Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Diplom-Biol. Dolle, Dirk-Dominik
born in: Groß-Gerau

Oral-examination: _____

Application of motif scoring algorithms for enhancer prediction in distantly related species

Referees:

Prof. Dr. Joachim Wittbrodt

Dr. Steffen Lemke

Acknowledgements

First of all I want to thank Dr. Laurence Ettwiller for giving me the chance to start my PhD in this field and for introducing me into various successful collaborative projects. Not any less I thank Prof. Dr. Jochen Wittbrodt for all his advice and help during my whole PhD and especially for “adopting” me to his group when things changed. Further thanks go to Dr. Steffen Lemke who accepted to be second supervisor for this thesis despite the short notice and the trouble this may have caused. I also want to thank Michael Eichenlaub for all the support he offered me in the lab, for his advice and patience during the three years. Special thanks also go to all people in the lab who were willing to sacrifice their time for injecting multiple enhancer constructs into hundreds of medaka embryos. Without their help I could not have finished my thesis in that time. Most of all, I want to thank Dr. Juan Luis Mateo Cerdan for his strict, critical but constructive, very supportive, and always helpful comments throughout the whole process of the underlying work of this thesis, and for all the time he invested in the hundreds of detailed discussions we had. Without his help I would have never made it that far! I further want to thank all the people in my former lab, the lab of Jochen Wittbrodt and all other labs around that shared space and equipment with us and created a nice working atmosphere. I also want to thank all the members of the Hartmut Hoffmann-Berling International Graduate School (HBIGS) for the supportive environment they provided. Of course I also want to mention my family and friends who were always willing to help and to listen whenever necessary. Finally, I also want to mention Rolf Jansson and his team who, without knowing, provided a spare time activity that allowed me to regenerate and focus all my energy on my PhD.

Summary

Although many studies proposed methods for the identification of enhancers, reliable prediction on a genome-wide scale is still an unsolved problem. One of the reasons for this is the highly flexible regulatory logic underlying a detectable enhancer activity. In each cell type or tissue and at any given time, a mostly unknown set of transcription factors activates specific regulatory elements by coordinated binding to the corresponding genomic region. Position, spacing, and orientation of the individual bound factors can thereby vary between different enhancers yet result in a highly similar spatio-temporal activity. Due to this inner flexibility, so-called “alignment-free” methods have been proposed for enhancer prediction, as they are able to cope with rearrangements by comparison of word profiles rather than linear sequence. However, the problems caused by allowing for permutation in sequence comparison have not been investigated so far. In this study I implemented several published alignment-free metrics and analysed, which parameters affect their ability to successfully predict regulatory regions. As results show, single point mutations and the increasing amount of spurious matches with decreasing word size pose the biggest challenge to alignment-free techniques, especially when applied on a genome-wide scale. Alignment algorithms usually solve these problems quite efficiently but cannot handle permutation. I therefore implemented a new technique for enhancer prediction that combines the advantages of both algorithm types and used it for the identification of regulatory regions in the teleost fish *Oryzias latipes* (Medaka) based on a set of known and validated human enhancers. Predicted medaka regions and human enhancers were subsequently used in an *in vivo* enhancer assay and analysed for their activity. In total, 12 predicted regions corresponding to 9 human enhancers showed clear enhancing activity in the fish. This shows that the principle implemented here is able to predict active enhancers at a high rate on a genome-wide scale even in species as diverged as human and fish. Furthermore, evidence for motif-level conservation between some of the human and medaka enhancers could be found that was invisible for most of the alignment-algorithms used for comparison.

Zusammenfassung

Obwohl bereits viele Studien Methoden zur Identifizierung von Enhancern vorgeschlagen haben ist eine verlässliche Vorhersage in ganzen Genomen noch immer problematisch. Eine Ursache dafür sind die zu Grunde liegenden, teilweise sehr flexiblen regulatorischen Mechanismen in Enhancern. In jedem Zelltyp oder Gewebe resultiert die spezifische Aktivität eines regulatorischen Elements aus der koordinierten Bindung eines meist unbekannten Sets von Transkriptionsfaktoren an die entsprechende Region im Genom. Verschiedene Enhancer können dabei sehr ähnliche Aktivitätsprofile zeigen, selbst wenn sich Positionierung und Orientierung der einzelnen Faktoren, sowie deren Abstand untereinander, stark unterscheiden. Auf Grund dieser inhärenten Flexibilität wurden in der Vergangenheit so genannte “Alignment-free” Methoden zur Identifizierung von Enhancern vorgeschlagen, da diese in der Lage sind, Permutationen über den Vergleich von “Wörter-Profilen” auszugleichen. Die damit verbundenen Schwierigkeiten wurden allerdings bisher nicht wirklich untersucht. In dieser Arbeit habe ich daher verschiedene bereits publizierte Methoden implementiert um herauszufinden, welche Faktoren eine zuverlässige Vorhersage beeinflussen. Wie die Analysen zeigen stellen Punktmutationen und zufällige Übereinstimmungen von “Wörtern” das größte Problem dar, ganz besonders im genomweiten Maßstab. Alignment-Algorithmen lösen diese Probleme zwar recht effizient, sind aber nicht in der Lage Permutationen zu kompensieren. Aus diesem Grund habe ich für diese Arbeit eine neue Methode entwickelt, welche die Vorteile beider Arten von Algorithmen zu verbinden versucht. Diese neue Technik wurde dann angewendet um unter Verwendung von bekannten Enhancern im Menschen ebenfalls regulatorische Regionen im Teleost *Oryzias latipes* (Medaka) zu identifizieren. Diese Regionen aus beiden Spezies wurden anschließend mittels eines *in vivo* Enhancer Assays auf regulatorische Aktivität untersucht. Ausgehend von 9 Enhancern im Menschen konnten so 12 Regionen in Medaka mit eindeutiger regulatorischer Aktivität entdeckt werden. Dies ist ein klarer Hinweis darauf, dass die hier verwendete Methode in der Tat in der Lage ist, aktive Enhancer mit hoher Erfolgsrate auf genom-weiter Ebene zu identifizieren – selbst in so

verschiedenen Spezies wie Mensch und Fisch. Weiterhin zeigen einige der getesteten Regionen Hinweise auf Konservierung von Sequenzelementen, die von den meisten der zum Vergleich verwendeten Alignment-Algorithmen nicht entdeckt werden konnten.

1. Introduction

1.1 Evolution of species

The huge diversity of species has at all times raised the question, how such extensive variation could have been achieved. Besides several other explanation attempts, the idea that species could originate from other, previously existing ones came up already in ancient times. But it took until 1858 for the first scientifically sound theory to be presented [1]. In that year, Darwin and Wallace together presented their theory of evolution in front of the Linnean Society of London, a theory that both had developed independent of each other. Since these days it became more and more accepted that species acquire new traits and phenotypes by chance, which are selected for once they provide an advantage compared to the parent species. Transmitted over several generations, the accumulated divergences might finally result in the creation of new species from former variants, which either exist in parallel to the parent species or replace it if superior. This theory of gradual divergence of species from a common ancestor also laid the foundation for the concept of homology, allowing restructuring the previously existing classification of species based on evidence for common ancestry. But until the rediscovery of Mendel's rules of inheritance in the beginning 20th century, there was no explanation how this process of transmission across generations could be achieved. In 1915, Thomas Hunt Morgan was the first to prove, based on his studies in flies, that the information for the observed phenotypes had to be located on specific macromolecules, the chromosomes [2]. The smallest unit of this information was called a "gene" although at that time it was rather a theoretical construct than a clear definition of a physical region. It took additional 14 years before Barbara McClintock showed that genes in fact are real objects located on chromosomes [3]. At that time, two possible carriers of the "genetic" information were discussed, as both are contained in large quantities within chromosomes: DNA and proteins. One year before, Griffith [4] had already shown that genetic information can be transmitted between species but as he used cell extracts that contained both substances it stayed unclear which molecule contained the information. This proof was finally given

by Avery in 1944 [5], who repeated the experiments, this time removing individual components of the extract to test for the effect. Digestion of DNA prior to exposure of bacteria of a specific strain (R-type) to extracts from S-type strains finally revealed that it had to be the DNA - and not the proteins - that encodes the specific properties of the S-type strain. This could be further confirmed by experiments on lambda phages performed by Hershey and Chase in 1952 [6]. One year later, Watson & Crick [7], Wilkins [8], and Franklin [9] published both, the structural model of the DNA and the experimental results proving it, and this way allowed to explain how the genetic information is stored in a physical molecule. This finally resulted in the discovery of the genetic code by Nirenberg and others in 1965 [10] and led to what is called the “Central Dogma” of molecular biology, describing the process by which genetic information stored in the DNA is transmitted into proteins that fulfil most functions of a living cell.

While this explained how information is stored and evaluated by living organisms, the question remained, how this would lead to changes in species and thereby evolution. Since 1941, when Beadle and Tatum had shown that mutations in genes can alter metabolic pathways and are therefore likely to affect the organism as a whole [11], it was widely accepted that mutations modify the function of the encoding genes and in turn result in phenotypic changes that might lead to speciation. Studies by Jacob and Monod in 1960 further supported this [12]. They used *Escherichia coli* (E. coli) to study two mechanisms of gene regulation that could be affected by mutations in a specific class of repressive DNA binding proteins. Their main achievement however, was that they provided the first example of experimental evidence for the regulatory potential of proteins. In 1975, King and Wilson [13] were among the first who suggested that mutations in the regulatory architecture and not in protein-coding regions might account for the observed interspecies differences. Jacob followed in 1977, stating that mutations in regions of the DNA bound by regulatory proteins might be a more likely mechanism to create phenotypic diversity [14]. But despite these early publications, mutations in genes were still thought to be the main mechanism of speciation for several decades. This started to change after the successful sequencing of

the complete human genome sequence in 2001 [15,16]. In the pre-Human Genome Project era, the number of expected protein coding genes had been assumed to be roughly 40,000 [17], although also ranges between 60,000 – 70,000 [18] or even 120,000 [19] had been hypothesized – numbers way higher than the ~22,000 protein coding genes contained in the most recent human genebuild (EnsEMBL, v68). This high difference partially derived from the assumption that the increasing complexity of organisms is achieved by an increased number of genes, which is in clear contrast to the numbers known today. Interestingly enough, genome sizes and number of protein coding genes are very similar between Human, Mouse, and Rat ***(Table 1)***.

Species	Genome size *(assembly)*	Protein-coding genes *(EnsEMBL v68)*
Homo sapiens	**3,100 mb** *(GRCh37.p8/hg19)*	**22,088**
Mus musculus	**2,700 mb** *(GRCm38/mm10)*	**22,662**
Rattus norvegicus	**2,700 mb** *(RGSC 3.4/rn4)*	**22,938**
Oryzias latipes	**800 mb** *(HdrR/MEDAKA1/ol2)*	**19,686**
Drosophila melanogaster	**169 mb** *(BDGP 5)*	**13,940**

Table 1 *Genome size and number of protein coding genes per species (state: EnsEMBL v68)*

Comparison between these mammals and the teleost *Oryzias latipes* ("Medaka") reveals that even across an evolutionary distance of ~450mio years, gene counts have only slightly changed despite a significant difference in genome size. But even when compared to the invertebrate *Drosophila melanogaster* ("Fruitfly"), the gene count is only halved. These numbers clearly point out that the morphological differences between those species are not accompanied by dramatic changes in the amount of protein coding genes. Many genes even exist as orthologous copies in species as divergent as Human and Drosophila. This indicates that, in contrast to initial assumptions, the observed diversity is unlikely to be the result of changes in coding regions. Further support comes from the findings of the Human Genome Project. Surprisingly, only 1.5% of the full genomic sequence contain protein-coding

information compared to 5% that are conserved in total [20]. This means that 3.5% of the genome is subjected to evolutionary selection but not translated into proteins. Although a fraction of these regions encode for functional classes of RNAs like tRNA (involved in protein translation), rRNA (crucial functional and structural subcomponents of ribosomes), snRNA (involved in splicing), snoRNA (part of RNA-editing complexes), and siRNA (host virus defence), the majority of them is likely to be involved in regulation. This further emphasizes that regulatory regions might provide more "evolutionary playground" than coding regions. Examples in the recent literature show that changes in regulators indeed contribute to phenotypic diversity, this way providing additional support for their importance for evolution. For instance, Prud'homme et al. demonstrated that repeated independent mutations of the same cis-regulatory element in multiple Drosophila species had led to a gain and loss of an expression domain of the *yellow* gene. This gene is involved in pigmentation processes in the flies. The changes in the regulatory region resulted in a gain and loss of a pigmented wing spot that is involved in male courtship display [21]. Another study in Drosophila showed that mutations in the dorsocentral enhancer (DCE) led to an expansion of its domain of activity resulting in posterior dorsocentral bristles in *Drosophila quadralineata* [22]. Also in vertebrates, cases of phenotypic changes as result of mutations in regulatory regions exist. Chan et al. [23,24] provided evidence, that repeated deletion of a regulatory element near the *pitx1* gene in *Gasterosteus aculeatus* had led to the loss of pelvic spines in several independent freshwater populations. Tung et al. were able to find a similar event in primates. Mutation of a regulator of the *FY* gene in yellow baboons (*Papio cynocephalus*) led to an altered resistance to a very common malaria-like parasite in this species, which might have provided a selective advantage [25]. These examples clearly highlight that regulatory mutations contribute to the phenotypic diversity of species, from insects through vertebrates and even up to primates, and show that they have been – and still are – one of the driving forces of evolution.

1.2 Mechanisms of gene regulation

Compared to *E.coli*, the organism studied by Jacob and Monod, gene regulation in eukaryotes, and especially vertebrates, is by far more complex and acts on several levels. This starts already at the genomic structure. Unlike in bacteria, vertebrate genomes are organized in several chromosomes, each consisting of a long linear DNA molecule that is wound around specific multi protein complexes, the nucleosomes. This state is also described as “30nm fibre”. During cell division, this fibre is further compacted into a highly coiled and dense structure that can be identified under the microscope as metaphase chromosomes, and expanded again afterwards. But even in terminally differentiated cells, regions of the 30nm fibre are still partially packed as a result of uneven nucleosome densities. Due to this variable packing density, certain areas in the genome are accessible for transcription (described as “euchromatin”) while others stay condensed and inactive (“heterochromatin”). Actively transcribed loci for example can be associated with certain modifications of specific amino acid residues (e.g. H3K36me3) in the N termini of histones, the protein subcomponents of nucleosomes [26]. However, which regions in the DNA are active or not is thereby not a general property of the genome but varies between conditions and cell types. Furthermore, not only transcribed but also regulatory regions are affected by differences in nucleosome-density. A study in Drosophila for example could show that predicted regulatory regions containing sequences indicative for nucleosome depletion were more likely to be active than regions without [27]. One explanation is that local binding of nucleosomes to the DNA can cover binding sites for a specific class of DNA-binding proteins, so called transcription factors (TFs), rendering them inaccessible for TF binding. The binding of these factors to their corresponding transcription factor binding sites (TFBSs) in specific regions of the genome, described as cis-regulatory elements (CREs), is one of the most important mechanisms of regulation of gene transcription [28].

CREs are named this way, as they are located “in cis” to their target gene while “trans” usually describes factors that bind to these elements. CREs can

be further subdivided into “proximal” and “distal” elements. Promoters, which are located around the transcription start site (TSS) of genes, are the most known class of proximal elements. They roughly reside within 2kb upstream (5’) and 500bp downstream (3’) of the TSS and contain the starting platform for the RNA polymerase II (RNA Pol II), which is responsible for transcription of protein-coding genes. At the same time they integrate all regulatory signals resulting in a specific spatio-temporal expression pattern. These inputs either derive from TF binding events directly in the proximal promoter region or from “enhancers”, the most common class of distal regulatory elements. The two classes do not only vary in the distance to their target gene but also have very different structural and positional properties. As already stated, promoters always reside 5’ around the TSS of their target gene. Furthermore, they contain several unique sequence features aside from the existence of TFBS, including C/G-rich clusters (“CpG-islands”), TATA- or CCAAT-boxes, DPEs (“downstream promoter elements”, [29,30]), and other specialized elements like XCPE1 & 2 (“X Core Promoter Element”) [31,32], with many of them having to be precisely positioned within the promoter region.

Enhancers on the other hand, are more flexible in their positioning and inner structure. They can reside almost anywhere in the genome: 5’ and 3’ of their target genes, in introns – and even exons – of flanking genes or the regulated gene itself [28]. Enhancers are also reported to be able to activate target genes across several intercalated “bystander genes” [33] and/or large (>1mb) distances [34]. Considering the highly folded and organized packaging of chromosomes in the nucleus [35], they might be even located “in trans”, as two regions on completely different chromosomes can be physically directly adjacent to each other. Besides this positional flexibility, they also possess a highly flexible inner structure. Usually about 1kb in size, they can span several kilobases or just a few hundred nucleotides. This is owed to the fact that many enhancers can be subdivided into cis-regulatory modules (CRMs), which are able to drive a certain expression pattern independently of other modules in the same enhancer [36,37]. It is therefore a matter of debate whether CRMs represent just a part of a larger CRE or whether CREs are just a description of genomic regions containing several CRM-enhancers. In general, two different

classes of enhancers are discriminated: densely clustered and highly structured arrangements of TFBS, the “enhanceosomes”, or loose groups of CRMs and individual TFBSs described as “billboard” enhancers [38]. While the structural rules in enhanceosomes are that strict that TF binding has to occur in a highly coordinated sequential manner [39] billboard enhancers are the clear opposite. For those, each module or site can be bound independently or in combination, depending on available interacting factors or other outer conditions [36,37,40]. This functional independence allows extensive permutation and reshuffling between the individual sites and modules within billboard enhancers without affecting the function [41,42], while enhanceosomes tolerate little to no mutation [39]. Despite these huge differences in the inner structure, the mechanism of gene activation is thought to be the same for the two classes. In both cases, activation starts by binding of the involved TFs to their binding sites. These can be independent proteins or complexes of one to many interacting factors which form the initial enhancer complex upon binding. This complex then recruits further co-factors like CBP/p300 [43] that can act as histone acetyl transferases (HATs) modifying amino acids in the N-termini of DNA-associated nucleosomes, the result of which is an open and accessible chromatin configuration. Recruitment of chromatin remodelling factors (e.g. SWI/SNF) and other complexes like TRAP/Mediator finally leads to looping of the activated enhancer complex to the promoter of its target gene and this way activates it [44,45]. Several enhancers can cooperate in this process, each of them activating the same gene in a different cell type/tissue or at a different timepoint or condition resulting in the partially highly complex and/or ubiquitous expression pattern known for many genes [46,47]. There is also evidence for the opposite case, in which a single enhancer controls several genes at the same time by looping all TSSs to one spot where it interacts with the enhancer complex [48]. Although further regulatory mechanisms set in after initiation of transcription and during or after translation, they all depend on the initial activation provided by enhancer complexes. Their ability to also remodel chromatin structures in and around areas containing their corresponding target genes by recruitment of specific co-factors places them at the basis of gene regulation in eukaryotes. Furthermore, due to the high

number of putative TFs and TFBSs and the integration of multiple inputs at the same time, they offer the highest flexibility and diversity of all regulatory mechanisms.

1.3 Enhancer prediction

As described, enhancers are highly flexible structures that can occur at various different places in the genome. Unfortunately, this high structural and positional flexibility makes them also very hard to detect. Even promoters are already hard to predict, although they are restricted to the close vicinity of TSSs. This results from the fact that the initial “one gene one polypeptide” hypothesis does not reflect the physical reality of genomes. Besides alternative splicing that can create a huge variety of different gene products from a single locus, gene transcription can also start at different positions in the genome resulting in different proteins without even involving splicing events [49,50]. As these shifts can sometimes change the position of the TSS by several hundred to more than thousand base-pairs, it is a challenging task to predict the correct promoter region without the precise position of the used TSS. But while promoters contain certain core elements that might allow their prediction even in that case, no comparable structural elements are known for enhancers. Additionally, the possible search space can literally be the full genome, which further complicates the task. Different approaches have been used in the past to predict the location and/or function of enhancers, which can be classified in two different categories: biological and computational approaches.

1.3.1 Biological approaches

1.3.1.1 Targeting transcriptional (co-)factors

This class of prediction techniques is largely based on chromatin immunoprecipitation methods (ChIP) that are followed by different subsequent evaluation steps. They can be further differentiated in methods targeting TFs directly, those that focus on known interaction partners involved in enhancer complexes, especially co-factors, and other methods that try to predict

enhancers indirectly by assessing the chromatin state of cells or tissues. ChIP-chip belongs to the first category and was also one of the first methods used. For this technique, TFs are first crosslinked to the bound DNA region, followed by dissection of the genome into small pieces by sonication, restriction digest or other methods. Antibodies targeting the factor of interest and linked to different substrates e.g. magnetic beads are then mixed with the sample. This allows precipitation of transcription factor-bound DNA fragments, which were protected against the dissection process by the protein. Reversal of the cross-linking process results in a genomic fraction highly enriched for fragments containing a binding site for the targeted factor. These fragments are subsequently detected by DNA-probes via hybridization. The major drawback of this method is its limitation to a very restricted set of genomic regions due to the designed probe set. In the pre-genomic era, this only allowed the detection of binding events in already known regions and was therefore mainly restricted to the detection of promoters and proximal regulators, a task for which it is still used [51]. The progress made in sequencing techniques allowed the extension of ChIP to further tasks. Precipitation followed by massive parallel sequencing (ChIP-seq) does not depend anymore on previous knowledge of the target sequence and therefore broadens the spectrum, including TF motif prediction and targeting of co-factors or histones. This for the first time also allowed detailed investigation of genome-wide histone modification profiles and opened the door for large-scale epigenetic studies. The main advantage however is that it allows assessment of distant genomic regions and thereby largely facilitates enhancer prediction. Unfortunately, the main strength of this technique comes with a major drawback. Compared to prokaryotic TFs, TFs in eukaryotes have rather promiscuous binding properties in general [52], allowing them to attach to regions in the genome that match between each other by only a few nucleotides. Although the apparently variable positions might actually be highly specific in a given context (e.g. a certain cell type or tissue), this inherent flexibility is likely to result in an increased amount of noise binding events in regions that have no regulatory function. Therefore, computational methods are necessary that allow discrimination between those regions and functional ones. Furthermore, regions not bound by the factor under

investigation also often end up in immunoprecipitated sample and need to be filtered prior to data interpretation. This is usually done by comparison to a background ("input") sample, removing all regions equally enriched in both fractions. Obviously, this filtering is only as good as the used reference. Unfortunately, in the beginning of ChIP-seq, either unreliable or no "input" samples were used at all. This makes the interpretation of the resulting data nowadays very difficult. But even with good backgrounds, the data analysis stays a challenging task due to the high binding noise. To reduce the problems inherent to TF-ChIP, other studies [43,53] focused on co-factors like p300, which only bind to fully assembled enhancer complexes. This approach was very successful, yielding a high number of putative enhancers of which many could be validated experimentally in the meantime. Other approaches to increase the significance of TF-ChIP data are based on the combination of different TFs. As enhancers usually need to be bound by a set of factors rather than just one, searching for dense clusters of highly significant binding events is a very powerful way for enhancer prediction [54]. This approach is of course limited by the availability of highly specific antibodies against the TFs of interest (especially if several factors of the same family are involved) and requires previous knowledge about the expression profiles of the used TFs.

