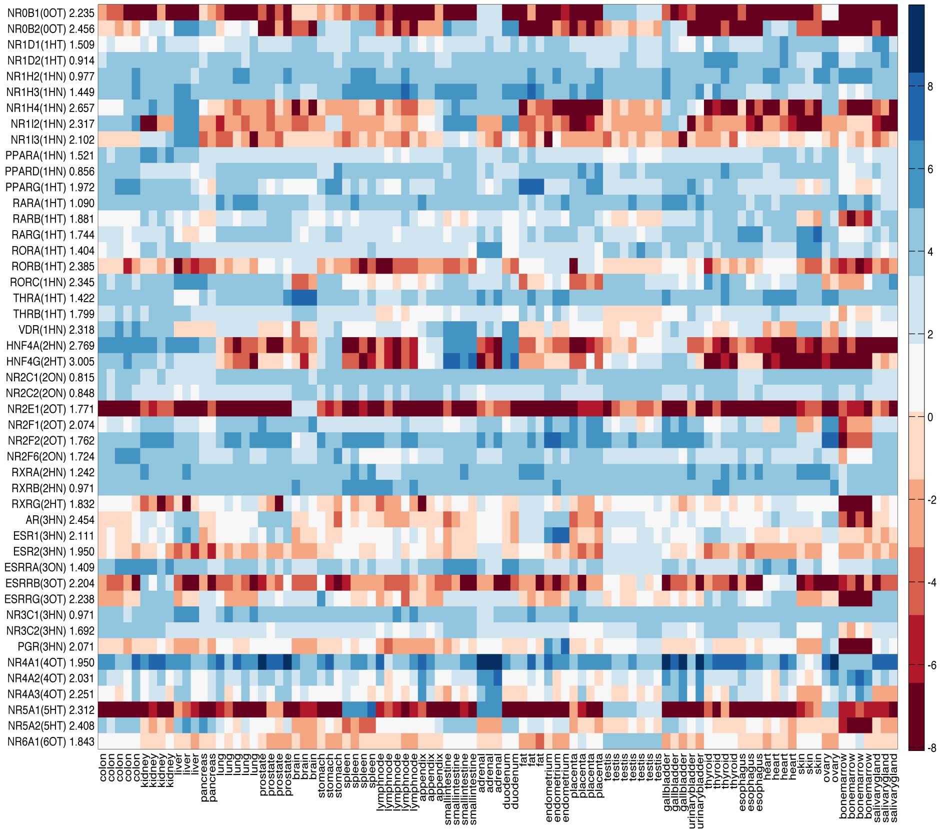


Figure 5. Boxplot showing entropy values for GRB targets and non-targets computed from log2 expression values of tissue-specific expression levels (see Figure 4).