1.3.1.2 Targeting histones

A more general approach focuses on histone marks associated with different chromatin states and/or regulatory elements. Jin et al., for instance, report that promoter regions correlate with H3K4me3, while enhancers either have H3K4me1/H3K27me3 ("poised") or H3K4me1/H3K27ac ("active") marks [55]. Targeting modified histones, the building blocks of nucleosomes, hence allows indirect identification of putative enhancer regions in a way that is not limited by previous knowledge of involved TFs and more significant due to less binding noise. But like for TFs, certain tasks require a series of experiments in the same cell type or tissue to allow conclusions about the predicted regions (e.g. discrimination between poised and active enhancers requires histone-ChIP for H3K4me1 and H3K27me3 or H3K27ac). Furthermore, although this technique is able to not only identify regulatory

regions but also, whether the identified regions are active in the investigated context, it provides no insights into the underlying regulatory logic. It is therefore comparable to ChIP on co-factors like p300 but without the limitation to cells expressing it. Its main drawback however, is that some histone modifications are not restricted to a narrow locus but tend to spread across a larger regions, leading to more blurred predictions compared to clustering of TF-ChIP.

1.3.1.3 Targeting chromosomal structure

Another ChIP-based approach used for enhancer prediction is ChIA-PET ("Chromatin Interaction Analysis by Paired-End Tag sequencing") [48], which combines the classical TF-ChIP technique with the analysis of physical interactions between genomic regions. It is based on the fact that enhancers activate their corresponding target gene by looping to its promoter upon binding of TFs [44]. Regions that are in close proximity to promoters and additionally bound by a specific TF in a given context are therefore likely to be functional enhancers. Unfortunately, this technique suffers from a problem common to all “chromosomal conformation capture” (3C) methods – which is limited spatial resolution [56]. Additionally, regions located near each other in linear DNA are also always physically in close proximity, independent of the chromosomal conformation. This leads to a high amount of reported non-functional interactions of close-by loci and results in a rather low signal-to-noise ratio. Due to that, filters like independent binding events of the same factor in both interacting segments (e.g. in enhancer and target gene promoter) are necessary to allow discrimination between real chromatin loops and false positive regions. However, a binding event within a reported loop does not necessarily mean that the detected interaction is also initiated by the TF under investigation and therefore might not correspond to a functional enhancer. ChIA-PET hence suffers from the same limitations as regular TF-ChIP but acts as a filter for noise binding events as it reduces the search space to interacting genomic loci. At the same time, in contrast to other methods it provides hints about the putative target gene of functional enhancers.

All methods described here have been used over the last years and provided unprecedented insights into regulatory mechanisms of eukaryotic genomes. As they are based on experimental data, they allow conclusions beyond the scope they were initially designed for and provide information that might turn out to be valuable for not yet asked questions. But as all experimental approaches they are time-intense, labour-intensive, and costly. They furthermore provide only a snapshot of the current state and are therefore limited by the accessibility and reproducibility of the outer conditions. To predict enhancers active in a specific tissue, this tissue also has to be available. It is therefore difficult to predict enhancers e.g. active in the human brain or during human embryonic development, as these conditions are not available for experiments. Model organisms can help in this situation but subsequently need methods that allow the identification of the functional analogous enhancer in the species of interest. Considering the flexible nature of many enhancers, this is a challenging task.

1.3.2 Computational approaches

Computational methods for enhancer prediction have several advantages compared to biological approaches. They do not depend on the availability of a specific cell type or tissue, are therefore not affected by the current state of the cell, provide information about all regions of the genome, and most importantly, are fast and cheap. Depending on the method, they are also not biased by selection of specific TFs and are therefore theoretically able to predict all kinds of enhancers, provided that the necessary information is available. These methods can be roughly classified in conservation-based, clustering or motif-scoring approaches.

1.3.2.1 Alignment-based detection of conservation

Since the beginning of the “genomics” era, started by the emergence of high-throughput sequencing techniques, comparison of multiple genomes of distantly related species by multiple alignments allows the detection of highly stable genomic regions outside of coding genes. These regions, which form up to 3.5% of the human genome, are under clear evolutionary constraint and

therefore might be of functional importance. After subtraction of all those regions that contain crucial non-coding RNAs, the function of the remaining is likely to be regulatory. It is therefore not surprising that deep sequence conservation across large phylogenetic distances, like between human and fish, has been suggested or used as predictor for enhancer regions in many publications [34,57–60]. The predictive strength of these approaches increases the more genomes are used and the larger the maximum evolutionary distance is across which the enhancer is conserved. This method therefore mostly detects enhanceosomes due to their packed and highly restrictive structure. Identified enhancers mostly reside next to crucial developmental or neuronal active genes and validation revealed that they indeed mostly show activity in neuronal tissues [34]. But as alignments depend on a conserved collinear arrangement of functional enhancer motifs they are unlikely to detect enhancers that follow the billboard model and have evolved by permutation or binding site turnover. A ChIP-seq based study aiming to identify heart enhancers could show that these regulators are in the majority only weakly conserved [53] and hence would be “invisible” to alignment-based techniques. Enhancer prediction based on conservation is therefore a powerful tool that made a huge contribution to research on regulation but provides only a limited spectrum of enhancer activity.

1.3.2.2 TFBS clustering

TFBS clustering approaches are the computational counterpart to ChIP-seq experiments targeting many different TFs. The basic idea is the same: the more TF binding events occur in a narrow region the more likely this region has enhancer activity. This is computationally assessed by prediction of the corresponding TFBS clusters. Both approaches thereby depend on a telling pre-selection of factors that are known to be expressed in the same tissue and interact to drive enhancer activity. The advantage of TFBS clustering is its genome-wide applicability without the restriction to a given context. While multi-TF ChIP will only reveal enhancers that are active under the current conditions, TFBS clustering can find all enhancers that have the potential to be active in that context but might be temporarily silenced due to the

experimental setup. As their experimental counterpart they suffer from the same penalty: motif/binding noise. As described, eukaryotic TFs often have rather variable binding specificities allowing them to bind to several, partially very different sites. In contrast to ChIP-seq however computational prediction can only make limited statements about the likelihood that the predicted site is also bound. TFBSs are usually described by position weight matrices (PWMs), which are compiled from all experimental reported bound sites for a specific TF. Percentage identity to a given PWM is normally used to give hints about whether or not a site is bound *in vivo*. But as many factors have different binding affinities depending on the given context, a fact that is not represented in PWMs, this method to predict binding events is highly error prone. This situation is further complicated by the inner structure of enhancers. As described, CREs can be composed of many CRMs and individual sites, separated by non-functional spacer sequences. This configuration can "dilute" a TFBS cluster by spreading it over a larger stretch of sequence. Even if an enhancer is mainly composed of CRMs, these modules might contain too few binding site to be recognized as a cluster of their own. Therefore, although not challenged by permutation and turnover, TFBS clustering is also likely to identify enhanceosomes, as these regions are dense clusters of many sites.

1.3.2.3 Motif scoring methods

Motif scoring methods are also called "alignment-free" techniques, as they, in contrast to alignment algorithms, do not depend on a collinear arrangement of functional elements in query and target sequence. This is achieved in most cases by dissection of a given enhancer sequence into a profile of overlapping, short sequence fragments, called "words". As multiple different permutations of the same set of words can lead to exactly the same word profile, these techniques are able to cope with enhancer evolution by permutation. The words thereby represent TFBS that might change their position by several mutational mechanisms, especially in billboard enhancers. Due to that, word sizes used by most algorithms range between 5 – 8nt, as this is the typical size of a TFBS. They therefore could be regarded as TFBS

clustering methods – or vice versa, TFBS clustering as “alignment-free”. Their main difference however is that alignment-free methods use an already known enhancer as input instead of a set of TFBS of factors known to be active in a specific tissue. This way, they can use the full sequence of an enhancer to extract words instead of using only a narrow set of sites. This allows inclusion of sequences close to TFBS that might have been conserved due to the functional importance of the binding site. This “conservation shadow” hence can contribute to the identification of corresponding enhancers in paralogous loci or orthologous species where TFBS clustering might fail. The fact that mostly perfect-matching words are used further reduces noise matches. One of the first alignment-free metrics published was the “d2” metric in 1994 [61], which was used for clustering of sequences into similarity groups to speed up database search. Further metrics were published in the following years, nicely reviewed in [62]. Since then, a huge variety of algorithms utilizing various types of input data and applied in many different species or conditions have been published (reviewed and compared in [63]). Most successful studies however, were performed in Drosophila [64–67], on narrow genomic regions [67,68] or using additional data like TFBS-sets or conservation-vectors [65,67–69]. It is thereby hard to assess how useful these methods can be in general, especially when applied in more complex organisms like vertebrates and on a genome-wide scale. Furthermore, only little is known about the general problems created by compensation of permutation using word profiles.

1.4 Aim of this study

As described above, enhancers are at the very basis of gene regulation, integrating multiple different inputs that in the end lead to complex spatio-temporal expression patterns of their corresponding target gene. Increasing evidence in the recent literature indicates that mutations in these regulatory regions significantly contribute to phenotypic evolution of species. It is therefore of great interest to get insight into the functional principle of enhancers, as it would allow drawing conclusions about the mechanisms by

which they can evolve. For this, it is of crucial importance to identify functionally similar regulatory elements in different species to determine the "rules" underlying a specific spatio-temporal activity. Unfortunately, this is so far restricted to highly conserved enhancers identifiable by pair-wise or multiple sequence alignment, although many more enhancers too divergent in sequence to be alignable but fulfilling similar functions might exist in different species. Methods for identification of these "corresponding enhancers" (i.e. enhancers that receive similar input leading to comparable activity) could hence contribute significantly to our understanding of regulatory mechanisms.

Two different classes of enhancers are known today which are described either by the enhanceosome or billboard model. While the first class has a high chance to be identifiable by alignment algorithms, the latter poses a huge challenge to these techniques, as they are able to keep their activity even after extensive reshuffling of their functional elements. This makes new detection techniques necessary that are able to cope with these changes. Alignment-free algorithms are the most promising methods to handle evolutionary permutation. To date, a large variety of implemented techniques exist but most of them depend on prior knowledge that is rarely available. Furthermore, the majority was used for prediction in species that have more compact genomes and regulatory elements than vertebrates and for which more detailed data about the specificity of the involved factors exists (e.g. Drosophila). To also apply these methods for enhancer prediction in vertebrates, a better understanding of the underlying problems of the alignment-free approach is necessary. Unfortunately, the modifications implemented in currently available algorithms make this task very difficult.

Recent studies [40,70] have further emphasized the fact that, despite their flexibility, billboard enhancers do not completely reshuffle their contained TFBS. Instead, some of these sites are permuted as groups or modules. These modules, which might be even full CRMs, are under such strong structural constraint that they reappear in almost the same configuration after complete turnover. This allows treating them as small-scale enhanceosomes that permute together with additional independent site within the same

enhancer. This mode of permutation is only insufficiently described by current alignment-free approaches, as they completely dissect a given sequence into words without paying attention to the structural constraints acting on potential modules. At the same time, these modules are likely to be too short and degenerate to be detectable by alignment algorithms. It therefore might be that this type of enhancer is missed by alignment-based as well as by alignment-free approaches.

As shown in [63], many recent alignment-free algorithms make use of additional information like conservation and TFBS. This limits their applicability to situations in which the necessary information is available. They furthermore are often restricted to narrow genomic regions or used for classification of comparably small sets of enhancers [71,72]. It is therefore unknown to what extent these methods are able to identify corresponding enhancers across large evolutionary distance on a genome-wide scale. This is of special importance as enhancers might be at remote places compared to their target gene due to the highly complex 3D structure of genomes in the nucleus.

The aims of this study are therefore:

I. Analysis of the problems created by introducing permutation into sequence comparison as done by alignment-free algorithms and development of possible solutions.

II. Implementation of a prediction principle based on the previous findings that take the modular but flexible structure of enhancers into consideration.

III. Application of this new implementation for the task of genome-wide prediction of enhancers, using only a known enhancer and the available information about the query and target genome.

To achieve this, I make use of a large set of known and validated enhancers, followed by implementation and testing of several basic alignment-free metrics. The lessons learned from these techniques are then combined and used for the development of a new algorithm, which is subsequently used for prediction of corresponding enhancers between Human and Medaka.

2. Results

2.1 Data set selection

2.1.1 VISTA Enhancer browser

To identify regulatory elements across a large evolutionary distance based on only sequence information, a highly reliable set of enhancers as input is necessary. The data set for this study was extracted from the VISTA Enhancer Browser [73] *[LINK1]* as it contains the largest, consistent set of validated human enhancers. These regions were initially predicted either by deep sequence conservation or by ChIP-seq on the transcriptional co-factor p300 and subsequently validated in vivo in a mouse enhancer assay. Of all the regions contained, only those were selected which had reported enhancer activity. This resulted in a total data set of 629 human regulatory regions, which were used for all subsequent analyses.

2.1.2 Subset extraction

To generate subsets for testing, I aligned all regions against the genome of the teleost fish Medaka (*Oryzias latipes*) using LastZ [74]. This species was selected as it is separated from the human lineage by ~450mio years of independent evolution. Furthermore, through relaxation of selective pressure by the whole genome duplication that has happened at the teleost-tetrapod split, enhancers might have been allowed to undergo evolutionary change and this way allow conclusions about mutational mechanisms. Among several alignment algorithms, LastZ was chosen mainly for two reasons. The first was that deep sequence conservation used to predict the majority of the enhancer set is based on PhastCons [75]. This algorithm uses multiple alignments generated by MultiZ [76] as input to calculate, in combination with a phylogentic tree, the evolutionary constraint acting on each individual nucleotide. MultiZ in turn is based on multiple pairwise alignments generated by LastZ. LastZ should therefore be able to detect all elements that have been conserved since the teleost-tetrapod split. The other important reason was its high specificity in aligning non-coding regions [77], which is crucial when searching for enhancers across large phylogenetic distances. In total, LastZ

could identify 252 of the 629 human enhancers in the medaka genome. These enhancers formed the "aligning" subset that served to assess the sensitivity of the tested algorithms. The remaining 377 regions were used as input for the prediction of corresponding enhancers in the medaka genome. Although many of these enhancers are likely to have lost their function due to mutations, some of them might still have enhancer potential.

2.1.3 Test candidate selection

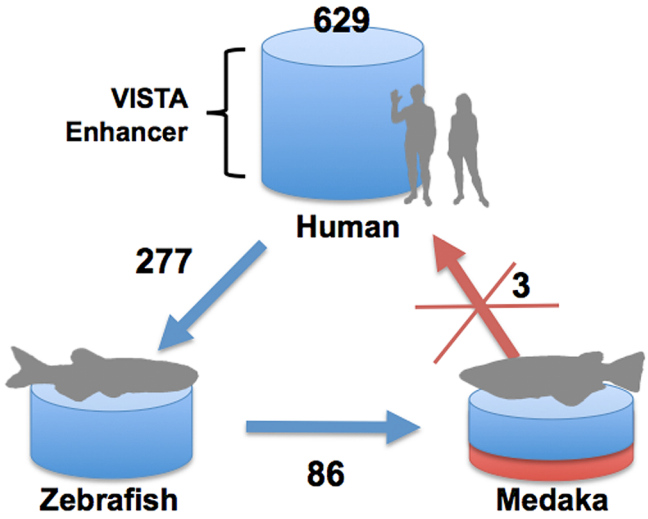


Figure 1 *Pariwise alignment pipeline from Human through Zebrafish to Medaka for all human VISTA enhancers. For three medaka loci identified in this way no direct Human-Medaka alignment was possible using LastZ.*

I further filtered the enhancer data set via an alignment pipeline from Human through Zebrafish (*Danio rerio*) to Medaka to generate a set of test candidates for assessment of the prediction capacity of several different techniques ***(Figure 1)***. This procedure served two purposes: first, using Zebrafish as an anchor, the likelihood to identify putative enhancers that are also functional in Medaka should be increased. Second, it should allow the detection of orthologous regions between Human and Medaka which cannot not be found by direct alignment on genome-wide scale. It this way facilitates the identification of regions, which on the one hand are strongly divergent in Human and Medaka but on the other hand share teleost specific mutations and innovations. Taher et al. [68] for example successfully applied a similar

"tunnelling" approach for the identification of divergent regions in Human and Zebrafish using Frog (*Xenopus tropicalis*) as intermediate step. Applied on my data set, this method resulted in three candidates (VISTA IDs: hs1022, hs692, hs20) for which the corresponding regions in Medaka identified by the pipeline do not produce significant direct alignments using LastZ on genome-wide scale. However, two of them still contain small aligning regions identifiable using another aligner, BlastN [78] *[Link2]*, at very sensitive conditions on the already identified loci. These alignments seem to be too weak to allow their identification by LastZ on genome-wide scale. The third locus contained no detectable alignment at all. All three human enhancers were subsequently used as candidates to test whether they can be detected by alignment-free prediction methods.

In addition to these three candidate loci, I also selected two enhancers (VISTA IDs: hs320, hs631) located in paralogous loci in Human next to the genes ZNF503 and ZNF703, respectively, which are still alignable between Human and Medaka. Both genes exist in Medaka in two copies together with several syntenic genes in their vicinity, indicating that these loci have been generated by the whole genome duplication in the teleosts. Interestingly, when using LastZ to align hs631 (next to human ZNF703) to the medaka regions identified by the pipeline for both enhancers, it maps at a higher identity to the ortho-paralogous enhancer locus for hs320 near medaka ZNF503 than to its orthologous counterpart. As the paralogous duplication obviously happened prior to the teleost-tetrapod split, one would expect that both loci acquired independent mutations, which subsequently were kept conserved between the corresponding orthologous loci after the split. This should result in alignments between the orthologs followed by weaker hits in the ortho-paralogous regions – at least given that the paralogous sequences are still similar enough to be alignable. Repetition using BlastN however further confirmed the ranking reported by LastZ. These enhancers were therefore good candidates to not only test the sensitivity of different alignment-free approaches but also their specificity. They furthermore served to adjust several parameters of the alignment-free metrics (e.g. word length, window size, target window overlap).

2.2 Alignment-free metrics

Obviously, more than 50% of the VISTA enhancer set cannot be identified by an algorithm specialized to identify non-coding regions in distant species. The easiest explanation would be that those enhancers have just acquired that many mutations during evolution that they lost any sequence similarity – and most likely their function as well. On the other hand, the functional motifs within the corresponding enhancer might just have been permuted rendering it invisible for alignment algorithms. In that case, alignment-free algorithms, which are designed to cope with rearrangements, should be able to reveal some of those eventually hidden (also termed “covert” [68]) elements.

2.2.1 Classical metrics

To investigate to what extent alignment-free operating metrics are able to identify putative enhancers purely by direct sequence comparison, several classical alignment-free metrics were implemented and examined. Some of these metrics (termed here “COSINE” and “D2”, for details see “Materials & Methods”) are taken from a comprehensive review on different alignment-free approaches [62]. Further metrics were proposed by [79] (“POISSON”) and [64] (“HEXDIFF). With the exception of the “COSINE” metric, each of the used metrics has already been used in alignment-free algorithms for enhancer prediction. The “POISSON” metrics for example form the basis of an algorithm for enhancer prediction in intergenic regions in different Drosophila species [67] while “D2” and “HEXDIFF” were both used for prediction of several classes of enhancers [66] in Drosophila and Mouse. In that study, the latter metric, although rather simple, was able to outperform many different modifications of the “D2” metric that implemented sophisticated statistical methods to calculate similarity. Nonetheless, due to its use for many different tasks [61,80,81], the “D2” metric was included in the set of tested metrics. It would have been possible to just use one of the already existing implementations of those metrics but several parameters argued against. First, as previously mentioned (see 1.3.2.3 and 1.4) many of those algorithms require additional information besides the input enhancer sequence (“query”) and the genome of interest (“target”). Second, it was planned to subsequently

use those metrics for genome-wide scanning. But as most implementations were only used on narrow genomic regions, rewriting of the source code would have been necessary to adapt them to this approach. Last but most important, these metrics had been modified in the respective algorithms to meet the specific needs of the project scope. As the aim of this approach is to identify possible problems introduced by allowing permutation for prediction, especially when having only limited information, these modifications might have interfered with the readout. I therefore implemented the aforementioned unmodified metrics myself and used them for prediction of the selected test candidates.

Enhancer	Target region	locus	LastZ	BlastN	HEXDIFF	D2	POISSON:Overrepresented	POISSON:Distinct	POISSON:Additive	POISSON:Product	COSINE
hs320 (ZNF503)	**ZNF503_1o2_ol2**	*ol2:chr15:20310742-20311499*	***30503***	***531 bits***				**1**			
	ZNF503_2o2_ol2	*ol2:chr19:9344622-9345379*	***21037***	***421 bits***				**2**			
	ZNF703_2o2_ol2	*ol2:chr12:3073907-3074664*	***6030***	***89.7 bits***				**4**			
	ZNF703_1o2_ol2	*ol2:chr9:25149830-25150587*	-	***68.0 bits***				**3**			
hs631 (ZNF703)	**ZNF503_1o2_ol2**	*ol2:chr15:20310742-20311499*	***9434***	***141 bits***	**4**				**2**		
	ZNF503_2o2_ol2	*ol2:chr19:9344622-9345379*	***9311***	***134 bits***	**3**	**4**			**3**		
	ZNF703_2o2_ol2	*ol2:chr12:3073907-3074664*	***6837***	***111 bits***					**1**		
	ZNF703_1o2_ol2	*ol2:chr9:25149830-25150587*	-	***96.9 bits***	**2**	**3**			**4**		

Table 2 *Ranking of orthologous/ortho-paralogous loci for alignable test candidates based on either alignment or implemented alignment-free algorithms. Dark marked loci are those with highest syntenic evidence. Interestingly, HEXDIFF prefers both orthologous loci for hs631 compared to all other metrics*

2.2.1.1 Test candidates

All implemented metrics are able to clearly identify the signal in the putative orthologous regions (531 bits and 421bits, locus specific BlastN alignments) when running the hs320 enhancer (ZNF503) against the four identified medaka loci (two for ZNF503 and for ZNF703). No alignment-free signal can be detected in the two ortho-paralogous loci (around ZNF703), which contain much weaker alignments (68 bits and 89.7 bits, locus specific BlastN alignments). Using the hs631 enhancer (ZNF703) as input, all except one POISSON metric (POISSON:Distinct) can find each aligning region ***(Suppl.***

Figure 1). Different to the result obtained by LastZ and BlastN, all metrics report one of the orthologous loci as first hit ***(Table 2)*** - interestingly, this region has less syntenic evidence compared to its duplicated locus. Furthermore, most metrics report the four loci in a ranking similar to that of the alignment algorithms, differing only by shifting the least-syntenic orthologous locus to the first position (but still scoring the second orthologous locus last). Only the HEXDIFF metric reports the loci in the order expected by the presence of the anchor genes (ZNF503 / ZNF703). However, due to the quality of the gene annotation in Medaka and the problems in assignment of correct orthologs between tetrapods and teleosts caused by the additional whole-genome duplication in the latter, ranking of the loci by their gene content might lead to the wrong conclusions. On the other hand, HEXDIFF is the metric with the best signal-to-noise ratio compared to all others. This is surprising as it is also one of the simplest metrics implemented but it confirms the results found by [66]. Although this would speak in favour for the HEXDIFF metric as the metric of choice, one has to keep in mind that even this metric fails to report a signal in the ortho-paralogous loci for the hs320 enhancer next to ZNF503 despite the fact that both loci still contain alignments.

This could indicate that the alignment-free metrics already reach their detection limit at levels of sequence identity for which alignment algorithms are still able to identify the region of interest – even on genome-wide scale. But this does not necessarily mean that regions invisible to aligners also produce no signal using alignment-free algorithms - individual motifs might have just been reshuffled. Detailed analyses of the performed tests support this. In all cases where alignment-free metrics produce a signal, this signal is already visible just by looking at the word counts. All peak regions have an increased word overlap between input and target window that is clearly distinguishable from the surrounding profiles (surprisingly, this signal is even more pronounced than the difference in the score profile of some metrics). Reshuffling of words within a given region would not interfere with that property and thereby still allows the prediction of corresponding regions while at the same time hiding the signal from alignment approaches.

I therefore also used the three human enhancer-candidates undetectable by direct LastZ alignment to Medaka as input for the alignment-free algorithms. As before, I checked several genes in the locus containing the pipeline alignment hit to identify putative additional syntenic regions in the medaka genome that might have been created by the teleost whole genome duplication. For two of the enhancers (hs1022, hs692), an additional putatively orthologous locus can be found containing multiple paralogous genes in syntenic arrangement ***(Table 3)***.

	Orthologous locus	Orthologous genes	Paralogous genes in loci
Hs1022 *(hg19:5:158486120-158487498)*	*ol2:10:16854890-17095899*	EBF1, IL12B	*ADRB2, ABLIM3, AFAP1L1*
	ol2:14:8007407-8271571	EBF1, RNF145, UBLCP1, IL12B	
Hs692 *(hg19:11:15587041-15588314)*	*ol2:3:17465666-17961587*	RRAS2, COPB1, PDE3B, CYP2R1, SOX6, PLEKHA7	*RIC3*
	ol2:6:2178703-2799303	PSMA1, Q05K84_ORYLA(CALCA/ CALCB), INSC, SOX6a, C11orf58, PLEKHA7	
Hs20 *(hg19:16:72738568-72740149)*	*ol2:3:18905429-19258428*	ZFHX3, DHX38, NFAT5	

Table 3 *Putative orthologous loci in Medaka for 3 selected candidates identified by the pipeline. For hs20 no second locus could be found.*

BlastN alignments on those loci do not show any significant sequence similarity to the enhancer. Interestingly, for hs1022 no similarity can be detected at all in any of the two syntenic loci. This indicates that the tunnelling of the human enhancer through the zebrafish genome is crucial for the identification of the corresponding region in Medaka. I performed the alignment-free runs as before and analyzed it to identify windows peaking above a set threshold (see “4.8 Alignment-free metrics”). Of the two weakly alignable candidates, surprisingly only the weaker aligning one (for hs692) can be identified although only by peaks at the lower end of the peak ranking ***(Figure 2)***. Neither the strongest aligning candidate nor the one without significant alignment overlaps with any reported peak. On the contrary, most metrics fail at least in one of the loci predicted by the pipeline to produce a

peak at all. Only the HEXDIFF metric is able to find peaks in all three pipeline-loci but surprisingly fails to identify aligning region for hs692 predicted by the other metrics. However, closer examination of the score levels reveals that the scores in this region are very close to the threshold ***(Suppl. Figure 2)***.

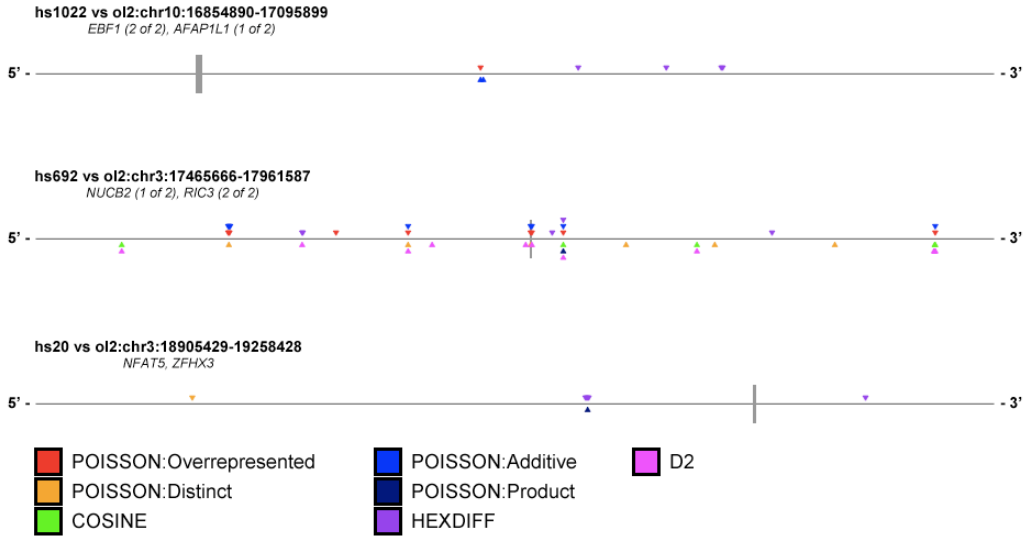


Figure 2 *Peak position plot for all implemented alignment-free metrics on the target loci identified by the alignment pipeline. Only for the medaka target locus of hs692 some metrics peak (coloured triangles) at the position of the identified alignment (vertical grey bar).*

In general, the signal-to-noise levels are 7- to 19-times lower than in the previous runs for the test candidates (hs320, hs631) and in many cases even below one ***(Figure 3)***. This indicates that the metrics indeed operate at the limits of their detection levels. In sum, results show that the implemented classical alignment-free metrics, if given nothing more than two sequences, reach their limit already at rates of sequence identity that still allow reliable identification of the region of interest by alignment of the enhancer against a given locus. On the other hand, the ability to detect a region and to discriminate it from other loci containing similar sequences is not directly correlated to the observed alignment scores, indicating that alignment-free metrics can utilize information that is either invisible or ignored by the used alignment algorithms.

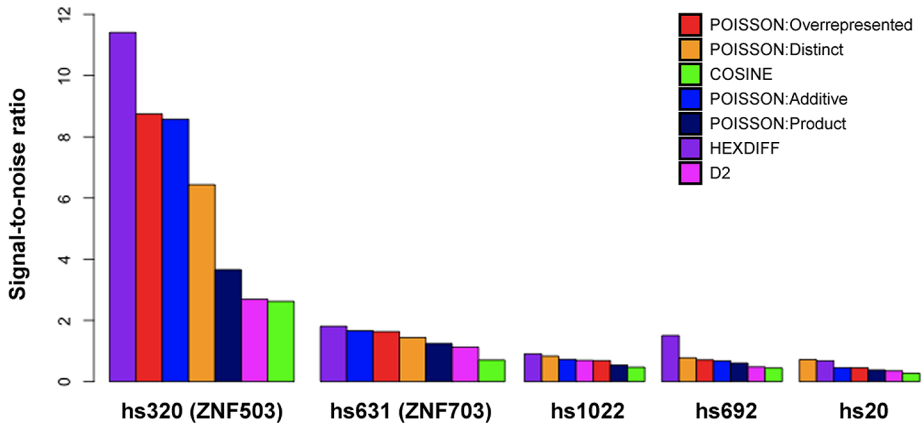


Figure 3 *Signal-to-noise ratios for all alignment-free metrics. Even for the weaker, still directly aligning (hg19 → ol2) test candidate (hs631) the signal to noise ratios are higher than for any of the regions only identifiable by the pipeline. Except for hs20, HEXDIFF always performs best.*

Analysis of the results shows that alignment-free metrics face two major challenges that both lead to signal loss. Single point mutations in the target sequence change all words overlapping them, which in turn changes the word profile. This way, even few mutations can alter the whole profile, especially in short input sequences like enhancers. This problem also occurs when two words detach and rearrange as it eliminates all motifs overlapping the break point. The other problem leading to signal loss is the significance reduction for single words caused by splitting them into many smaller overlapping fragments. Long words obviously have a higher significance than the small words they are composed of. As these words are not required to occur in the same overlapping fashion in the target sequence, each individual nucleotide in the query is virtually multiplied by the words overlapping it. This way it might be repeatedly scored in the target independent of its preceding or succeeding nucleotides. But as these words are smaller than the word they derive from, this rather increases noise than signal. Due to that, even a region containing a strong alignment signal resulting from a long stretch of perfect sequence identity can “drown” in noise peaks predicted in unrelated genomic regions if the enhancer size is large enough to allow the generation of an equally scoring amount of small “noise-words”.

Most recent metrics solve this problem by assigning different weights to individual words. Word filters just keep certain subsequences defined a priori while ignoring all others, which is equivalent to setting their weight to zero ("direct filtering"). Others selectively increase the importance of specific words ("balancing"). This, for example, is done by the POISSON and HEXDIFF metrics. Only the HEXDIFF however also uses "implicit filtering" by scoring only words that exist in the profiles of query and target sequence, all other metrics score at least all words in the query profile. This fits well to the observation that HEXDIFF has the best signal-to-noise ratio, followed by the POISSON metrics, D2 and COSINE ***(Figure 3)***. D2 seems to benefit from the fact that it scores profiles for different word sizes simultaneously but the extent of which is rather low. Implemented POISSON and HEXDIFF metric are theoretically also able to score words of different size at the same time but face two problems in practice. First, motifs used for assessment of similarity are normally derived from additional data, which is not available in this approach. Second, both require the calculation of word background frequencies, which is computationally expensive for larger words (>10nt), especially if the maximum size of a word matching between query and target is not known a priori.

2.2.2 Extended metric

Based on these findings I implemented a new metric that uses "implicit filtering" of variable-sized words and "balancing" simultaneously. This should show whether these concepts could help compensating the adverse effects of permutation on sequence comparison. The extraction of variable-sized words however is different from the strategy used by the D2 metric. For D2, both sequences are individually split into word profiles of a certain range of sizes. Profiles of the same word size are then independently compared and the scores for each word size added. Here, instead of dissecting input and target sequence individually into words of various fixed sizes, both sequences are used at the same time to extract the longest possible word-matches between them (for details see "4.8.5. Modified metric"). Words are then filtered afterwards by allowing only words of the same size to overlap in the target, as

they basically are equally significant. Other overlapping words are either truncated or discarded if they drop below the minimum word size. This way, only words contained in both sequences are scored (i.e. “implicit filtering”). As mentioned earlier, the extraction of words of variable size at runtime does not allow calculation of exact word frequencies a priori. Therefore, frequency values are assigned to all nucleotides in the target genome. The weight of a specific word (“balancing”) is then calculated by averaging across the mapped nucleotides (for details, see “4.5 Frequency track” and “4.8.5 Modified metric”).

2.2.2.1 Orthoblocks

I first tested the new metric on the three candidate enhancers identified by alignment pipeline. These could not be detected before in an alignment-free manner although two of them still contained BlastN alignment hits (64.4 bits and 78.8 bits for hs692 and hs20, respectively). This time, a signal is detectable in each of the three target loci but not in the alternative syntenic regions. This signal is not only clearly distinguishable from the surrounding sequence but also overlaps in all cases with the previously reported alignments ***(Suppl. Figure 3)***. This shows that it is possible to use variable-sized motifs in alignment-free techniques even if they are not specified by additional data a priori. Subsequently, I also applied this metric for all enhancers that were not directly detectable when aligning the human sequence against the medaka genome using LastZ. First, for each enhancer a set of regions is extracted from the medaka genome based on syntenic arrangements of genes orthologous to those in the surrounding of the human element. Then, regions are scanned and the highest peak for each enhancer selected. Peaks next to at least one orthologous flanking gene are selected as putative candidates. In total, for 18 out of 377 (4.8%) enhancers a putative candidate can be identified. 10 of 18 (56%) can be further confirmed by locus-specific BlastN alignments ***(Figure 4)***. The remaining eight have no obvious sequence similarity to their input enhancer. Previous approaches, which also focused on orthologous intergenic regions, had shown that weakly alignable regions only identifiable in that narrow scale still had detectable enhancer

activity [82]. But even these studies would have missed the additional elements found here due to the lack of alignable sequence. One of those additional candidates is even located between both orthologous flanking genes and in comparable relative distance. This is only the case for 2 additional peaks among the 18 candidates (3 of 18 in total). Comparison of the alignment strength found in the BlastN-overlapping candidates identified by this metric to those in the test loci, which are identifiable by the implemented alignment-free metrics, shows that using variable sized motif can clearly increase sensitivity (weakest alignment found: 51.8 bits vs. 96.9 bits). 3 out of 10 overlapping candidates (30%) were even close (51.8 bits) to the significance threshold for BlastN (50 bits).

2.2.2.2 Genome-wide

I then tested whether this sensitivity is still achievable on genome-wide scale. For this, all 377 human enhancers were scanned against the full medaka genome. Unfortunately, this time only five candidates could be identified by the metric, all overlapping BlastN hits, whereas the previously identified novel regions were all lost – even the double-flanked candidate. Surprisingly, BlastN can identify all five even on genome-wide scale although LastZ, which is specialized for aligning non-coding regions, did not report any of those regions. More detailed analysis of the lost and still found enhancer candidates showed that there is no correlation between the contained alignment strength and the fact that a candidate was lost.

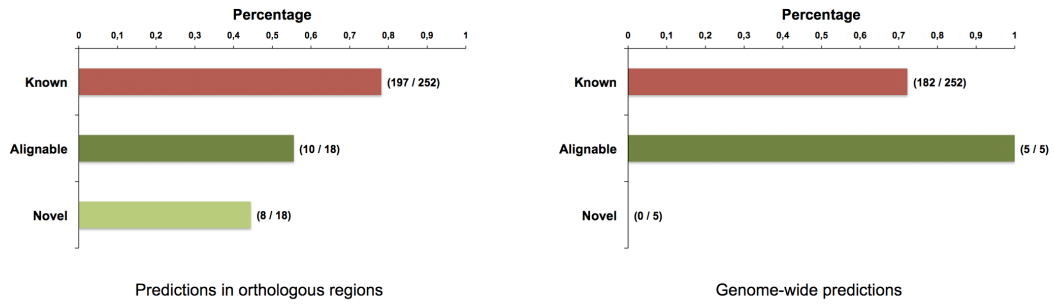


Figure 4 *Result summary for modified alignment-free metric. All novel candidates identified in Orthoblocks are lost on genome-wide scale. The remaining candidate regions are all detectable by BlastN alignments.*

While 2 of the 3 weakest aligning regions (51.8 bits) were still found, the strongest one (122 bits) and several other significantly aligning regions were lost. To test whether this also affects the set of 252 regions for which LastZ did report an alignment, this set was scanned as well against the medaka genome, resulting in only 182 (72%) enhancers that could be identified. Repeating the same run against orthologous regions computed in the same way as for the non-aligning sub set, increased the overlap by only 6% (197 of 252, 78%) ***(Figure 4)***.

2.2.3 Conclusions

Among the regions hidden to the alignment-free approach are also three regions of the highest alignment category defined by BlastN (i.e. >200 bits). These regions have a length between 300 – 600nt, a percentage identity of 68% to 76%, and contain only few gaps (2 - 15%). Sequences are partitioned into perfect matching motifs of various sizes mainly by single point mutations or short mismatching regions (longest motif between 14nt and 19nt) ***(Figure 5 and Suppl. Figure 4)***. The regions predicted by the alignment-free metric for those enhancers have motifs between 21-23nt. As both region sets always also contain additional smaller motifs that balance each other, the final decision for one or the other region is based on the longest contained motif. While this explains why the most likely wrong region is preferred by the metric it does not explain why those motifs exist. Analyses of the predicted loci show no evidence for orthologous genes at these positions. Although this could be the result of wrong orthology mapping between Human and Medaka, the fact that two of the aligning loci are clearly identified when scanning the computed orthologous regions argues against that explanation. A perfect matching 21-mer between human and medaka however should be a very unlikely event outside orthologous regions.

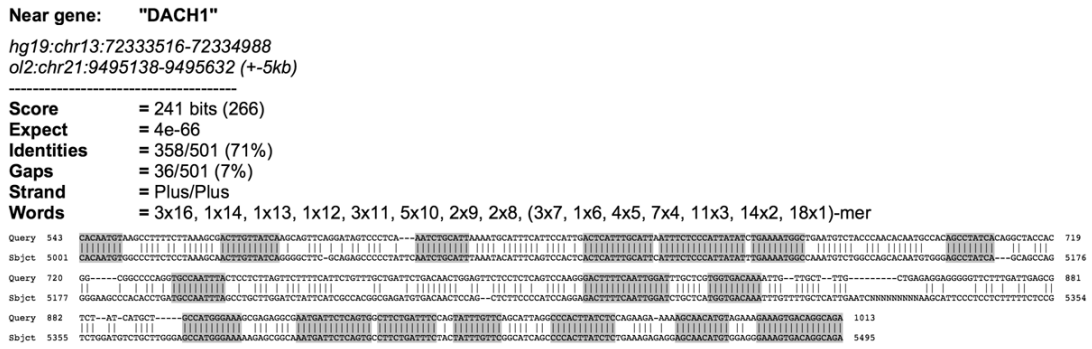


Figure 5 *Example for word pattern in highly significant aligning region. Individual perfect matching words (grey areas) are mostly separated by only a few mutated nucleotides. Resulting words are too small to carry enough signal that would allow identification on genome-wide scale.*

To test how strong deviations from the expected word frequencies might influence the correct identification of the corresponding region by the modified metric, I analysed the repeat masked sequence of the smallest medaka chromosome (i.e. chr19, ~23.5mb, assembly: ol2/MEDAKA1) for occurrence of all possible 20-mers. A 20-mer has a likelihood of 4^-20 ~= 9e-13. This means, it should occur only once ~1,300 full medaka genomes (genome size: ~800mb, ol2/MEDAKA1). As ***Table 4*** shows, ~340,000 different words (~2.1% of all 20-mers found) are contained more than once, the most frequent ("GATTTCATGTAATCCATGGA") occurs even in 189 copies – and that on a chromosome which contains only ~3% of the medaka genome.

This clearly shows that likelihood assumptions for motifs above a certain size largely deviate from the genomic reality. Long motifs should therefore be given less weight to avoid artifacts. Short motifs on the other hand occur in such high amount per scanned target window, that their signal drowns in the noise. This can be shown by analysis of 14 still alignable enhancers selected at more or less equidistant positions in the median score distribution (0.011 – 0.077, stepping ~0.005). The median score represents the genomic background signal and is correlated with the enhancer length as more noise motifs are accumulated. This also results in reduced signal-to-noise levels.

	Repetitive	Most common
6-mers	4096	**AAAAAA** *(25677 copies)*
8-mers	65536	**AAAATAAA** *(4251 copies)*
10-mers	966642	**AATAAATAAA** *(580 copies)*
12-mers	3640346	**ATAGCACAAGGG** *(359 copies)*
14-mers	1666708	**ATAGCACAAGGGTT** *(301 copies)*
16-mers	572522	**GATAGCACAAGGGTTA** *(275 copies)*
18-mers	390041	**TTTCATGTAATCCATGGA** *(199 copies)*
20-mers	336044	**GATTTCATGTAATCCATGGA** *(189 copies)*

Large motifs in ol2:chr19:1-23451325		
GATTTCATGTAATCCATGGACACCAGTGA	**29-mer**	*~180 copies*
CTAGTGGGTCTAGATGACCC	**20-mer**	*175 copies*
TAACCCTTGTGCTATC	**16-mer**	*254copies*

Table 4 *Word counts for different word sizes extracted from medaka chromosome 19. This chromosome obviously contains several large words that differ drastically from likelihood assumption and affect all word profiles of a size smaller than these artifacts.*

To determine what leads to this reduction, the relative abundance of words of a certain size in all scanned windows as well as the amount of words of that size per window were analyzed. As it can be seen in ***Figure 6***, independent words of size 12 and below not only exist in more than 50% of all windows on average but also accumulate per window with increasing enhancer length. As windows overlap by 25% during scan (see “4.8.5 Modified metric”), this means that the target genome contains a matching motif of that size at least every second enhancer length. Unfortunately, many strong alignments are dissected into motifs of that size due to point mutations and/or small indels. Therefore, the signal contained in those regions drowns in the noise of randomly matching motifs. Together with the previously described problem of artifacts generated by words of 20nt or more this is a huge challenge for alignment-free predictions on a genome-wide scale.