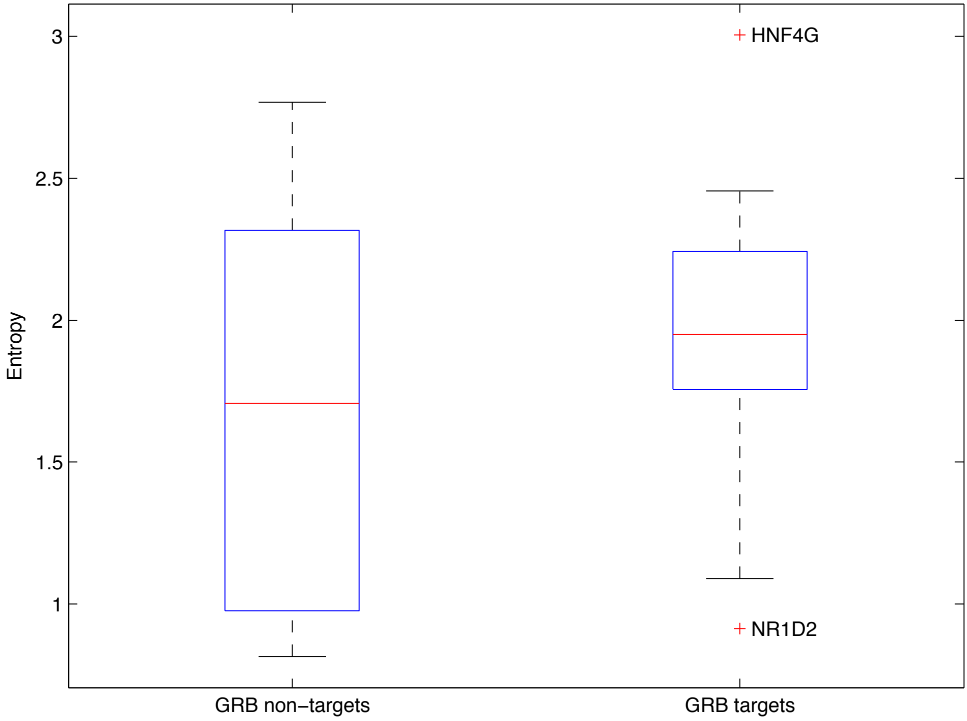
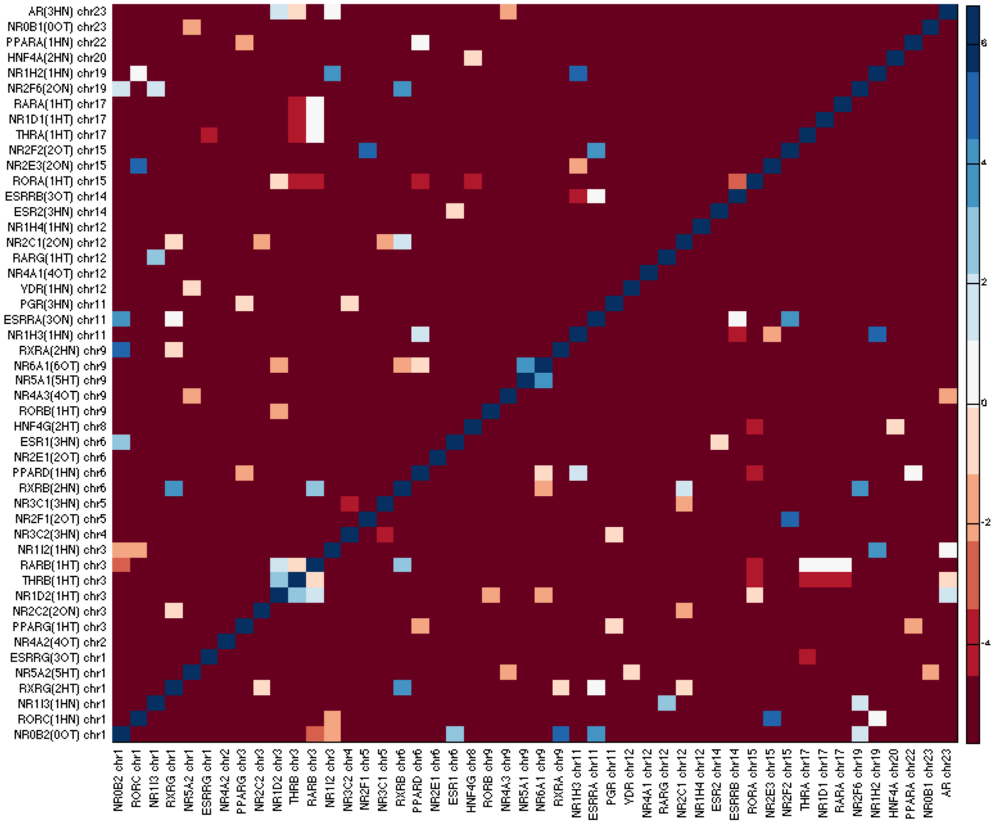
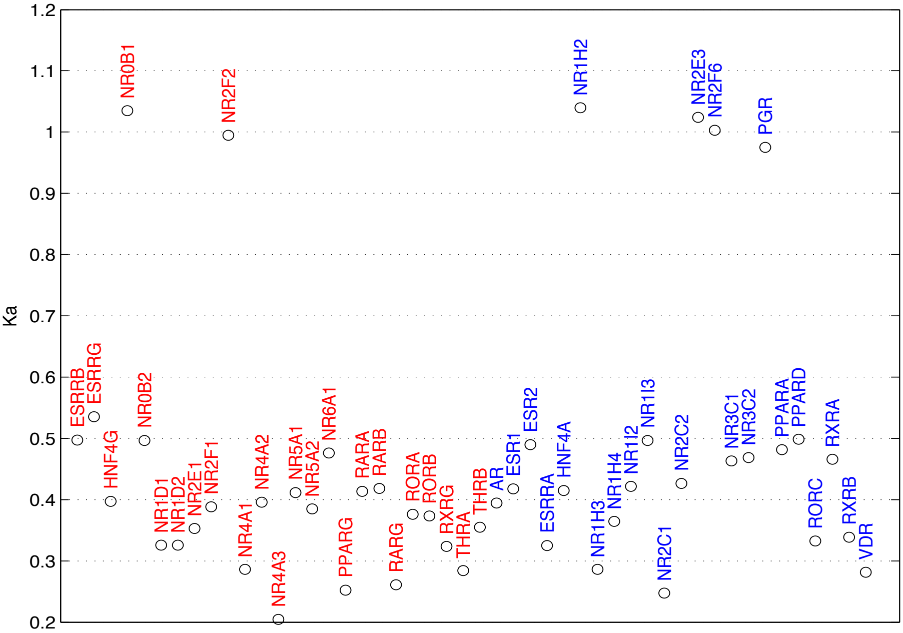