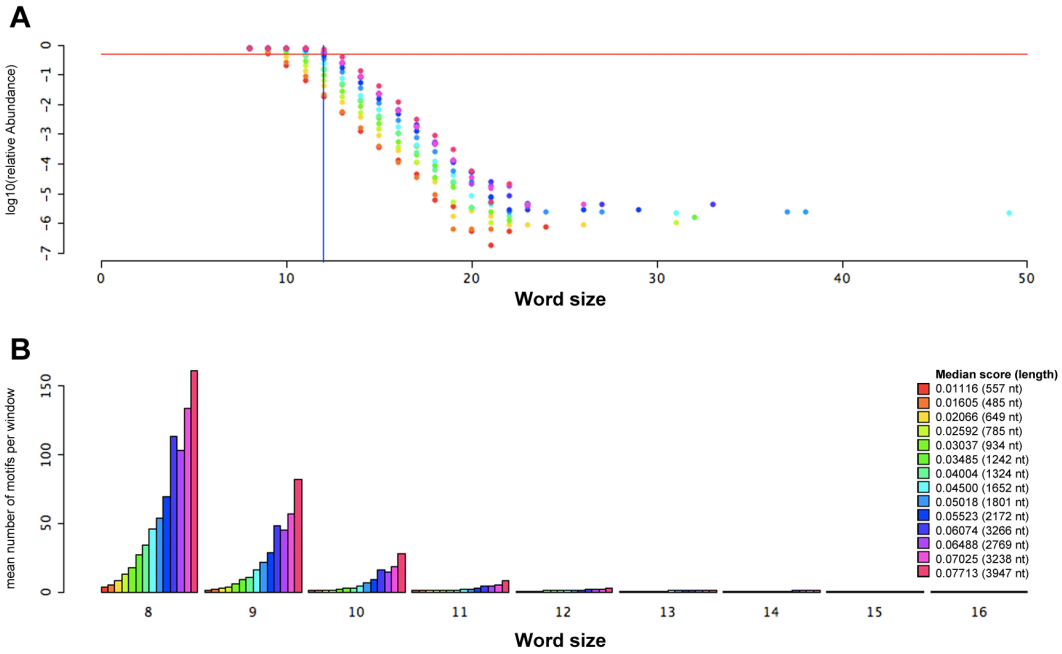


Figure 6 *Word accumulation analysis. (A) Relative abundance of words in windows for different enhancer (i.e. window) sizes. Words of size 12 (blue vertical line) and below occur in more than 50% of the windows (red horizontal line) for most enhancer sizes. (B) Words of size <= 12 not only occur in 50% of the windows but also accumulate per window.*

2.3 NASCAR

The results obtained using alignment-free metrics show that it is possible to identify candidate regions of likely enhancer activity in the genome just based on the comparison of profiles of matching words between two sequences, and that even on genome-wide scale. No additional data like TFBS sets or conservation profiles are necessary. Although none of these regions was tested in an in vivo assay, previous studies could show that regions of similar properties identified using alignments were able to drive a reporter construct in an enhancer-like fashion [82]. The observations also highlight that the main problem generated by allowing rearrangement in sequence comparison is the loss of significance of the individual matching segments by fragmentation into smaller pieces. Small words are more likely to occur at higher frequencies, offer more possibilities for permutation and thereby increase noise. While this still can be handled when focusing on narrow regions around orthologous genes, it causes substantial problems on genome-wide scale. Increasing the significance of individual words by perfect match extension can reduce adverse effects but is challenged by long matching segments that exist much

more frequent in the genome than expected. Furthermore, extension is interrupted in many cases by short interspersed mismatch mutations that dissect regions of high sequence similarity into many small, noise prone segments. Alignment algorithms solve these problems very effectively but aim to detect single regions that are significant by themselves. This is contrary to alignment-free approaches, which are based on the simultaneous scoring of multiple matching elements within a given region. As the results for the variable perfect-match metric show, a combination of both principles is possible in theory. Mismatch extension allows combination of small perfect matching words into larger mismatch containing motifs, which can then be scored in an alignment-free manner allowing motif reshuffling. This principle has the potential to work for all classes of enhancers: enhanceosomes, which are forced by their working principle to keep their functional elements in a collinear arrangement, could be indentified even if they are rather short as the metric would treat them as a whole instead of dissecting them into small pieces and thereby dispersing the signal. Changes in variable positions could also be handled as long as enough sequence identity is kept. Rearrangement, which is a prominent feature of billboard enhancers, can also be handled by such a metric as it allows motifs to occur at rearranged positions without affecting the score. The recently describe type of structured enhancers [40,70] should benefit the most from this principle. Swanson et al. describe that enhancers are composed of individual binding sites on the one hand and groups of sites following a strict grammar on the other. These groups contain sites, which need to be precisely positioned to allow interaction of the bound TFs. In this way they should form longer motifs that could be recognized by mismatch extension. They could also show that these groups can rearrange as a whole and even vanish and reappear by binding site turnover. This would also be covered by the principle described here.

2.3.1 Principle

To test whether the combination of alignment and alignment-free properties can be used for enhancer prediction, I implemented them in the “NASCAR” algorithm (*“Non-linear Alignment SCoring AlgoRithm”).* Knowledge gained

from previous alignment-free approaches was thereby included in the design. Key features of this new algorithm are: window-based scanning of a given target sequence, usage of mismatch containing motifs to increase signal of each individual word, weighting and filtering of motifs according to their relative importance, sum scoring of all matching elements contained within a given window, and allowance for full permutation (rearrangement & strand swap) of the motifs involved. In short, motifs are detected in a similar way as in the alignment-free metric used for genome-wide prediction, just that instead of allowing only perfect-match extensions, mismatches are included using a scoring function known from alignment algorithms. The detected motifs are then weighted, filtered and used to generate a similarity score between the input enhancer and each individual window in the target sequence (for details see “4.9 NASCAR”).

In addition to the basic metric, I also implemented a pattern detection technique to increase signal in non-permutated but far-spaced regions. This was thought to be of special help for the detection of structured enhancers as the CRMs that form the full enhancer are likely to be farther spaced than the sites within one module. Mutation of those module spacers, especially in large regulatory elements would leave the functional elements intact but split them in several, still collinear arranged fragments. Knowing that sequence within functional elements or groups evolves at slower rates than the spacers between them [40], this would not only reduce signal strength but at the same time increase the noise level by generating new, random motifs matching just by chance to different parts of the input region, thereby blurring the signal. To compensate these effects, motifs still arranged in a collinear fashion are strengthened. This way, even motifs below the noise-threshold (see 2.2.3 and 4.9.2) can be considered. This technique should also help to reduce permutation noise that can occur even without the generation of new motifs. In an algorithm like the proposed here, every single motif is independent form each other in the sense that they neither overlap in the input nor the target sequence. This way they are free to permute as long as they finally reach another non-overlapping configuration. But as all possible permutations result in the same score, a still collinear arrangement, which is a way less likely

event to happen, cannot be discriminated from fully mixed up arrangements. The described pattern detection technique however enhances the score of collinear regions above mixed sequences and thereby allows their discrimination.

2.3.2 Sensitivity

I first tested the final algorithm on the aligning data set to assess its sensitivity and specificity. When run against the full medaka genome it can identify 246 of 252 (~98%) of all alignment hits found by LastZ, only 238 of 252 (~94%) however also overlap with BlastN alignments. This is partially due to the fact that LastZ and BlastN overlap by only ~97% (246 of 252). Nonetheless, NASCAR reaches an overlap of >90% in general which is a clear improvement compared to the 72% reached with the purely alignment-free technique. As NASCAR is a hybrid between an alignment and alignment-free algorithm this result was expected for the aligning enhancer set.

2.3.3 Prediction

Having successfully identified most of the aligning subset, I subsequently used NASCAR to detect enhancer candidates for elements in the non-aligning set. As most candidates for the aligning set are directly flanked by at least one orthologous gene (~77%), this criterion was also used to call candidates for the non-aligning enhancers. In total, NASCAR predicts putative candidates for 30 of 377 enhancers (~8%), 10 of them located even between both their orthologous flanking genes. Due to the fact that many motif permutations can lead to more or less the same score, NASCAR cannot be as specific as alignment algorithms, meaning that candidate regions might not necessarily score first (similar to the situation for only weakly aligning regions). Furthermore, scores are strongly correlated to the input size as always a whole window is scored. This makes it very difficult to define a score cutoff for significant candidates like the bit score of BlastN. I therefore tested until which position in the peak ranking flanked candidates can be found. More than 50% of all flanked candidates are already found as first peak, ~75% until position 12. To include even weak signals in the candidate set, I set the threshold to

25 (~84%). Application of this threshold on the alignment set increases the overlap with LastZ further, reaching 99% (249 of 252). For the non-aligning set, this results in 48 candidates, 14 still between both orthologous flanking genes. To assure, that no candidates are selected for further testing, which could have been identified by an alignment algorithm on genome-wide scale, I used BlastN on the medaka genome and selected the 25 highest alignments per input enhancer, including even those that were not considered as significant hits (bit score <50). Afterwards, I filtered all enhancers and their candidates further by removing all enhancer-candidate pairs for which the suggested medaka peak is next to an orthologous gene and within 5kb around a BlastN hit. This assures that each selected candidate is indeed uniquely found by NASCAR and results in a final set of 9 human enhancers for which at least one peak is found next to an orthologous flanking gene (see ***Figure 7*** and ***Table 5***). For hs882:hg19, even two peaks very close to each other (4kb) can be identified (hs882:ol2-1, hs882:ol2-2). This second peak overlaps the aligning region of another VISTA enhancer (VISTA ID: hs431), which was not included in the compiled dataset as no enhancer activity was observed in the initial mouse assay. To rule out that they may be paralogous enhancers, I aligned those candidates against each other in all possible combinations (hg19→hg19, ol2→ol2, hg19→ol2, ol2→hg19) using BlastN. This did not show any significant similarity. This candidate is thereby very interesting as it might be a "redundant" enhancer [83] with highly similar activity.

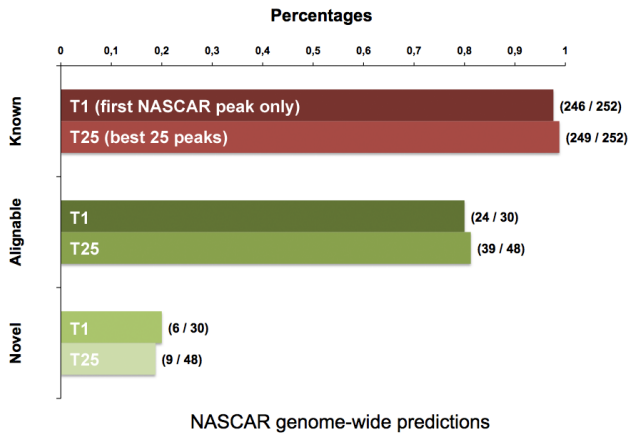


Figure 7 *NASCAR results for genome-wide prediction of aligning and non-aligning VISTA enhancers.*

VISTA ID	Loci	Name	Rank	Score	Pattern score	Pattern motifs	Status	Orthologous genes
hs1831	*hg19:7:95236622-95240458*	*hs1831:hg19*	3	14075	-	-	SF	COL1A2, CASD1, SGCE, PPP1R9A, ASB4, **PDK4 * DYNC1I1**, SHFM1, DLX6, DLX5
	3837 nt							Gene order: 1, 2, 3, 4, 5, **6**, **7 * 8**, 9, 10
	ol2:11:9757778-9761614	*hs1831:ol2-1*						COL1A2, CASD1, SGCE, PPP1R9A, ASB4, **PDK4**, **DYNC1I1 *** SHFM1, DLX6, DLX5
hs865	*hg19:6:50685244-50686237*	*hs865:hg19*	4	3918	-	-	DF	**TFAP2D * TFAP2B**
	994 nt							Gene order: **2 * 1**
	ol2:24:19553675-19554668	*hg865:ol2-1*						**TFAP2B * TFAP2D**
hs1535	*hg19:2:60498057-60502013*	*hs1535:hg19*	7	11386	-	-	DF	VRK2, **FANCL * BCL11A**, PAPOLG, REL, PEX13, KIAA1841, XPO1
	3957 nt							Gene order: 4, 5, 6, 7, 8, 1, **2 * 3**
	ol2:15:7226825-7230781	*hs1535:ol2-1*						PAPOLG, REL, PEX13, KIAA1841, XPO1, VRK2, **FANCL * BCL11A**
hs590	*hg19:18:34719386-34720720*	*hs590:hg19*	9	5121	-	-	SF	**C18orf10 * CELF4**
	1335 nt							Gene order: **1**, **2 *** ()
	ol2:5:16698603-16699937	*hg590:ol2-1*						**C18orf10**, **CELF4 *** ()
hs882	*hg19:13:71533037-71534195*	*hs882:hg19*	12	4452	-	-	SF	**KLHL1 * DACH1**
	1159 nt							Gene order: () * **2**
	ol2:21:9414695-9415853	*hs882:ol2-2*						() * **DACH1**
hs394	*hg19:2:59746377-59746992*	*hs394:hg19*	23	2666	-	-	DF	VRK2, **FANCL * BCL11A**, PAPOLG, REL, PEX13, KIAA1841, XPO1
	616 nt							Gene order: 4, 5, 6, 7, 8, 1, **2 * 3**
	ol2:15:7017013-7017628	*hs394:ol2-1*						PAPOLG, REL, PEX13, KIAA1841, XPO1, VRK2, **FANCL * BCL11A**
hs882	*hg19:13:71533037-71534195*	*hs882:hg19*	1	5814	1790	4	SF	**KLHL1 * DACH1**
	1159 nt							Gene order: () * **2**
	ol2:21:9408987-9410145	*hs882:ol2-1*						() * **DACH1**
hs848	*hg19:16:51491799-51493025*	*hs848:hg19*	1	6629	2811	4	SF	TMEM188, HEATR3, PAPD5, ADCY7, BRD7, NKD1, NOD2, CYLD, **SALL1 * CHD9**, RBL2, AKTIP
	1227 nt							Gene order: 12, **10**, 11 * **9**, 8, 7, 6, 5, 4, 3, 2, 1
	ol2:3:29251639-29252865	*hs848:ol2-1*						AKTIP, **CHD9**, RBL2 * **SALL1**, CYLD, NOD2, NKD1, BRD7, ADCY7, PAPD5, HEATR3, TMEM188
hs1049	*hg19:5:92314781-92316083*	*hs1049:hg19*	1	5775	2048	3	SF	POLR3G, MBLAC2, GPR98, LYSMD3, **ARRDC3 * NR2F1**, FAM172A, ANKRD32, KIAA0825, MCTP1
	1303 nt							Gene order: ***(5?)*** * **6**, 7, 9, 9, 8, 10, 5, 3, 4, 1, 2
	ol2:9:15278121-15279423	*hs1049:ol2-1*						***(ARRDC3?)*** * **NR2F1**, FAM172A, KIAA0825 (1 of 2), KIAA0825 (2 of 2), ANKRD32, MCTP1, **ARRDC3**, GPR98, LYSMD3, POLR3G, MBLAC2
hs1344	*hg19:3:193660817-193662478*	*hs1344:hg19*	2	5131	2031	4	DF	**OPA1 * HES1**, ATP13A3, TMEM44, LSG1, FAM43A, C3orf21, ACAP2
	1662 nt							Gene order: 8, **2 * 1**, 3, 4, 5, 6, 7
	ol2:4:13569031-13570692	*hs1344:ol2-1*						ACAP2, **HES1 * OPA1**, ATP13A3, TMEM44, LSG1, FAM43A, C3orf21

Table 5 *Candidate regions predicted in Medaka for 9 human VISTA enhancers. For hs882, two regions very close to each other (hs882:ol2-1, hs882:ol2-2) could be predicted but only hs882:ol2-1 contains a collinear motif cluster. Four peaks are located between both orthologous flanking genes ("double flanked"; DF), 6 still next to one ("single flanked"; SF).Orthologous flanking genes are marked in bold face, relative peak location is indicated by asterisk. Gene order depicts arrangement of orthologous genes in the medaka locus compared to Human. Empty parenthesis indicate non-orthologous genes flanking the peak*

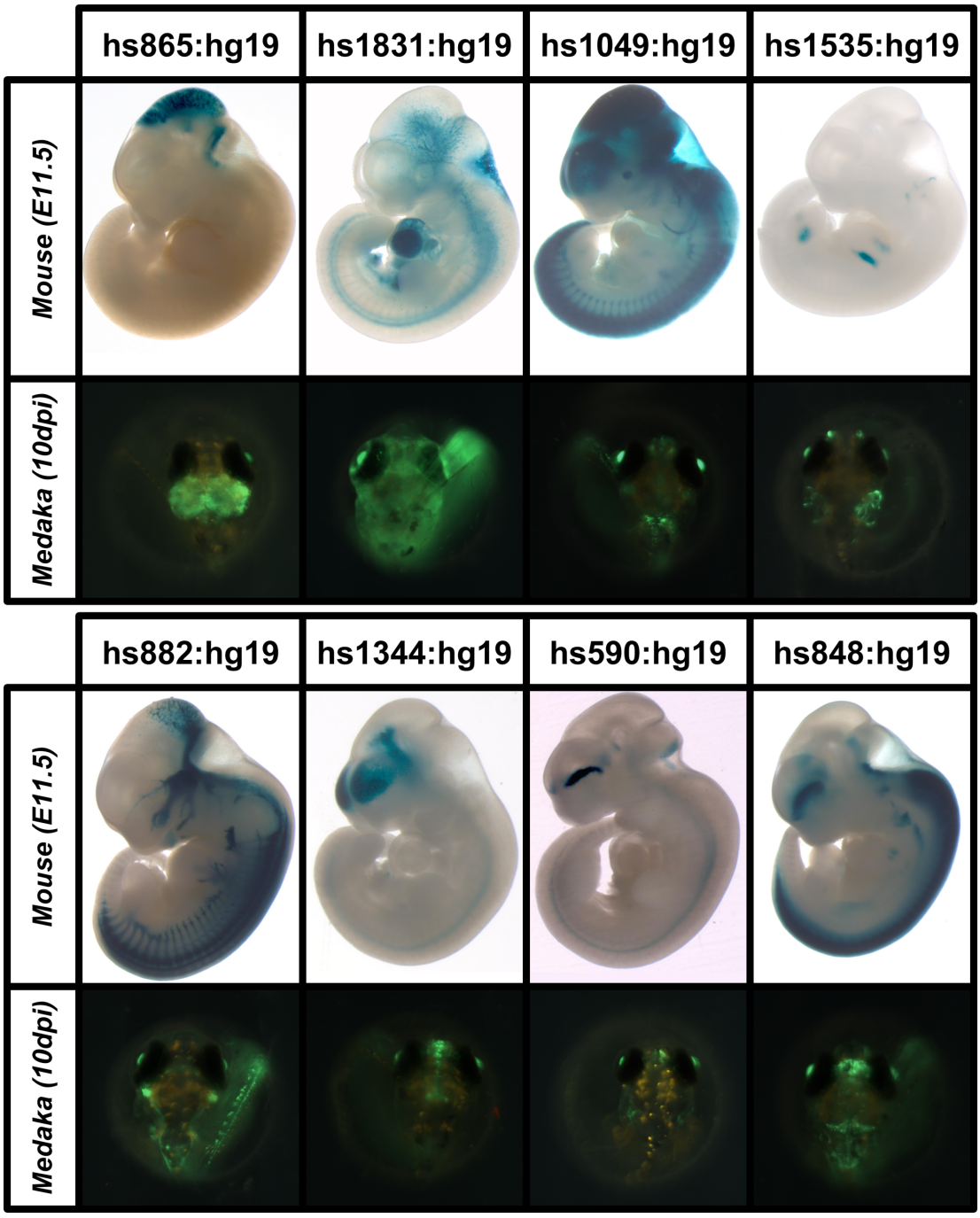


Figure 8 *Comparison of activity patterns oh human VISTA enhancers in Mouse and Medaka (dpi=days post injection). Hs394 is not shown as it has no activity in Medaka. Mouse pictures taken from the VISTA Enhancer browser [LINK1].*

2.3.4 In vivo validation

The final set of candidates consists of 9 human enhancers and 10 corresponding medaka regions ***(Table 5)***. I cloned each of these regions from the human or medaka genome into a reporter construct and tested the construct in the medaka fish for enhancer activity. This in vivo enhancer assay was established in the lab and has also previously been used to study the activity of enhancers [84]. Eight of 9 human enhancers, which already had been validated in Mouse, also show activity in Medaka in comparable structures, mainly in the brain and other neuronal tissues ***(Figure 8)***. Of the corresponding medaka regions, 7 of 10 show activity as well ***(Figure 9)***. For 3 of them (hs865:ol2-1, hs848:ol2-1, hs882:ol2-1), activity is even partially similar to the corresponding human enhancer as they drive reporter expression in comparable tissues. Hs882:ol2-2 (the second peak for hs882:hg19) was of special interest, as it is not only predicted using the same enhancer as input but also located very close to the first candidate, which makes it a likely redundant enhancer candidate. In contrast to the aligning enhancer in the VISTA data set, which was negative in the initial validation in Mouse, the predicted medaka region indeed shows enhancer activity. However, in contrast to the expectation it shows a different activity pattern compared to the first candidate. The predicted medaka region for the negative human element hs394 is negative as well as expected.

2.3.5 Conservation analysis

Most enhancers that show activity in medaka also contain blocks of sequence that are conserved within the teleosts but interestingly not between the teleosts and placental mammals. I therefore analysed the motif profiles of all enhancer-candidate pairs to explore, whether these motifs might be conserved between Human and Medaka but just too short to allow the detection of deep sequence conservation across that phylogenetic distance. For this, I used PhastCons scores [75] for placental mammals and teleosts downloaded from UCSC *[LINK3, LINK4]* and calculated the average conservation score per motif for Human and Medaka.

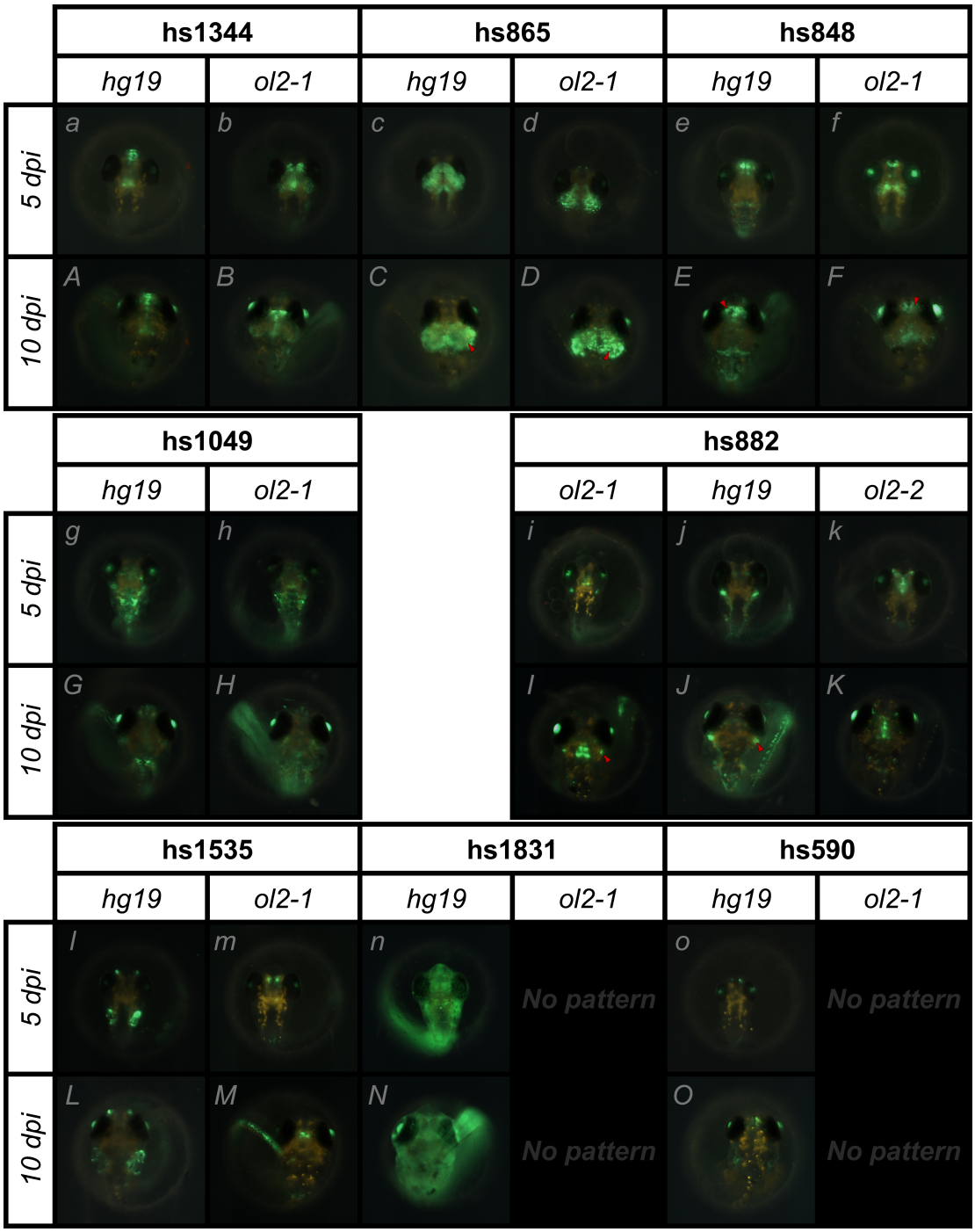


Figure 9 *Validation results for NASCAR-predicted medaka regions and the corresponding human VISTA enhancers. For some constructs, red arrows mark structures that are similar in the activity patterns caused by the human and medaka region.*

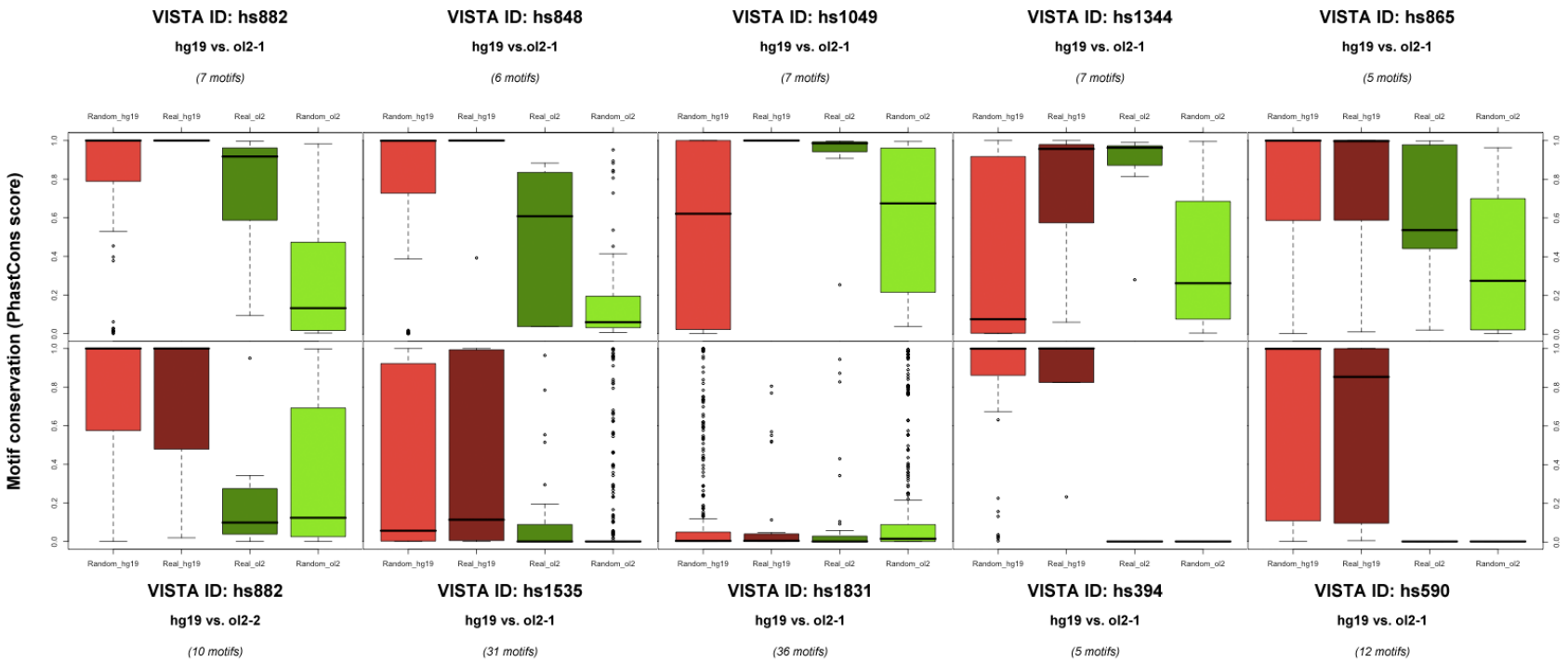


Figure 10 *Motif conservation boxplots. Bars in brighter colours on the left and right side of each plot show results for randomly selected motifs in the corresponding human or medaka loci. Darker bars in the centre show real motifs. Median motif conservation (horizontal black bar) in medaka in higher than in any random set picked from the same locus for the five upper enhancer-candidate pairs, for hs882:ol2-1, hs1049:ol2-1, and hs1344:ol2-1 even significantly higher (p-value < 0.01, wilcoxon rank sum test). Situation in Human is similar, with median conservation as high or higher than in random sets (hs882:hg19 = p-value <0.05; hs1049:hg19 = p-value < 0.01; wilcoxon rank sum test). Hs1344 for instance shows very high conservation of NASCAR motifs, which is much stronger than for the corresponding random sets. Furthermore, median levels in randomisations show that there is a large amount of non-conserved sequence in both loci. This clearly indicates that, although the majority of the sequence in both regions has acquired mutations within its clade, the indentified motifs were kept conserved.*

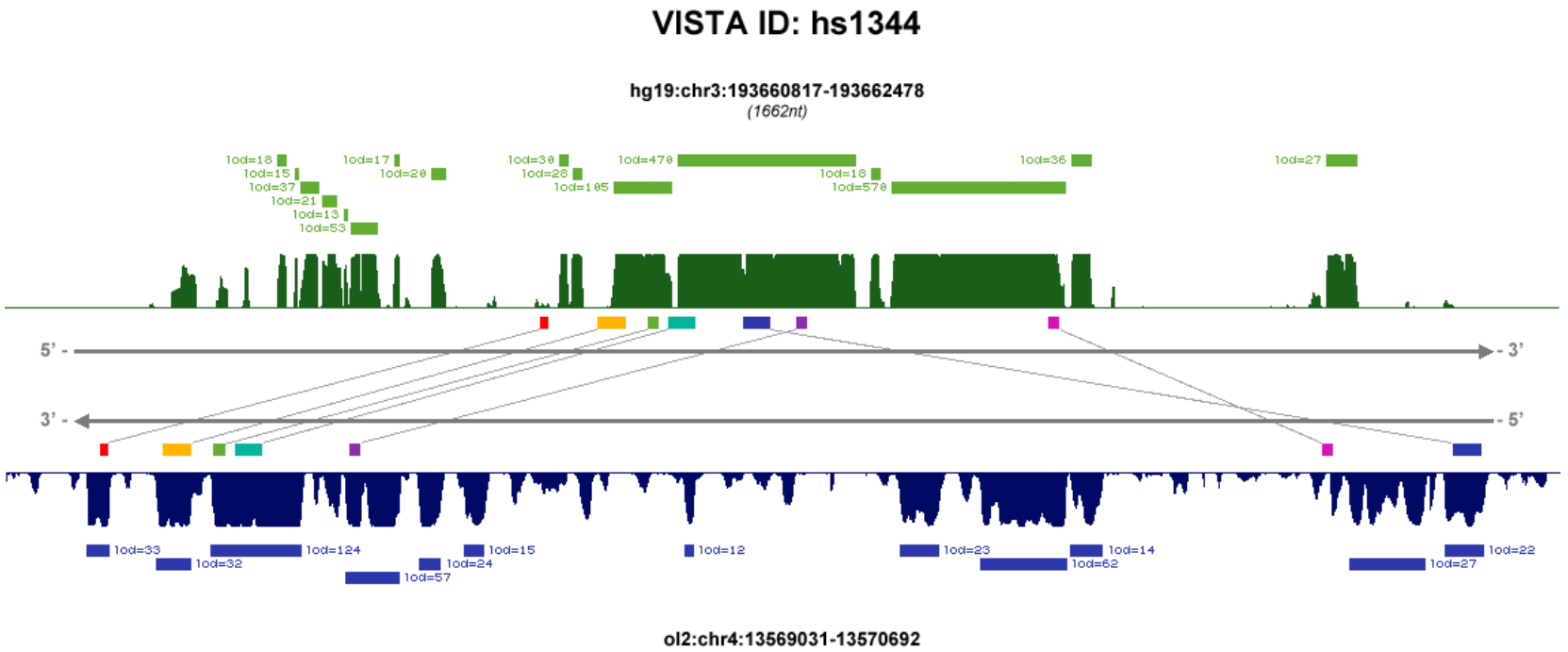


Figure 11 *Detailed analysis of NASCAR-motifs for hs1344 enhancer-candidate pair. PhastCons scores for placental mammals including Human are shown in dark green, scores for teleosts including Medaka in dark blue. Brighter coloured blocks above and below show predicted conserved elements with attached lod-scores ("log odds"). NASCAR-motifs are shown in rainbow colours between both tracks and corresponding motifs are connected (grey lines). Although drawn next to each other, there is no relation between the two sequences except those regions forming the motifs. One collinear cluster with almost identical spacing of motifs is clearly visible. Interestingly, the red motif is not conserved in Human but located in a small conserved block in Medaka. The blue motif, which seems to have changed its relative position compared to the others, is also located in one defined conserved block although there is enough non-conserved sequence. This indicates that these motifs contain functionally relevant information that, although rearranged, was kept conserved independently in both clades.*

To rule out that eventual high conservation levels just derive from the fact that the full regions (Human, Medaka, or both) are under constraint, I extracted 10 random motif sets from the same region and also analyzed them for their conservation. For 5 of the 10 candidates, the median conservation of the involved motifs in Human and Medaka is clearly as high or even higher than for the random sets, partially also at lower scattering around the mean (see ***Figure 10***). Three of them even show no conservation scattering of human motifs although the corresponding random motifs are variable in their conservation. This indicates that the NASCAR motifs in these human regions are specifically conserved compared to the surrounding sequence. In sum, those 5 enhancer-candidate pairs show clear signs of motif level conservation in both species although PhastCons cannot find any direct conservation from Human to Medaka ***(Figure 11)***.

2.3.6 Motif analysis

As a next step, I analysed the motif arrangement in Medaka for each of the candidates by using Human as reference. All 5 enhancer-candidate pairs that show motif level conservation also show clearly collinear motif arrangements ***(Figure 12)***, 3 of them (hs848:ol2-1, hs882:ol2-1, hs1344:ol2-1) as dense clusters of 4 to 5 motifs. This further strengthens the conclusion that small-scale conservation throughout the vertebrates was missed in those cases. One candidate (hs1049:ol2-1) shows a rather loose arrangement of motifs with two motifs near the 5' end that seem to have undergone a local inversion. The last of these (hs865:ol2-1) is harder to assess, as it seems to have 3 motifs still in collinear arrangement that are interrupted by a reshuffled motif. As NASCAR predicts two possible candidates for one of the enhancers (hs882:hg19) it was of special interest whether the involved motifs also overlap. This would argue for a similar function of both enhancers. Both peaks however seem to utilize completely different motifs in the human enhancer ***(Figure 13)*** which is in accordance with the validation results.

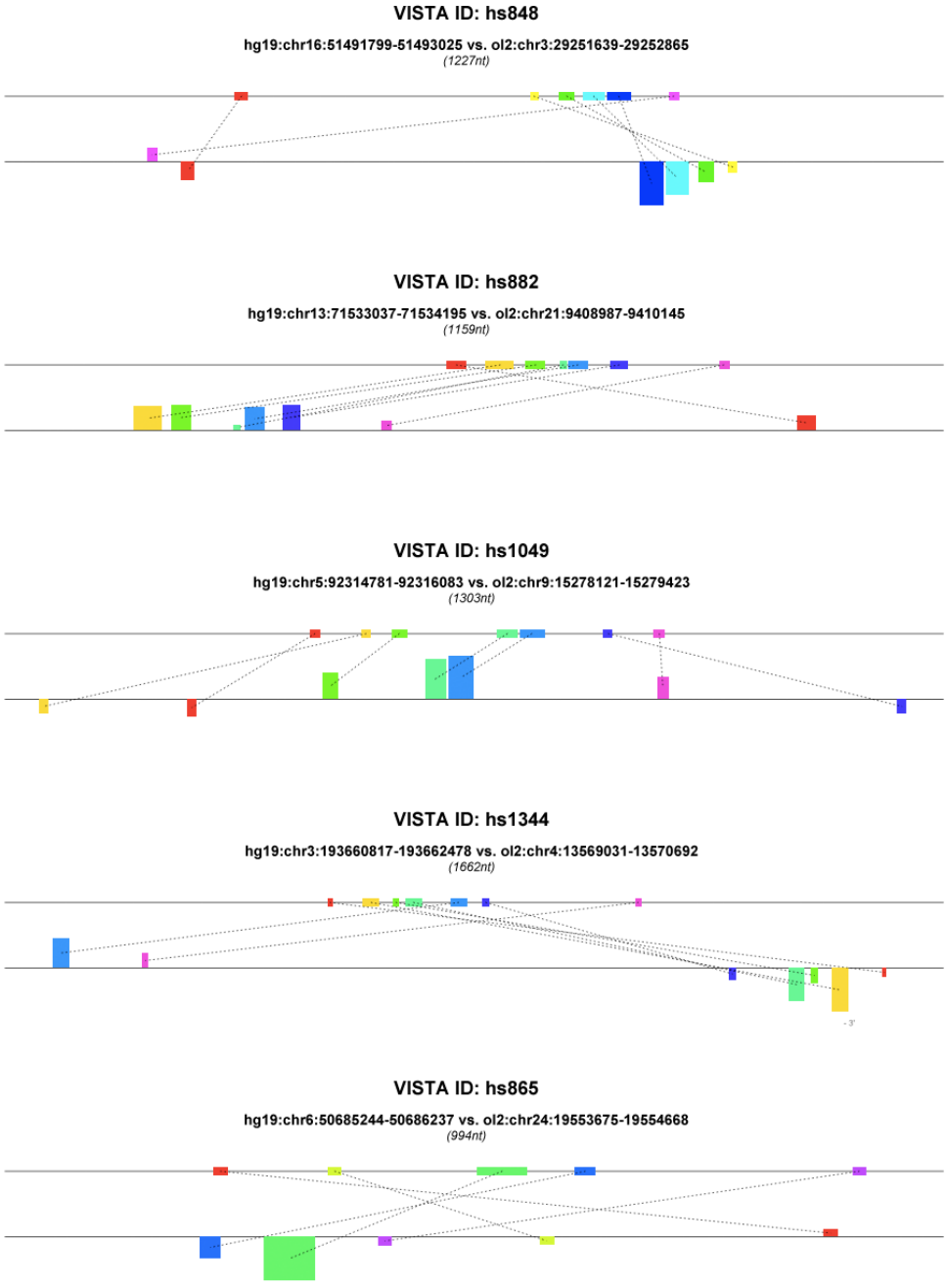


Figure 12 *NASCAR-motif arrangements for all enhancer-candidate pairs that contain collinear patterns. Pattern in hs865 was not found by the implemented detection technique as it requires at least three motifs and the yellow motif is too far away.*

Of the 4 candidates that showed no conservation within the teleosts (hs394:hg19, hs590:hg19, hs1535:hg19, hs1831:hg19), none showed any collinear motif arrangement but completely mixed profiles ***(Figure 14)***. This could explain why 3 of them show no enhancer activity for the medaka region. On the other hand, hs882:ol2-2 shows strong activity despite a completely mixed profile – as well as hs1535:ol2-1.

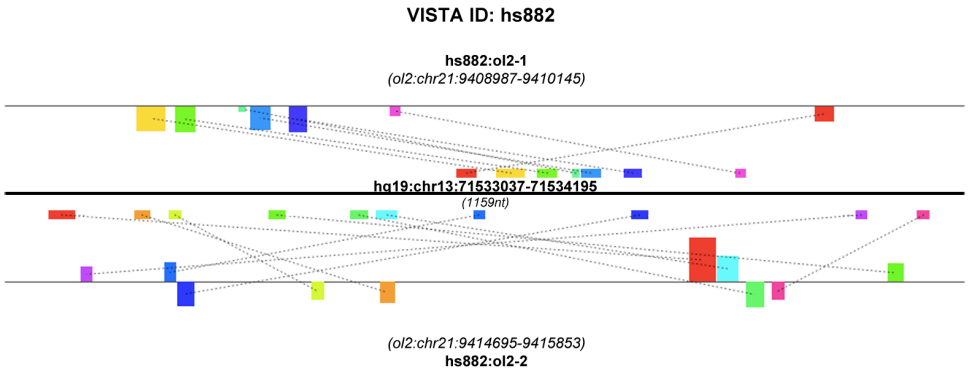
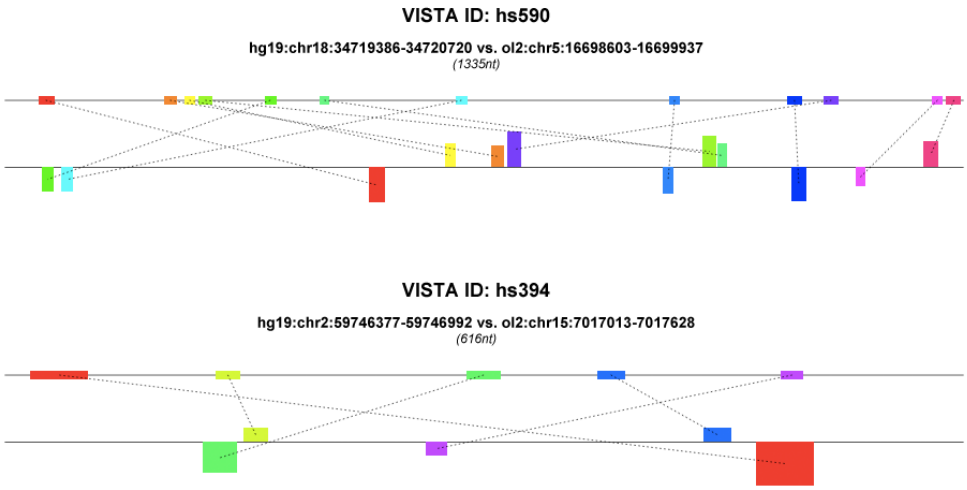


Figure 13 *Query motif usage for both hs882 enhancer-candidate pairs. Obviously, completely different regions in the query (middle two motif rows) were used for prediction of hs882:ol2-1 and hs882:ol2-2.*

Figure 14 *Examples for shuffled motif profiles. Hs590:hg19 showed activity in the enhancer assay but the predicted medaka region (hs590:ol2-1) did not. Hs394 (hg19 & ol2-1) was completely inactive.*



To test whether the found conserved motifs are also functional, I scanned the remaining 24 peaks per enhancer for a second candidate sharing the highest possible motif overlap with the already tested region whilst excluding as many collinear motifs as possible. These candidates should serve as kind of natural

deletion constructs and their activity compared against the previously tested candidate for which the collinear cluster of motifs was artificially deleted ***(Figure 15)***.

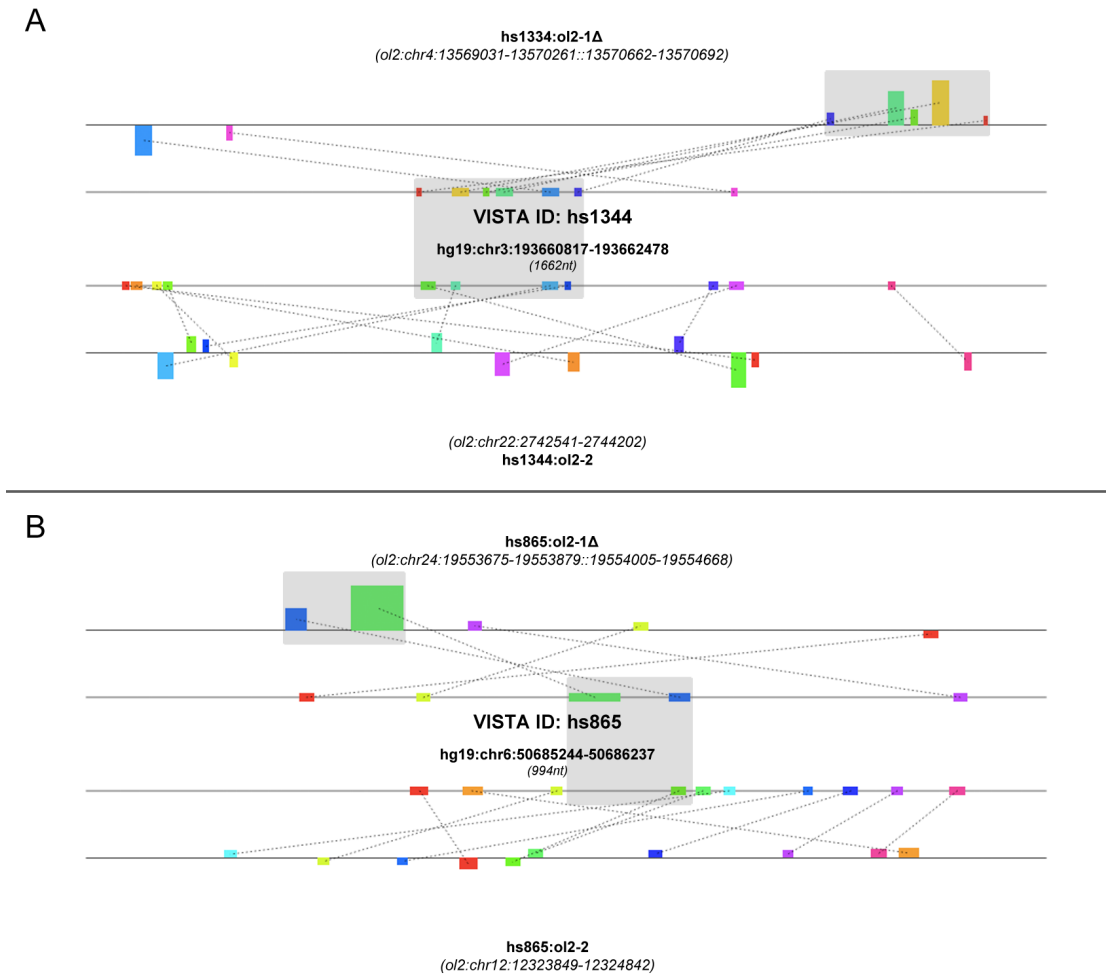


Figure 15 *Deletions and alternative constructs for hs1344:ol2-1 (A) and hs865:ol2-1 (B). Motif usage in Human is always shown in the central two lines. Motifs deleted in the medaka construct are highlighted in grey. Alternative constructs in both cases share some motifs inside and outside of the deleted area.*

Surprisingly, the patterns of both deletion constructs (hs865:ol2-1Δ and hs1344:ol2-1Δ) remain rather stable and show little to no reduction as result of the deletion ***(Figure 16)***. Quite the contrary, an additional domain seems to emerge in hs1344:ol2-1Δ. Furthermore, none of the additional peaks (hs865:ol2-2, hs1344:ol2-2) is similar to either the tested candidate or the deletion construct. Hs1344:ol2-1, -1Δ, and -2 have overlapping activity in the optic tectum, but this is only very faint in hs1344:ol2-1 while very strong in hs1344:ol2-2.