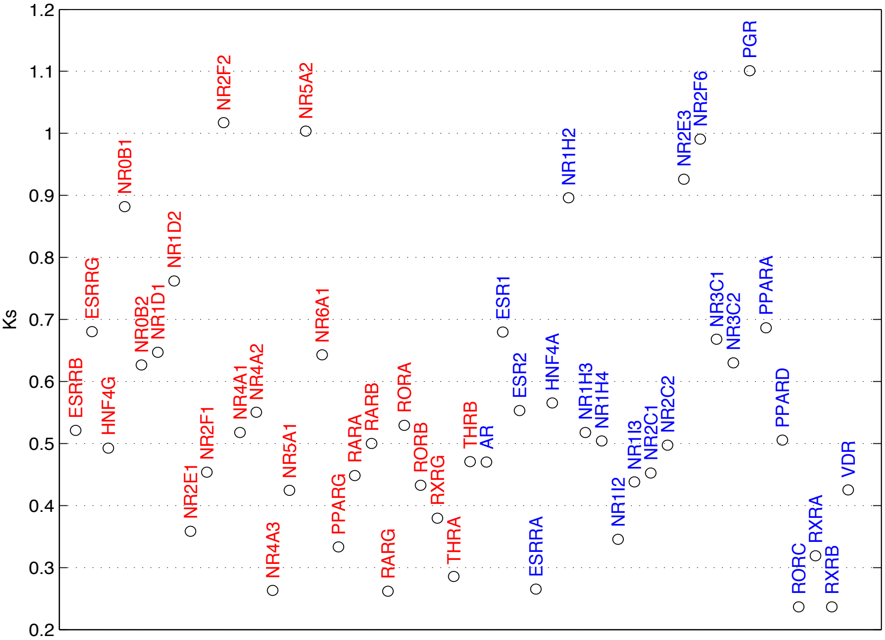
Figure 6. Selfchain alignment of the nuclear receptor genes in human. Both x and y-axis represent nuclear receptor genes.