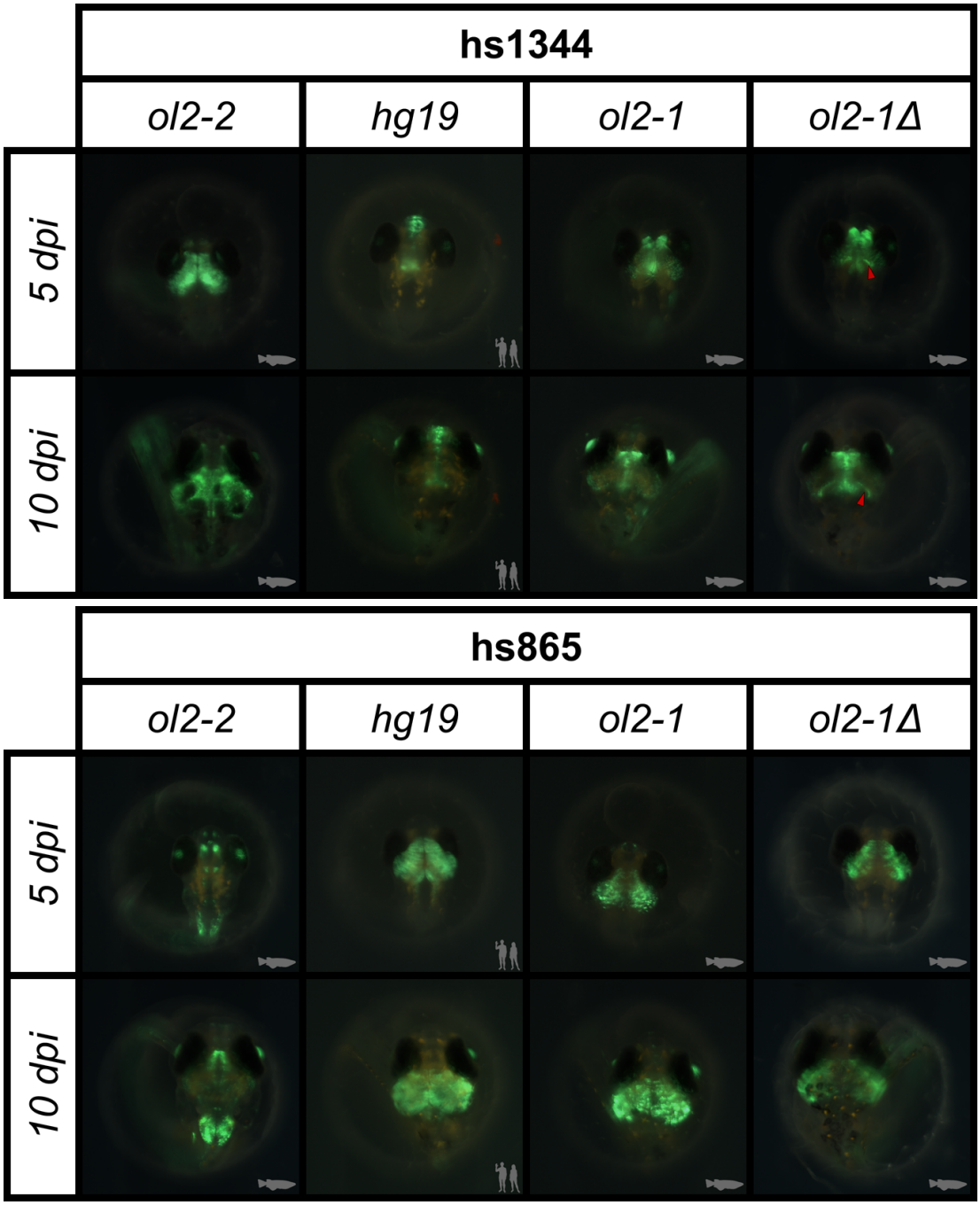


Figure 16 *Comparison of original constructs with deletions and alternative regions. Hs1344:ol2-1Δ shows an additional domain (red arrow) but otherwise looks like the undeleted construct. Both alternative constructs (hs1344:ol2-2, hs865:ol2-2) show strong enhancing activity but not similar to any of the other tested regions.*

2.3.7 TFBS analysis

As TFs are the main actors for enhancer activation, I tested whether patterns of exclusive TFBS enrichment can be found in the motifs predicted by NASCAR. As expected, even at a conservative threshold of 90% similarity to the used PWMs, hundreds of putative TFBSs are found. Even when

restricting the search space to the NASCAR motifs only, still 50 to 100 TFBSs can be identified. This is partially due to the fact that different factors of the same family bind to the same motif and are reported as individual hits but also caused by overlapping sites of unrelated factors. For 18 regions (8 in Human and 10 in Medaka), motif-specific TFs can be identified but do not differ significantly when compared to the set of randomly selected motifs (p-value: 0.22 for human and 0.19 for medaka active enhancers, wilcoxon rank sum test). I also tested whether the same TFs are restricted to the identified NASCAR motifs in both species. If TFBSs depletion is one mechanism of TF-guidance and the identified TFs involved in the enhancer function, sites for those factors would have to be motif-exclusive in Human and Medaka. Indeed, a significant enrichment can be found (p-value: 0.003, wilcoxon rank sum test) for regions with validated enhancer activity. Due to the small number of positive candidates (only 7 candidate pairs showed activity for cloned regions of Human and Medaka) this result might be strongly influenced by a single outlier. Closer inspection of the identified TFs shows that for one pair 7 restricted TFs are identified but all overlapping exactly the same spot. This artificially increases the amount of TFBSs found. I therefore removed this candidate pair from the data set and repeated the analysis. This time, the p-value drops to 0.013 (wilcoxon rank sum test) but still shows a significant enrichment.

3. Discussion

3.1 Data set selection

To date, the VISTA Enhancer Browser provides one of the largest collections of in vivo validated enhancers. Most of the contained regions were either predicted by ChIP-seq on the transcriptional co-factor p300 or deep sequence conservation. The fact that this co-factor is involved in many enhancer complexes [43,53] allows generation of sets of tissue-specific enhancers based on very different underlying enhancer logics. This way, the resulting sets allow the study of enhancer mechanisms in an unbiased fashion. Most importantly, it does not introduce biases to enhancer sequence structure, as it binds to already assembled TF clusters rather than to the sequence itself. Conservation on the other hand introduces a strong bias to sequence structure. In this thesis, I made the attempt to predict enhancers by a hybrid method situated between alignment and alignment-free techniques. This should especially allow detection of enhancers that have undergone structural changes and permutations during evolution. Due to these changes, this class is likely to be missed by alignment-based approaches using deep sequence conservation but at the same time of great interest for studying regulatory evolution of species. Using the enhancers in the VISTA database as starting material for the prediction of permuted regulatory regions in a distant species might however come with several problems. Although there might be many reasons why the corresponding counterpart in Medaka is not detectable by LastZ, the most likely explanation is that it just mutated to an extent that destroyed its function. It is of course also possible that subunits rearranged and therefore hid the enhancer, but the fact that it was initially found by deep sequence conservation somehow argues against that. Why should an enhancer region, kept conserved from Human down to Chicken or Frog, suddenly be allowed to rearrange? Mutational events in one species or clade of course do not depend on events in any other lineage, but for some reasons this did not happen in a large variety of different vertebrates, most likely because the enhancer under investigation is of crucial importance for the organism carrying it. However, the whole genome duplication that happened

shortly after the split of the teleosts from the tetrapods [85] might have provided the playground for mutational changes as it relieved the selective pressure on at least one of the copies. In this case, the second copy should still be detectable by alignments. Otherwise, one enhancer would have had to mutate for some time while still two copies existed of which the not mutated one was lost at some point, leaving no trace of any of the two copies. Although this scenario is possible it seems to be unlikely.

Another possible scenario would be that the enhancer was neither copied nor lost but mutated in a fashion that hides it from alignment algorithms without affecting its function. This does not necessarily have to be permutation-related. Billboard enhancers are known to be partitioned in individual TFBSs or functional modules (CRMs), separated by intercalated spacer sequences [36,38,70]. These sequences evolve much faster than the functional modules, at a rate that is close to the genomic background [70]. This could result in "erosion" of sequence similarity in the spacer regions, leaving "islets" of regulatory function behind (see ***Figure 17***). Furthermore, the modules themselves might acquire mutations in the spacers between the TFBSs or even at variable positions within. In the end, the accumulated changes can reduce sequence similarity to an extent that does not allow identification by alignment algorithms anymore. Considering that enhancers are rather short in general, this might happen quite easily. This scenario does not require any permutation or even complete turnover of functional regions as only functionally unimportant positions are affected. It therefore might be a more likely explanation why some of the input enhancers do not seem to have a counterpart in Medaka.

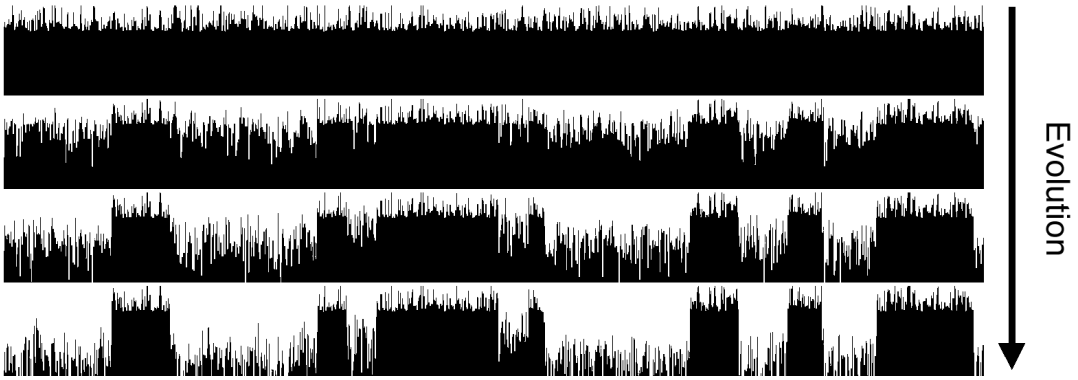


Figure 17 *"Sequence erosion": accumulated mutations in non-functional regions of an enhancer result in small, functionally conserved blocks which are too small to produce a significant alignment on their own but at the same time too far spaced to be combined by gapped alignment.*

As described, the data set extracted from the VISTA Enhancer Browser might not be the ideal starting material to search for permutation-based evolutionary changes. The results found in this study somehow support this conclusion, as no clear signs of permutation or turnover are visible in the validated enhancers. This of course assumes that the applied detection method is able to find permutation and that the motifs predicted in the enhancer regions are indeed involved in the function. Nonetheless, as long as no large validated enhancer sets are available that were generated by less biased prediction methods, the VISTA Enhancer Browser represent one of the most valuable sources of starting material for prediction approaches.

3.2 LastZ vs. BlastN

LastZ [74,77] is a widely used and generally accepted fast alignment algorithm applied especially for the alignment of non-coding regions, partially because of its high specificity [77]. Due to elevated rates of sequence evolution in non-coding regions, reliable identification of the "correct" ancestral region is crucial. This is even more important when considering segmental and even whole genome duplications, which have occurred several times during the evolution of life. In combination with its speed, LastZ is therefore the ideal tool for comparative genomics, especially for the assessment of deep sequence conservation over large evolutionary distances. LastZ (more

precisely its predecessor BlastZ) hence forms the basis for multiple alignment algorithms like MultiZ [76] and this way also for methods used to calculate scores of evolutionary constraint (e.g. PhastCons [75], PhyloP [86]) or tools for visualisation of evolutionary conserved regions (ECRs) (e.g. zPicture [87]). Deep sequence conservation on the other hand has been very successful in the identification of enhancer regions in the past [34,59,60] and is still used for enhancer prediction. In a highly dynamic sequence environment like non-coding regions, stable segments are likely to carry important function for the organism as they were preserved by negative selection. This makes LastZ also very valuable for the analysis of gene regulation and regulatory evolution. In this context, it is quite surprising to notice several highly aligning regions in the VISTA data set being missed by this algorithm. Of the 629 regions extracted from the VISTA Enhancer Browser [73], only 252 produce a direct alignment hit between Human and Medaka on a genome-wide scale while for 377 enhancers no significant alignment is detected. BlastN [78] however is able to identify a significant hit for additional 55 regions while missing only 4 previously found by LastZ if only the best BlastN hit is used. Extending the BlastN search beyond the first hit (like for LastZ), BlastN reports 336 highly significant (>=80bits) and even 480 significant (>= 50bits) hits for the full VISTA set – compared to 371 for LastZ. Although this could be the result of different parameters specified for both algorithms, the difference is too large to be purely explained in this way – especially as 4 of the first hits have a bit score of >100 and additional 7 still >80. Furthermore, for both algorithms I used parameters that are optimized for the detection of distant homology. In fact, these are the same settings that were used for the LastZ runs underlying all MultiZ and PhastCons scores including the teleost fish *[LINK7]*. It therefore might be possible that regions conserved from Human to Medaka have been missed in these comparisons and are therefore not reported by PhastCons or PhyloP. Based on these observations, it seems that LastZ sacrifices sensitivity for specificity. This might lead to exclusion of even strong signals and hence to signal loss during the preparation of multiple pair-wise alignments. MultiZ for example uses small aligning regions as anchors, which are subsequently combined into chains of alignment blocks [88]. In case of multiple possible chains, the highest scoring chain is selected first. Other

overlapping chains are then either truncated or discarded. Missed alignment blocks can influence this process with the result that the "wrong" chain is favoured, just because alignment blocks in the "right" were missed. In the end, an unknown fraction of deeply conserved (non-coding) regions could be invisible for conservation-based enhancer detection approaches. It therefore might be advisable to rather perform genome-wide pair-wise alignments using BlastN at sensitive settings than using LastZ for enhancer prediction - especially for distantly related species.

3.3 General problems of the alignment-free principle

Although alignment-free algorithms have properties that make them promising for the detection of enhancers they suffer from some certain drawbacks that make their usage difficult for such narrow-scale approaches.

3.3.1 Relative word significance

This starts already with the generation of word profiles: the selection of an appropriate word size is crucial to obtain a sufficient signal-to-noise ratio. 5- or 6-mers are typical sizes proposed by several studies [64,66,72] but also 8-mers were suggested [71]. Among those, 5-mers have the most limited pool of possible words, which is 4^5 = 1024. Theoretically, a sequence of 1028nt (1024 + word size – 1, as the last word starts on the 1024th nucleotide) could harbour all of them in an overlapping fashion – which of course is a rather rare case that likely does not exist in all currently sequenced genomes. Another estimate would be 5120nt (1024 x word size) if they were placed next to each other ***(Figure 18)***. Although also this case might be rather rare, it shows that windows scanned for a given enhancer sequence of ~1kb might contain many different, perhaps even most of all possible words. This leads to the fact that the calculation of similarity using 5-mers is mainly based on the instances of a given motif rather than on its mere presence. But knowing the rather limited set of 5-mers, how telling is it whether or not there is one instance more or less in a given sequence?

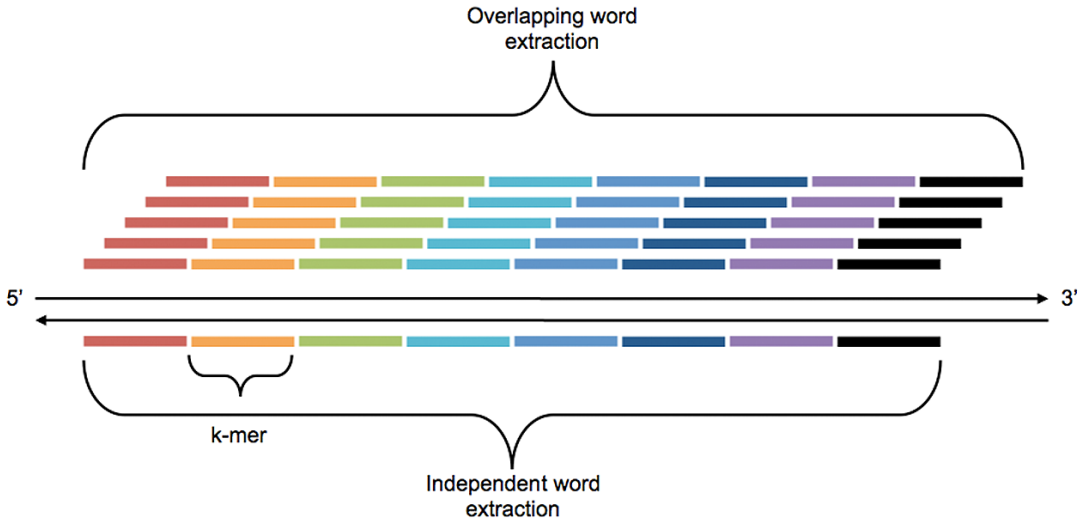


Figure 18 *Word profile generation from a given sequence. Words can either be extracted independently or in an overlapping fashion*

It has to be kept in mind that the used words serve as a kind of substitute for TFBSs within an enhancer. Although there are reports about repetitive occurrences of the same site in certain types of enhancers [89] it is not known whether or not this is a common feature of cis-regulatory elements in general. It therefore might be completely unimportant for the function of an enhancer how many copies are present as long as at least one site exists. Furthermore, sequences of low complexity also contain repeated copies of the same motif without having regulatory activity. Both types of motifs are indistinguishable as long as nothing but a query and a target sequence are given to the algorithm. This is one of the reasons why many published approaches make use of TFBS repositories which either are additionally provided together with the query sequence [69,90] or replace it completely [68], transforming the alignment-free algorithm into a kind of TFBS-clustering method. Without this additional support, thorough masking of repeats is crucial for metrics using small word sizes.

Another way increase word-significance is to simply increase word size. Using 10-mers, the set of possible words exponentially increases to 4^10 = 1048576. In a genome of ~3gb like for Human, each word would have an expected occurrence of 3*10^9 / 4^10 =~ 2860 copies – or better, an

occurrence of one per ~1mb assuming a uniform background distribution. Taking the combinatorial space into account, formed by the amount of different words to the power of all positions for word extraction in an enhancer or given sequence window, this should be more than sufficient to reliably indentify the correct region. However, due to the huge amount of different words possible, each individual word will likely occur only once within a given sequence. This makes accurate likelihood estimates a crucial prerequisite, as similarity will be mainly based on whether or not a word is present and not on the number of its instances. Metrics such as POISSON [79], HEXDIFF [64] and several modifications of the D2-score [66] approach this problem by using statistical or empirical background distributions to adjust the importance of individual words. But as tests on medaka chromosome 19 have shown, reality deviates largely from the assumptions. The Poisson distribution for instance is used to describe rare events that happen independently of each other. This is challenged by the fact that even very large words (up to 20-mers) can occur quite frequently on single chromosomes. Furthermore, as the analysis has shown, these words not always occur independently of each other but also in a highly overlapping fashion (e.g. in case they all derive from one longer word like a 29-mer in this case; see ***Table 4***). As a result, enhancers that by chance contain any of these words will produce extremely high scores in each region harbouring the longer “source” word. Another problem of overlapping word extraction is that words are usually generated at every single nucleotide position. This way, each word has all but one nucleotide in common with the previous word. Therefore, every abundant word strongly affects the frequencies of all overlapping ones, and by that, leads to complex dependencies, which are very difficult to handle.

3.3.2 Word background distribution

As just mentioned, one of the most prominent problems of alignment-free techniques is the completely unknown background distribution of words in the genomes of interest. While the real background frequencies per word can be computed from the genomic sequence, their physical distribution along the chromosomes of the genome cannot. The majority of the genome is

composed of many highly repetitive elements like transposons [91], ancient retroviral insertions, and ancestral segmental duplications [88]. Furthermore, different lineages have encountered several whole genome duplications followed by genomic restructuring and sequence loss. This results in a very clustered distribution of the same set of words in different regions of the genome that is almost impossible to approximate. One consequence of this is that words very frequent on a genome-wide scale might occur in dense clusters forming only a small portion of the genome while more or less rare motifs can be evenly distributed across all chromosomes and thereby are contained in almost every randomly extracted fragment. As statistical models always assume a standard background distribution, being it Gaussian, Poisson, Binomial or others, this very uneven distribution of motifs poses a huge challenge to the calculations. This shows that the reality of genomes and genome evolution can have a tremendous effect on statistical and likelihood-based prediction models.

3.3.3 Mutation

The usage of word profiles comes along with another type of problem. Point mutations and small indels are abundant events in the genomes of all living organism (and even viruses) and one of the driving forces of evolution. Unfortunately, they represent one of the major problems for purely sequence-based alignment-free enhancer detection. Alignment algorithms can partially compensate this, as a single seed (i.e. a k-mer) is enough to start an alignment, which allows to bridge individual or small groups of mismatches - at least as long as enough matching nucleotides follow. Alignment-free algorithms on the other hand suffer from the fact that a single mutation affects all words overlapping it. Although word-loss increases only linear with word-size, its impact on the score increases exponentially due to the exponential increase in significance of a single word. In the extreme case, one nucleotide mutated at every 5^{th} position between to sequences can completely mask any similarity based on 5-mer profiles by affecting every possible word in the profile. This would hide relations between sequences already at 80% similarity – for 10-mers even 90%. Such a regular mutation pattern might be pretty rare

but clearly shows the impact of single point mutations on the similarity score. Using an intermediate motif size is by far not sufficient to mitigate these effects as even the smallest suggested motif size might still fail at 80% identity.

One possibility to deal with point mutations would be to allow mismatches in words. While this might sound easy to do it comes with a severe problem. In the perfect-match word-space, each word in the query sequence is associated with only two words in the target: the identical word and its reverse complement. If words are treated as sense-antisense-pairs, this turns even into a one-to-one relation. Introducing just a single mismatch at an arbitrary position in the query word extends this to a one-to-many problem. In this context, a 5-mer in the query would map on average to 15 different words (so called a “1-neighbours” [71]) in the target – and vice versa (special case of palindromic or low-complexity words is not discussed here for simplicity reasons). This of course includes that several words in one of the sequences can also map to the same 1-neighbour in the other. By that, overlapping “word-spaces” are created which are hard to resolve and most likely lead to signal blurring. In the past, several approaches have been published that allowed for mismatches in their word profiles. While some solved the blurring problem by using aligned sequences of multiple species [65] or reduced it by focusing on sets of enhancers active in a restricted tissue [72], only one did not require any additional information [71]. With their method, they were able to reach a significant overlap between a set of known regulators and their genome-wide peaks but also detected a high number of noise matches. However, these enhancers contained dense clusters of TFBSs and were searched in the comparably small drosophila genome. A test run on a single human chromosome using a set of known human regulators resulted in an even higher rate of false positives. As this approach only allowed for a single point mutation, it is questionable whether it can be extended to full vertebrate genomes using enhancers that contain less dense binding sites and higher rates of mutations. Different ways of dealing with the mismatch problem are therefore necessary.

3.3.4 Permutation

One of the main reasons for suggesting alignment-free algorithms was their inherent ability to cope with permutations as they are based on word profiles rather than recognition of long collinear stretches of DNA. But allowing for permutation in general comes along with several problems beyond those generated by the composition of genomes or the occurrence of point mutations. The most obvious is their main property, namely to score all different permutations of a given input at exactly the same level. Considering the very special case that the input is clearly subdivided into completely independent words, e.g. achieved by insertion of “N”s in the DNA sequence. In this case the number of all possible combinations is n! (n = number of words). Using a 5-mer profile for a region of a typical enhancer size like 1kb leads to approximately 1000 / 5 = 200 words (not considering the positions occupied by “N”s). If each word occurs only once, this leads to 200! different permutations, which is too large to be computed on regular computers (largest value calculated by R 2.8.1: 170! = 7.257416e+306). In case words occur more often, this reduces the number of combinations approximately to u! (u = number of unique words), which largely depends on the used word size “k”. Words forming the word-profile for a given sequence however are generated in an overlapping fashion, meaning that each k-mer depends on the k-1 words before. This leads to even more words and at the same time makes estimations very difficult. Assumptions of Binomial [71] or Poisson [79] distribution of words, as used by several metrics, are therefore only rough estimations. They also rely on another assumption, which is that a profile of very similar or even almost identical composition is very unlikely to occur at any other position in the genome. Considering the patchy composition of genomes and the huge amount of segmental duplications, this is definitely not the case. The fact that most metrics also treat all words equally, and thereby cannot distinguish between completely different profiles as long as the amount of overlapping words is the same, complicates the situation even further. Hence. it is not surprising that the peaks obtained with the classical, unmodified alignment-free metrics were much clearer when looking only at the counts of overlapping words. It also explains why weighting of individual

motifs improves the sensitivity and specificity of alignment-free approaches. This might also be the reason why those methods work best for medium-sized enhancers. For small regions, one word more or less makes a huge difference leading to very spiky scores and many artifact peaks. Large regions on the other hand contain that many words that the increase in matching words in the "correct" region is hardly noticeably against the noise.

3.3.5 Usage of additional information

To overcome the problems described above, many studies included various kinds of additional information into their prediction models to improve sensitivity and specificity of their algorithms.

3.3.5.1 TFBSs

TFBSs represent the smallest known functional subunit of regulatory elements like enhancers and promoters. It is therefore not surprising that many different approaches utilize TFBS data for enhancer prediction. The main benefit of TFBSs for alignment-free methods is that they act as word filter that helps to reduce noise. As previously described, purely alignment-free techniques a priori cannot distinguish different target sequences by their word content as long as query and target profile overlap by the same amount. TFBSs allow filtering of these profiles for those words that match to the specified sites and thereby help to discriminate target regions from noise. By assessing the percentage similarity of a given TFBS and a word in the profile they might also allow to distinguish between regions that have very similar word profiles. High numbers of words that match very well to a specific set of sites might also enrich for regions that are not only bound or active but even very similar in activity to the given query region. As TFBSs are normally specified as PWMs, they help to handle point mutations as well. But at the same time, PWMs are also the main problem. Although they represent the compiled binding information derived from biological binding data (e.g. ChIP-chip, ChIP-seq, EMSA), the combination of many different binding sites into a unified position weight matrix leads to loss of combinatorial information in variable sites. Certain nucleotides might always occur in highly specific combinations in

variable positions of a TFBS and never in any of those other patterns that can be derived from the computed PWM. This co-occurrence might also be very cell type specific. Hence, using PWMs might result in predicted regions that show sites specific for a certain cell type and lead to a highly significant prediction although the input enhancer is active in a completely different tissue. Due to this degeneracy, matching putative TFBSs in query and target sequence will occur more often than perfectly matching words. The usage of TFBSs for enhancer prediction might therefore be counteractive. Furthermore, they can only be used, if the set of factors binding to a given enhancer is known but this is rarely the case. In most instances, the used PWMs were not even determined in the species of interest, assuming that the binding specificities are the same as in the source species for the PWM. It is therefore hard to tell how useful TFBSs are for enhancer prediction.

3.3.5.2 Sequence conservation

Sequence conservation is another commonly used word filter [65,67]. In case multiple alignments across large phylogenetic distance are available for the enhancer of interest, this information can be used to extract words that seem to be under evolutionary constraint. This implicitly assumes that the observed conservation is due to the function. Although this approach also helps to filter motifs and in turn reduces noisy hits during prediction, it is limited to regions that are diverged enough to result in a significant word profile restriction but at the same time are conserved enough to still allow the extraction of a telling word set. However, deep sequence conservation makes it unlikely to find rearranged enhancers. If the observed conservation is indeed functional, strong constraint was acting on the non-mutated regions. It is therefore hard to explain why these sites suddenly should have been allowed to rearrange if no segmental or perhaps whole genome duplication has happened in the meantime. But in this case, one of the copies should be still detectable by alignments. Another downside of this method is that the binding specificities of TFs allow for variable positions that can mutate to some extent without affecting functional binding of the factor. Focusing only on fully constraint regions might therefore miss the wanted binding sites. Sequence

conservation can also be used even if the enhancer is not directly alignable, as long as close by sequence anchors are. In that case, the sequences of many species between those anchors can be analysed for words occurring in the majority of them. This approach makes it more likely to allow detection of rearranged enhancers, but again only if those words carry at least part of the function. It also allows extraction of mismatch-containing motifs [65] in a similar way as TFBS-prediction on ChIP data sets [92,93]. In sum, although conservation data might act as efficient word filter, this information is mostly not available for those enhancers that have rearranged during evolution and therefore are the most interesting.

3.3.5.3 Functional conservation

Functional conservation (e.g. used by [64,72,79]) is by concept very similar to sequence conservation. Instead of a set of aligning sequences ("sequence similarity"), regions of known comparable activity ("functional similarity") are used to extract enriched words or mismatch-containing motifs. This approach suffers from the same drawbacks as sequence conservation with the addition that they derive from the same species and therefore are not further supported by evolutionary constraint.

Summed up, it can be stated that the largest benefit of additional information for alignment-free prediction is the reduction of the extracted word profile to a set of – putative – telling words and by that not only reduces noise but also increases the significance of the remaining words. At the same time, it allows assigning individual weights to words and thereby shifts prediction from a pure overlap count to a word-specific prediction that also allows inclusion of mismatch-containing motifs. However, the feasibility of this approach largely depends on the quality and availability of the necessary information, which strongly limits the possible scope of those techniques.

3.4 Algorithm selection: alignment vs. alignment-free

Knowing the weaknesses of alignment-free algorithms the question arises how helpful these approaches can be in general. As discussed, their main advantage is also their main weakness: dissecting a linear sequence in a set of relatively small words introduces a lot of uncertainty about the significance of individual motifs while at the same time creating a huge permutation space that hinders precise identification of even collinear regions that align very well. They also require some kind of preprocessing step to discriminate between words as they otherwise only measure the amount of overlapping words rather than the occurrence of a specific word pattern. This preprocessing is in most cases achieved by providing additional data that comes with further assumptions and penalties. Alignment-free techniques are furthermore very susceptible to small-scale mutations. As TFBSs in most cases seem to shift their positions by turnover rather than small-scale translocations, it is unlikely that newly generated binding sites are perfectly identical to the lost one. In this case, little to none of the words generated at the new site will correspond to those at the old as they all differ by one to many nucleotides. Although all this can be taken as arguments against alignment-free methods there are little alternatives. Alignment algorithms may be able to handle mutations to some extent and by that allow detection of regions of low sequence similarity. But as binding site turnover in enhancers is a well-proven fact [36,37,41,42,70,94–96], they are only able to find a very limited set of all existing regulatory elements. This might explain why most enhancers detected by deep sequence conservation were close to important developmental genes or regions encoding TFs [59]. Due to their importance, no mutations are allowed that would remove a crucial binding site. This at the same time strongly limits the space for generation of redundant sites and thereby inhibits binding site turnover. On the other hand, as mutations in variable sites are less likely to affect functional binding of the TFs involved, these sites might be still allowed to evolve. This makes them ideal candidates for alignment-based prediction while the contained mutations mask the signal against alignment-free approaches. Obviously, using conservation in such mutated regions as motif filter is unlikely to recover functional motifs. A closer look at the principle of

alignment and alignment-free techniques reveals that both actually are two extremes of the same method: both start with extraction of small, perfect matching words or “seeds” contained in query and target sequence. But while aligners continue with seed extension (perfect match or mismatch) to find a single, highly significant subsequence, alignment-free methods try to filter the “seed” profile to extract a significant pattern that consists of many small sequences. It is therefore not a question of using one or the other but of finding means that allow the combination of both approaches in one technique. In the past, several publications have already proposed methods that make use of both principles [68,69,97]. Some of them were even successfully used for enhancer prediction [68,69]. The main difference between those approaches and the principle used here is that they always required additional data and were applied on narrow genomic regions between alignment anchors, either direct or indirect, instead of being used genome wide. The effect of this locus restriction is described in the next section.

3.5 Search space

Most techniques that have been published so far share the common feature that they were applied on specific regions of the genome. These regions were either delimited by direct or indirect alignment anchors [68,69] or anchored at or between genes [67,82]. A similar approach was chosen here for the analysis of alignment-free metrics by focusing on regions near orthologous genes. As the results show, the restriction to a certain region in the genome helps to reduce the number of false positive predictions and this way facilitates the identification of putative candidates. This finding can be explained by the properties of alignment-free metrics and the genome structure. As mentioned, alignment-free metrics suffer from the fact that a priori all words are indistinguishable. At the same time, the amount of different motifs is limited by the small alphabet of only four bases. This makes noisy matches at any position in the genome very likely as even words up to 11-mers accumulate with increasing enhancer length. The duplicated nature of

most genomes makes this even more difficult. It thereby greatly reduces the chance of false positive hits if the majority of the available sequence is not scanned. If the aim is to find the putative orthologous enhancer for a given input, it is reasonable to search for it near its orthologous genes. A peak in such a narrow region is then likely to be the enhancer of interest. The same holds true when looking for enhancers of similar activity in the vicinity of co-expressed genes. The drawbacks are that signals in unexpected regions are missed, as they are not included in the search space. And as results for aligning enhancers show, strong alignments can indeed be found away from any orthologous gene. Furthermore, genome-wide approaches require reliable significance thresholds as it otherwise is hard to determine whether a found peak is still within the genomic noise level. Unfortunately, due to the complex dependencies in word profiles such thresholds are more difficult to obtain than for alignments.

3.6 NASCAR

An enhancer detection algorithm has to solve the problem of detecting genomic regions that consist of small clusters of moderate sequence similarity (the regulatory modules) that are intercalated with stretches of sequence, which evolve almost at the rate of the genomic background [70]. Unfortunately, these spacers mostly contain more sequence than the modules they separate. They therefore hide enhancers from alignment algorithms as they evolve beyond recognizable sequence similarity while the modules themselves provide too little significance to be detected on their own. This is made even more complicated as these modules, and even individual binding sites, can rearrange by several mechanisms. Alignment-free techniques, which are thought to be able to cope with these permutations, are often blinded by too many noise motifs derived from these spacers as they score almost everything in a given window. This strongly reduces the signal-to-noise ratio and hides the region of interest in a huge amount of false-positive calls on a genome-wide scale. While many published approaches try to compensate this by providing additional information, the approach I chose in

this project tries to integrate several principles of alignment and alignment-free techniques to allow enhancer prediction on genome-wide scale based on sequence information only.

3.6.1 Mismatch extension

The initially applied alignment-free approach showed that extension of words to their maximal possible length could be used to increase the signal of individual motifs even without additional biological information. Longer words are just less likely to occur at random region in the genome and thereby have a higher weight than any set of small motifs they can be dissected in. This concept is well known from alignment algorithms but also applicable for word profiles. However, mutations in or between TFBSs can stop the extension process at word sizes that still occur more frequent in the genome than expected. This leads to artifacts that hinder the identification of the corresponding region. Alignment algorithms solve the problem of motif-internal mutations by a match-mismatch scoring function, which also allows bridging changes occurring in the motif spacers. As recent studies have shown, the structural rules governing regulatory modules can be strong enough to be kept even after full module turnover [70]. This could allow their detection by mismatch extension of words, which therefore is the most important concept implemented. Module spacers on the other hand cannot be handled in this way as they evolve at higher rates than the short spacers between motifs and span larger distances in the genome.

3.6.2 Motif weighting

As mismatch-containing motifs lead to more noise matches than perfect-match words, a more rigorous filtering is necessary. The threshold was again obtained from previous alignment-free attempts. They showed that small motifs up to 11-mers occur almost every enhancer length and thereby contribute rather to the noise than increasing the signal. I therefore set the threshold for words to the score of a perfectly matching 12-mer. This threshold is still rather permissive than restrictive: a 14-mer containing one mismatch contributes to the score although 19 different variants exist while a

perfect matching 14-mer is only 16-times less likely than a 12-mer. As mismatch motifs do not allow using the relative likelihood of individual words like applied for the alignment-free approach, I weighted motifs by simply multiplying their identity score by their length. Motifs weighting is therefore the second concept taken from the alignment-free approach.

3.6.3 Scoring

The most important aspect of a metric is of course the score calculation. Already very early in the history of sequence comparison, authors suggested to not only score individual matches, but to combine several matches in a single score [98]. Although this was initially meant for protein sequences it was implemented in a modified version of Blast developed at the Washington University (WU-Blast). Although changed to a commercial version in the meantime, a free version is still running at the EBI [99] and the basis of the EnsEMBL Web-BlastN service accessible via the EnsEMBL Genome Browser *[LINK5]*. Several statistical models to compute such a score were proposed which all can be activated in the web interface. Results obtained with the implemented classical alignment-free metrics on the other hand showed that simple counting of matching words can provide a clearer signal for comparison of word profiles than a specific statistical model (like for the POISSON metrics). This can be explained to some extent by the fact that the occurrence of words in the genome does not follow any statistical distribution, as genomes are a patchy, clustered composite of duplicated or even highly repetitive fragments. However, this has no influence on the fact that a spurious match between two sequences becomes less likely the more matching words can be identified. The scoring function of NASCAR is therefore a simple sum scoring of the detected words. Tests show that 4 validated enhancer pairs detected by NASCAR, 2 of them even having a quite similar pattern, are not detectable by WU-Blast within the set thresholds even if the implemented statistics are used (EnsEMBL BlastN-Parameters: profile “distant homologies”, modified using -statistics "-sump", -W 8, -M 2, -N -1, -Q 5, -R 2). This of course does not make any statement about the quality of the used metric, but it shows that very simple scoring techniques using little to no

previous assumptions are sufficient to detect regions of similar activity and might even allow the detection of strongly diverged regions.

3.6.4 Pattern detection

The last and most experimental concept implemented in NASCAR tries to deal with the special situation of functional clustering in enhancer regions. As described, enhancers are composed of to some extent functionally independent modules that are allowed to change their relative distances, positions and orientations. In some cases however, these modules might keep their relative positions and orientations while the spacer sequences mutate. Interestingly, the effect of unchanged spacing but high sequence turnover in the module spacers is more severe for alignment algorithms than insertions/deletions of the same size. While affine gap scoring allows a rather "cheap" bridging of longer gaps, mismatch scoring is more restrictive. LastZ for example uses 3.25 times the mismatch penalty as penalty for opening a gap – the gap extension penalty however is only ~0.24 times the mismatch penalty by default *[LINK6]*. This means that gaps >3nt (4nt gap = 3.25 + 3 x 0.24 = 3.97) would have a lower impact on the score than a mismatching region of the same length. Alignment algorithms would thereby miss correctly spaced regulatory modules even in case of perfect sequence collinearity when separated by highly diverged spacers ***(Figure 19)***.

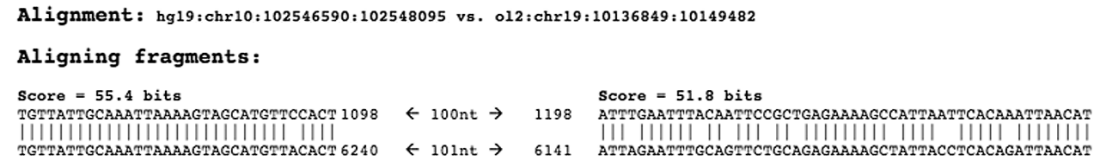


Figure 19 *Example for almost perfectly spaced matches in the same region separated too far to be recognized as a single hit by alignment algorithms. Although still alignable as such, a few more mutations would hide them in genome-wide alignments while a pattern detection method could still pick them up.*

The same holds true for alignment-based gene prediction due to their exon-intron structure. The main difference to the situation in enhancers however is that exons are dense packages of perfectly "beaded" codons that are only allowed to mutate at every 3rd position in order to keep the resulting amino

acid chain unchanged. As two matches normally score higher than a single mismatch this situation can be perfectly handled by alignment algorithms – especially as exons normally are longer than regulatory modules. The challenge was now to implement a mechanism that can detect the collinear patterns of preserved regulatory modules against the noisy background generated by the spacer sequences. Due to the fact that these arrangements can be considered as one to many blocks separated by simultaneous gaps, gap detection techniques were investigated for their potential to handle the problem. For this, a modified gap detection technique initially published for the CHAOS [100] aligner was used. CHAOS is a fast local aligner used in the preparation of alignment hits as input for the Shuffle-LAGAN algorithm [97]. This algorithm was designed for the correct alignment of larger genomic regions that have undergone local reshuffling. A very similar approach was chosen here just at a much smaller scale using less significant – and ungapped – modules. For this, the distance and shift parameters for motifs were extended and scored in an elliptic region upstream and downstream of the starting module instead of using a short, parallel section. The reasoning for this “shape” was that close by modules are less likely to have a large relative shift due to the fact that this mutation very likely would have affected one or even both modules and thereby impaired the function. Far spaced modules on the other hand were restricted to only small shifts, as too many artificial clusters would have been generated otherwise. Modules at about half the size of the input enhancer were allowed to shift the most. This method is able to identify one candidate that shows strong enhancer activity similar to the activity of the input enhancer that would have been missed otherwise. As a side effect, this technique is also able to recover gapped alignment hits that would be missed by the basic score without a significant speed reduction of the algorithm.

3.7 Prediction results

3.7.1 Alignment-free

Results obtained for alignment-free metrics show that these approaches are able to identify putative corresponding regions across large evolutionary distances even on a genome-wide scale. However, all of the regions identified are also predicted by a significant alignment when using BlastN. Focusing on more narrow, orthologous regions, determined by the genes in synteny, increases the amount of putative corresponding enhancers and even allows to predict candidates that are invisible to the alignment algorithms used. Initially, alignment-free techniques were used for database searches [61] or large-scale similarity comparisons [101]. For both tasks they perform well as they do not have to report one significant hit but to either exclude regions that are not similar at all of give an estimate about the degree of similarity of two known genomic regions – or even full genomes. Especially for the latter task they are very well suited as segmental duplication, inversion, permutation and other mutational events are common features of large genomic regions (>>1mb) and make approximations based on global alignments difficult. Motif noise, which is one of the major challenges of alignment-free methods on a narrow scale, has only little to no effect in that context as the large window sizes allow generation of densely occupied word profiles of long matching words. These carry way more signal than the short words that have to be used within enhancers. When used for enhancer prediction, it is therefore not surprising that a lot of putative corresponding regulatory regions, which can be found in a restricted locus, are lost in a genome-wide approach. This explains why most published alignment-free algorithms include supporting data into their scoring schemes and/or focused on narrow regions of confirmed orthology. In the light of the aforementioned problems inherent to the alignment-free methodology it is therefore interesting to note that 72% of all aligning regions, among them even weakly aligning regions (50 – 60 bits), could be identified by the implemented alignment-free approach on a genome-wide scale without specifying additional information.

3.7.2 NASCAR

3.7.2.1 Candidates

In this study, 377 in vivo validated enhancers, for which no significant LastZ hit is detectable, were used to identify corresponding enhancers in the medaka genome. For this set, 30 putative enhancer candidates could be selected defined by their position next to orthologous flanking genes. This selection criterion was derived from observations in the aligning sub set, showing that the majority of all still identifiable regions (194 of 252, ~77%) is either located next to one (133 of 252, ~53%) or even between both orthologous flanking genes (61 of 252, ~24%). Twelve (~5%) are still near their orthologous flanking genes and additional 32 (~13%) near at least one gene orthologous to the human enhancer locus. In total, ~94% of all alignment hits occur in the vicinity of orthologous genes. This clearly shows that the presence of orthologous genes in the vicinity, especially if they are still in flanking positions, is a strong indicator for a corresponding putative enhancer. Shifting the cutoff from only the first peak to the first 25 increases the number of putative candidates to 48. This concession had to be made due to the allowance for motif permutation in the metric. A scoring scheme that produces a stable score in case of permutation events results in a lot of different combinations scoring at the same level. Furthermore, such a metric has to include a lot of independent motifs. As each of these motifs is rather weak compared to the significance levels of alignment algorithms, this leads to inclusion of random matches that do not correspond to functional motifs. This, in turn, increases the number of equally scoring combinations even further. It is therefore necessary to include more than just the highest scoring peak when looking for enhancer candidates. To be able to discriminate between rearranged and collinear regions, an additional pattern detection technique was implemented. As no metric alone can assess both structural models at the same time, the two metrics are used in parallel. More restrictive or permissive cutoffs than 25 would also be possible. Analysis of peak location suggests that most putative candidates would be included at this level. The lowest ranked peak, which is at position 23, shows no enhancer activity and thereby might be a false positive hit – but the human counterpart,

which is strongly conserved within the placental mammals, shows no activity as well. It is therefore possible that the necessary trans environment has changed since the split of Human and Medaka. The medaka region however is not conserved and therefore might well be a false positive prediction. The next lowest candidate peak showing enhancer activity for both regions (Human and Medaka) is at position 12. This shows that even a more conservative threshold would not result in a loss of active candidates. But with an average number of ~40.000 peaks per enhancer, 25 is already rather conservative. The highest loss of candidates was caused by BlastN alignment on the non-aligning data set. This shows that 55 still alignable regions are missed by LastZ, 39 of them overlapping with predicted NASCAR peaks in the vicinity of orthologous genes. These were therefore removed from the candidate list.

In vivo validation of the remaining 8 putative enhancer candidate pairs results in 6 pairs (~75%) showing activity of both corresponding regions, the 2 remaining ones have no enhancer activity in the medaka region. This result is very similar to a TFBS-based approach on narrow regions in Zebrafish [68], which resulted in 88% (7 of 8) regions showing enhancer activity. In contrast to this approach, the regions tested here are predicted on the full medaka genome without any additional information like TFBSs. This shows that the simple principle of sum-scoring short alignment hits that are insignificant by themselves is able to reveal enhancers that would have been missed by commonly used alignment algorithms. Even when removing those that can be identified by specifically modified WU-Blast settings, still 60% (3 of 5) show enhancer activity. This is at the rate of active elements expected for enhancer prediction by deep sequence conservation (44%, [34]).