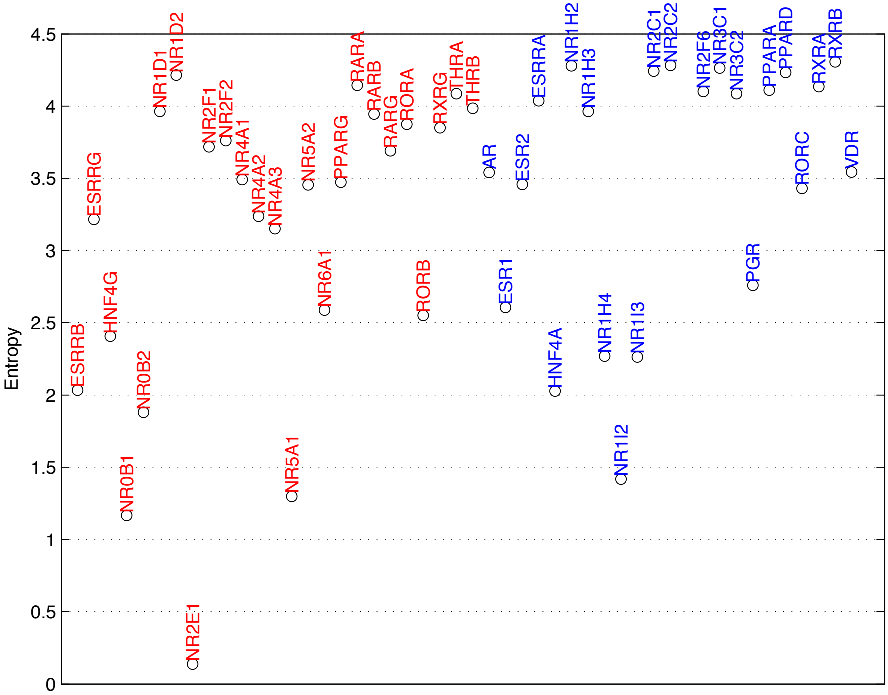
Supplementary Figure 1a: Non-synonymous substitution rate (K_a) of nuclear receptor genes (GRB targets, GRB non-targets)



Supplementary Figure 1b: Synonymous substitution rate (K_s) of nuclear receptor genes (GRB targets, GRB non-targets)



Supplementary Figure 2: Entropy values of nuclear receptor genes (GRB targets, GRB non-targets).



Supplementary Table 1. The list of genes in clusters obtained using HCNE based analysis in the GRB model. Genes in Cluster 1 are putative targets of the GRB model.

Gene Name	Cluster ID	Homology-based subfamily	Mechanism of action
NR1D1	1	I	NHR
RARA	1	I	NHR
THRA	1	I	NHR
NR4A3	1	IV	NOR
NR6A1	1	VI	NOR
NR1D2	1	I	NHR
RARB	1	I	NHR
THRB	1	I	NHR
RARG	1	I	NHR
HNF4G	1	II	NHR
NR0B1	1	0	NOR
NR2E1	1	II	NOR
NR5A1	1	V	NHR
RORA	1	I	NHR
RORB	1	I	NHR
NR0B2	1	0	NOR
NR4A2	1	IV	NOR
ESRRG	1	III	NOR
NR5A2	1	V	NHR
NR2F1	1	II	NOR
NR2F2	1	II	NOR
NR4A1	1	IV	NOR
RXRG	1	II	NHR
PPARG	1	I	NHR
ESRRB	1	III	NOR
NR1H3	2	I	NHR
AR	2	III	NHR
NR2C1	2	II	NOR
RORC	2	I	NHR
NR2E3	2	II	NOR
NR1I2	2	I	NHR
NR1H4	2	I	NHR
ESR1	2	III	NHR
ESR2	2	III	NHR
NR2C2	2	II	NOR
NR1H2	2	I	NHR
PGR	2	III	NHR
RXRA	2	II	NHR
NR3C1	2	III	NHR
NR3C2	2	III	NHR
PPARD	2	I	NHR
VDR	2	I	NHR
NR1I3	2	I	NHR
RXRB	2	II	NHR
NR2F6	2	II	NOR
PPARA	2	I	NHR
HNF4A	2	II	NHR

ESRRA	2	III	NOR

Supplementary Table 2. List of multi-gene GRB locus along with their gene names.

Serial Number	Gene Names
GRB 1	THRB, RARB, NR1D2
GRB 2	THRA, RARA, NR1D1
GRB 3	NR5A1, NR6A1

Supplementary Table 3. CGN scores of nuclear receptor genes in human (hg18) compared to mouse (mm9) and zebrafish (Zv7) genomes. Here NT and NC denote the total number of neighbouring genes within 2Mb region of nuclear receptor (NR), and, number of neighbouring genes with orthologs in the mouse (resp., zebrafish) genome with 2Mb of the NR ortholog. $CGN = NC/NT$.

Gene	NT	NC (mm9)	CGN (mm9)	NC (Zv7)	CGN (Zv7)
NR2F2	3	1	0.33		
NR2F1	6	3	0.5	2	0.33
NR2F6	56	52	0.93	8(3)	0.14 (0.05)
ESR2	24	19	0.79	6(3)	0.25(0.12)
ESR1	15	14	0.93	3	0.60
AR	5	5	1	2	0.22
PGR	9	7	0.78	2	0.22
NR3C2	5	5	1	4	0.80
NR3C1	7	6	0.86		
ESRRG	5	4	0.8	3(2)	0.60(0.40)
ESRRB	25	24	0.96	3	0.12
ESRRA	70	59	0.84		
NR6A1	20	12	0.6	1	0.05
NR5A2	11	10	0.91	4	0.36
NR5A1	19	12	0.63	4(1)	0.21(0.05)
NR1H3	37	33	0.89	4	0.11
NR1H2	87	73	0.84		
NR1H4	15	11	0.73	2	0.13
THRB	7	7	1	4	0.57
THRA	65	51	0.78	3(1)	0.05(0.02)
VDR	30	21	0.7	3(1)	0.10(0.03)
NR1I2	23	22	0.96		
NR1I3	50	44	0.88		
RARB	5	5	1		
RARA	82	52	0.63	4(3)	0.05(0.04)
RARG	63	53	0.84	11(3)	0.17(0.05)
NR4A3	12	12	1		
NR4A2	4	4	1	2(1)	0.50(0.25)
NR4A1	49	38	0.78		
RORB	7	6	0.86	1	0.14
RORA	11	9	0.82	3	0.27
RORC	69	50	0.72		
NR1D2	6	6	1	4(1)	0.67(0.17)
NR1D1	66	51	0.77		
PPARD	33	29	0.88	3(2)	0.09(0.06)
PPARA	21	18	0.86	3(2)	0.14(0.10)
PPARG	14	12	0.86		
NR2C2	21	12	0.57	2	0.10
NR2C1	15	14	0.93	6	0.40
RXRB	53	32	0.6	6(1)	0.11(0.02)
RXRA	17	12	0.71	4(4)	0.24(0.24)
RXRG	12	9	0.75	2	0.17
HNF4G	3	1	0.33	1	0.33
HNF4A	40	35	0.88	6	0.15
NR2E3	19	14	0.74	2	0.11
NR2E1	14	12	0.86	9	0.64
NR0B2	48	43	0.9	4	0.08
NR0B1	12	7	0.58	1	0.08

Errata for
Nuclear Receptor Genes – Regulation and Evolution

Yogita Sharma

Thesis for the degree philosophiae doctor (PhD)
at the University of Bergen

Yogita

(signature of candidate)

(signature of faculty)

12th February 2016

Errata

Page 79 "For several nuclear receptors ()" – For several nuclear receptors (*NR2F2, NR2F1, RARB*)

Page 79 "We also identify the cases of fast evolving nuclear receptor genes ()" - We also identify the cases of fast evolving nuclear receptor genes (*NR2F6, NR0B1*)