In addition to the candidates tested, more than 50 enhancers have NASCAR peaks near at least one of their orthologous flanking genes although it is not in flanking position anymore. Two of these peaks are even reported at first position. Comparison to randomly generated peak sets shows that this is within the range of what would be expected by chance. The aligning set on the other hand also has 12 peaks in that category. It is therefore possible that

those 50 peaks might contain several additional candidates. One possible explanation for those regions would be that local inversion or insertions between the former flanking gene and the enhancer led to their separation in the fish. Genomic changes might also have masked some good candidates by the deletion of “bystander genes” [33], initially located between the enhancer and its target gene, in the teleosts. If one of the other genes kept in synteny is the target gene, these enhancers could still be detectable but would have been assigned to the “orthologous gene near” category. But as more than 100 peaks within the top 25 are contained in this sub set, selection of likely enhancing peaks is very difficult.

Interestingly, for almost each of the tested candidates peaks can be found that score even higher. The reason for this is unknown. It might be that still unmasked regions in the enhancers produce high amounts of random matches in certain regions, which lead to a high NASCAR score. As the expected motif sizes are rather small (<35nt) noise is still a problem. Larger motifs might already produce an alignment on their own and are therefore unlikely to occur in the word profiles generated by NASCAR. Another possible explanation is that those regions might be redundant enhancers, which are expected to contain similar motifs. In that case, at least some of those peaks should occur near the tested peak although too far away to be assigned to the orthologous gene. As enhancers are known to act over large distances and even if several genes are in between [28], the chosen assignment threshold might just be too conservative to show that relation. Location analysis performed on the 100 highest peaks for each of the tested candidates however could not reveal any clustering pattern around the tested candidates. Nonetheless, several peaks among the 25 highest per human enhancer may be regulatory elements. For 3 VISTA enhancers (hs882:hg19, hs1344:hg19, hs865:hg19) additional NASCAR peaks further down in the ranking were tested, 2 of them not even near any orthologous gene. Despite that, all 3 showed (partially even strong) enhancing activity (see ***Figure 9*** and ***Figure 16***). However, none of them was similar to either the human enhancer or the first tested region. While this argues further against a redundant enhancer activity it indicates that more than just the predicted region near the

orthologous gene might be an enhancer – especially those regions that score higher.

In total, 9 out of 12 predicted NASCAR peaks have clear enhancing activity in the used reporter assay, even 2 peaks that are not near any annotated orthologous gene and despite a lower score than the orthologous candidate. Further validation of the remaining 25 highest-scoring NASCAR peaks for each of the 9 human enhancers might therefore reveal even more regulatory elements. This indicates that NASCAR is able to predict active regulatory elements at a high rate based on sequence features alone. However, only 3 of the tested peaks resulted in an at least partially similar pattern. But as studies have shown, even clearly orthologous enhancers conserved between Human and Zebrafish display a similar pattern in only 30% of the cases [102].

3.7.2.2 Conservation analysis

Detailed inspection of the tested candidate loci revealed that the majority of them are conserved within their clade or at least contain some conserved blocks. This is interesting as no conservation between the clades is reported. It also indicates that large changes must have happened shortly after the split, which were fixed individually within each clade. Afterwards, the majority of sequences mutated independently in each species retaining only some regions that were under constraint. These clade specific conserved blocks are either too short or too diverged to result in significant alignment hits between the clades, otherwise the corresponding human enhancers would have been contained in the aligning set. It is not very surprising however that the motifs used by NASCAR overlap to a large extent with the conserved blocks in both species. As the remaining sequence is too diverged to align even within the clade it is unlikely to match to human. Four candidates show motif-level conservation at higher levels than for any other random motif set selected in either Human or Medaka. This can be taken as evidence for motif-specific conservation and makes it likely that those motifs are involved in the function. For 2 candidates, several motifs in the medaka sequence were deleted to validate their importance for the enhancer. Unfortunately, the results obtained

by the deletion experiments do not allow any clear conclusion. Although apparently specifically conserved in Human and Medaka, the deletion of several identified, and even still collinear, motifs did not eliminate the enhancer activity of the initial construct. One could therefore state that those motifs are of no functional importance for the enhancer - but this would leave their clear conservation unexplained. On the other hand, it is well possible that due to the transient context (none of the injected embryos were raised to form a stable line), variations in spatio-temporal activity caused by the deletion are just not visible. The results obtained for hs1344:ol2-1Δ can be taken as support for this hypothesis. This construct shows a new domain neither contained in the human enhancer nor the undeleted construct, arguing for a repressive function of the deleted motifs. However, the new domain is only visible in roughly 25% of the pattern-positive fish. Further analyses of the full enhancer regions show that they contain additional conserved blocks in both species, which do not contribute to the NASCAR profile. These blocks may either contain binding sites that have diverged too far to be recognized even by NASCAR or contain lineage-specific innovations that help to stabilize the activity of the enhancer. Hence, they might have compensated the loss of the deleted motifs. However, without additional test using just the deleted motifs it is impossible to tell whether or not they are of functional importance for the enhancer. Nonetheless, it can be stated here that the used detection principle is able to reveal small-scale conservation that is missed by the tested alignment algorithms.

3.7.2.3 Motif analysis

The motif profiles of all enhancer-candidate pairs were analysed for patterns that would allow conclusions about the inner organisation of enhancers. As it seems, most candidate regions in Medaka that drive recognizable enhancer activity have very collinear motif arrangements, some even as dense clusters that are still alignable using BlastN. These alignments however are too weak to peak high enough in a genome-wide alignment ranking to be within the significance cutoff set here. Only WU-Blast is able to rank three of them high enough to be within the cutoff when sum statistics are used. This indicates

once more that consideration of multiple hits within a given region is an appropriate way of detecting otherwise lost enhancers. The fact that there is only little evidence for motif rearrangement could be explainable by the used enhancer set. As previously mentioned, most enhancers in the VISTA set have been initially detected by deep sequence conservation and are therefore unlikely to have rearranged in Medaka. The observed patterns fit more to the assumption of “sequence erosion” by mutation of the intercalated spacer sequences. This is further supported by the motif-specific conservation detected in some cases. A few motifs that co-occur with collinear motifs show mutation patterns that would fit to local inversion. But as the ancestry of these sequences cannot be traced back they might also be the result of a turnover mechanism or just spurious matches. Without further functional tests it cannot be stated whether these motifs are responsible for the observed enhancer activity. The patterns for two of the used enhancers could not be interpreted, as these regulators are roughly 4kb in size, which results in a very complex motif arrangement of which most are likely random matches.

3.7.2.4 TFBS analysis

The binding of TFs to their corresponding binding sites within the enhancer is thought to be the first step in the process of enhancer activation. But it is still unclear how TFs find the correct binding sites. As TFBSs are usually rather short and degenerate, many possible sites can exist at any given location. One possible guidance mechanism could therefore be the depletion of possible additional binding sites within an enhancer, restricting their occurrence only to correct positions that allow functional binding and activation of the enhancer. To search for patterns that argue in favour for this, a TFBS analysis was performed on the NASCAR motifs but did not show conclusive results. While statistical testing shows significant enrichment of specific TFBSs in motifs when compared between the species, this is very likely the result of the underlying sequence similarity. Tests, using the conserved blocks in teleosts and placental mammals as control could not reveal any significant motif enrichment between the species. But as these blocks are larger and at the same time of lower sequence similarity than the

NASCAR motifs, it is hard to tell which of the two parameters was responsible for this result. An equally sized negative set of NASCAR-predicted regions for comparison against the validated enhancers would be the best control but unfortunately does not exist. More detailed analysis of the predicted TFBSs indicates that the observed enrichment might be caused by the fact that these sites are rare in general. Most of the identified sites are between 10 and 17nt long, which makes another instance of the same TFBS in a small region like an enhancer rather unlikely.

Another hypothesis contrary to the aforementioned is that enhancers may be formed in regions that are enriched for TFBS precursors instead being depleted. As soon as some of these sites mutate into a configuration that allows reliable binding of a TF, this might create a beneficial enhancer activity, which in turn leads to its fixation. In this case, the identified TFBSs within the NASCAR motifs should be more similar to those in their surrounding spacers than to the spacer sequences in different enhancers. Analysis of all possible motif-spacer combinations however did not reveal any significant difference within or between enhancers.

In sum, although both analyses do not show a clear result this does not rule out that specific TFBS enrichments exist. The identified NASCAR motifs might just not appropriately reflect the functional fraction of the investigated enhancers. Furthermore, PWMs are not the best predictor for TFBSs as they neglect combinatorial co-occurrence of nucleotides at specific variable positions. It is therefore impossible to make a final statement without further functional tests or more reliable control sets. In any case, larger numbers are necessary to allow meaningful conclusions.

3.8 Conclusion

In this study, I developed a new enhancer prediction method based on a combination of alignment and alignment-free principles as each technique has strengths that can compensate some of the weaknesses of the other. The resulting algorithm was subsequently used for enhancer prediction in Medaka and able to identify active regulatory elements at a rate comparable to the rate achieved by alignment-based deep sequence conservation. This shows that motif-scoring techniques can successfully be applied for genome-wide enhancer prediction in vertebrates even without providing any additional information and in species as diverged as Human and Medaka. However, the importance of the identified motifs for the observed enhancer activity remains unclear. Despite partially even significant motif-level conservation only little to no effect was visible as consequence of motif deletion in two tested constructs. Several explanations for this result are possible, but further experiments are necessary to test these hypotheses. It is therefore not clear yet whether the algorithm indeed used the regulatory logic of the given enhancers to predict additional candidates. Furthermore, it needs to be tested whether the used principle can also be applied for regulatory elements that are less conserved in general. Most of the activity of the tested candidates was observed in neuronal tissues and the question remains whether the regulatory logic in other tissues like heart or muscle might utilize more promiscuous transcription factors with more degenerate binding specificities. This might hide important motifs even to an algorithm like the one used here. However, the results achieved by NASCAR are promising and further refinements might allow the identification of additional enhancers in even more distantly related species.

4. Materials & Methods

4.1 VISTA Enhancer set

I used the VISTA Enhancer browser [73] (state 2010-12-07) to generate a set of validated human enhancers. For this, I used the internal search routine to extract all tested hg19 regions, which showed enhancer activity at stage E11.5 in Mouse. I then further filtered the obtained set by removing of all overlapping entries to create a set of fully independent human enhancers. This resulted in 629 human regulatory regions. I split the full set further into an aligning and a non-aligning set based on LastZ. Then, I retrieved repeat masked sequences of all human enhancers via the EnsEMBL API (v63) and aligned them against the masked medaka genome retrieved in the same way (LastZ command-line parameters: --noytrim, --inner=2000, --masking=40, --chain. Other parameters see “FishScoreFile.txt” in the thesis data folder). Afterwards, I filtered the results by extracting all human enhancers for which LastZ reports at least one alignment. In total, the aligning set contains 252 of 629 human enhancers. The remaining 377 enhancers form the non-aligning data set for enhancer prediction.

4.2 Pairwise alignment pipeline

I uploaded the compiled full VISTA dataset to Galaxy *[LINK8]* [103–105] and retrieved all MAF-blocks (“Multiple Alignment File”, Galaxy parameters: Split blocks by species=”Do not split”) that contained aligning regions of the teleosts *Tetraodon nigroviridis* (assembly: tetNig2), *Takifugu rubripes* (fr2), *Gasterosteus aculeatus* (gasAcu1), *Danio rerio* (‘Zebrafish’, zv8), and *Oryzias latipes* (‘Medaka’, ol2) to Human (“Homo sapiens”, hg19). I subsequently filtered these blocks, keeping only those for which an alignment to Zebrafish is possible (always the same LastZ parameters used as before). Afterwards, I exported the resulting zebrafish regions in bed-format for further processing. For this, I removed all regions mapping to non-assembled scaffolds or contigs in the reference assembly (zv8). I then extended the centres of the remaining regions by 5kb in both directions and combined the overlapping segments into

longer, non-overlapping genomic regions. These regions were used to perform a direct hg19 to zv8 alignment for all enhancers in the VISTA set. For this, I downloaded sequences for both region sets via the EnsEMBL API (v63) and aligned them with LastZ. Next, I processed the alignment results and stored them as bed-files, containing the exact coordinates of all alignment hits in zv8 for every VISTA enhancer. For each of these zv8 regions I identified the corresponding genomic location in Medaka via EPO ("Enredo, Pecan, Ortheus" alignment pipeline, data retrieved via EnsEMBL API v63) and compiled a region set in the same way as previously described for zv8. I then aligned the zebrafish sequence of the Human-Zebrafish alignment hits against those ol2 regions using LastZ and again extracted the coordinates of resulting zebrafish-medaka hits. This way, an alignment chain from Human through Zebrafish to Medaka was established for every element that can be identified by LastZ in all three species. As a last step, I performed reciprocal LastZ alignments (ol2 vs. hg19, hg19 vs. ol2) between each ol2 element and the hg19 VISTA enhancer in the same alignment chain to test whether a direct alignment is still possible.

4.3 BlastN

I also downloaded BlastN included in the Blast+ Suite (v2.2.25, *LINK2*) from the NCBI. I then performed alignment runs against the repeat masked medaka genome (ol2) using the hg19 sequences for all enhancers in the VISTA Enhancer set as input (BlastN command-line parameters: -reward 2, -penalty -3, -gapopen 5, -gapextend 2, -word_size 7, -dust "20 64 1", -soft_masking TRUE). This served as control whether these regions are indeed non-aligning. In total, I did two independent runs using different e-value limits and subsequently filtered them to retrieve either only the highest scoring (parameter: -evalue 5e-1) hit or the 25 highest (parameter: -evalue 1) hits, respectively.

4.4 Background word counts

I obtained genome-wide word counts for different word sizes from genomic sequence of most recent assemblies (Human: hg19; Medaka: ol2) provided by EnsEMBL (v63). As the word frequencies should be later used to filter words that are likely to be repetitive, I used unmasked sequence. I first downloaded the full genomic sequence using the Perl EnsEMBL API (v63) and dissected it afterwards into words of size 5 to 10nt at nucleotide resolution. Words containing unspecified or masked nucleotides ("Ns") were excluded. These word profiles served as input for the POISSON and HEXDIFF metrics as well as basis for the calculation of frequency tracks.

4.5 Frequency tracks

To assess the word background frequencies in a given genomic region for variable word sizes I calculated genomic frequency tracks. For this, first the genomic frequencies for all 10-mers contained in the target genome were determined. Then, the background distribution of all possible frequency products (frequency of each word multiplied by the frequency of every other word) was calculated and binned into 256 categories to allow efficient storage. Afterwards, preview and review scores per nucleotide were calculated by multiplying the frequencies of the words starting at the current nucleotide and 5nt upstream or those ending at that nucleotide and 5nt downstream, respectively. The average of those scores was compared to the computed product background distribution and the corresponding frequency category assigned to the current nucleotide ***(Figure 20)***. As boundaries (e.g chromosomal start/end or repeat masked regions) allow calculation of only one of the two scores, this score was used to assign the category. The very first or last 4nt next to a boundary, for which neither of the scores is calculable, are assigned based on the squared frequency of the word starting/ending at the nucleotide. All scores were stored bitwise as binary file with each bit representing the frequency category of the corresponding nucleotide in the given sequence. The whole procedure is implemented in the "CompileFrequencyTrack.pl" script.

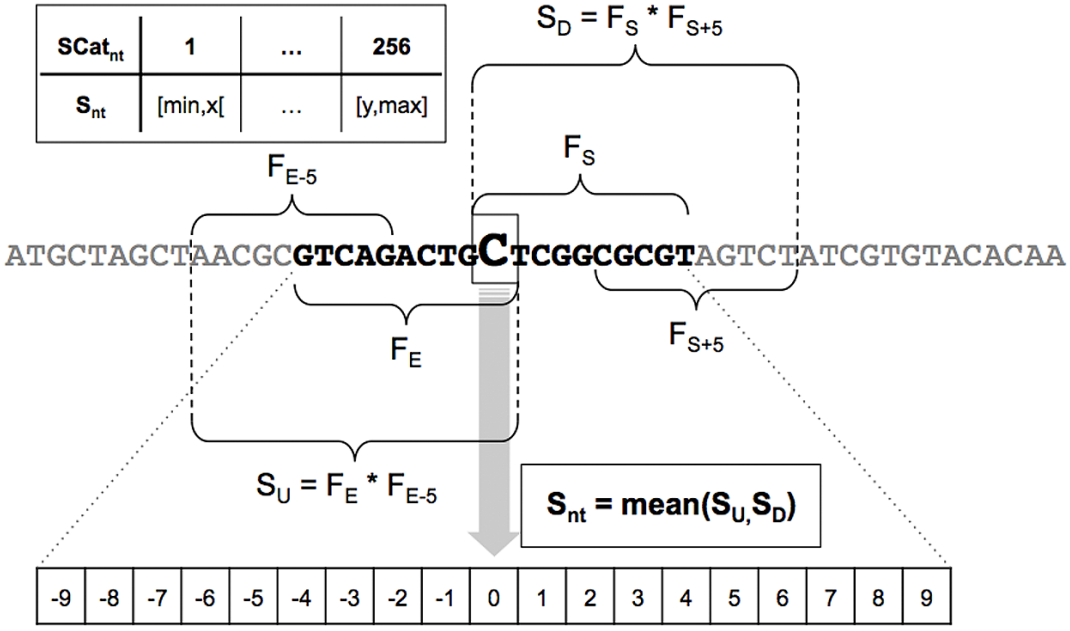


Figure 20 *Frequency track generation. For the current nucleotide in the sequence ("C" in the centre box) the frequencies of four words (F_{E-5}, F_E, F_S, F_{S+5}) are extracted from the background distribution to calculate the upstream (S_U) and downstream score (S_D). Those are used to compute the final value (S_{nt}) which is looked up in the table containing the boundaries for each of the 256 categories (upper left). The corresponding category value is than written in the frequency track at the position of the corresponding nucleotide.*

4.6 Gene sets

I obtained sets of all protein coding genes for Human and Medaka from EnsEMBL via the Perl API (v63) using the "GeneDownloader.pl" and "FilterByBiotype.pl" scripts. For this, I first downloaded all genes per chromosome and filtered them for protein coding genes afterwards (command-line parameter: --biotype protein_coding). These gene sets were later used for the generation of orthologous regions and the evaluation of peaks predicted by the implemented algorithms.

4.7 Orthoblocks

Orthoblocks are regions on a chromosome in the target species that start and end with an ortholog of a gene in the query species. To generate these regions, I used EnsEMBL Compara via the EnsEMBL Perl API (v63). Method

is implemented in the "Orthify.pl" script. First, the 5 closest flanking genes up- and downstream of each human enhancer in the generated VISTA data set were identified based on their transcription start site (TSS) using the previously generated gene sets. Additionally, all genes within 1.5mb up- and downstream were retrieved (command-line parameters: --min_genes 5, --min_range 1500000). For each gene in the union of both sets all orthologous genes in Medaka were retrieved using the API. Known ortho-paralogous genes were not taken into account. The identified orthologous genes per enhancer were subsequently grouped into sets for each represented chromosome and the coordinates for the most upstream/downstream boundary of the most upstream/downstream gene determined. These coordinates were subsequently further extended by 1.5mb. These chromosomal regions form a set of restricted search spaces for each enhancer in the input set.

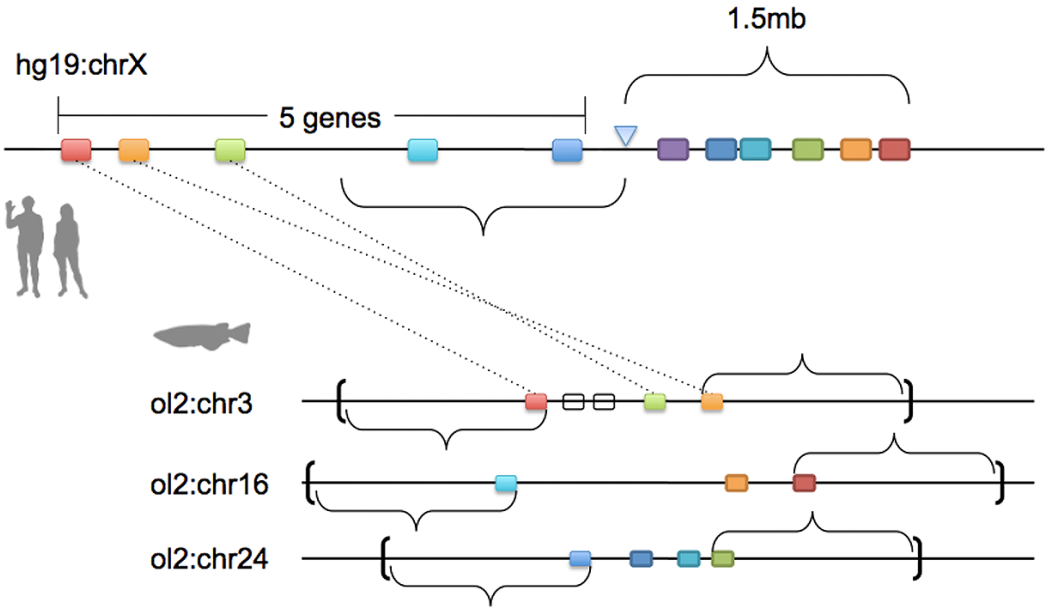


Figure 21 *Orthoblocks (brackets indicate orthoblock boundaries). Orthoblocks can contain interspersed non-orthologous genes (empty boxes) and permuted gene order (dotted lines).*

4.8 Alignment-free metrics

To test the advantages and disadvantages of alignment-free techniques, I implemented different classical metrics in the "MultiMetricScanner.pl" script. Scores for those metrics for a given sequence comparison were obtained the following: first, the query sequence was read and dissected into words of

specified size starting at every single nucleotide to generate a word profile of the input sequence. As for the word profiles, words containing unspecified nucleotides ("N"s) were excluded. Profile was then transformed into a double-stranded profile by combining the counts for each word and its reverse complement ("rcWord") in the same strand and assigning this number to the respective word-rcWord pair. Target sequence was then read and the maximum number of windows for given window size and stepping calculated (default: window size is equals to the query sequence length, stepping is 25% of used window size). Per window, the word profile was generated in the same way as for the query sequence. To calculate word background frequencies (for POISSON) and ratios (for HEXDIFF), word counts previously extracted from the target genome were given to the program. After score calculation, all continuous regions of a score higher than: median + 3 * median absolute deviation (per metric) were considered as peaks.

Definitions for equations: Q = Query genome, q = query (enhancer) sequence, T = Target genome, t = target (window) sequence, k = word size, c = word count, f = word frequency, m = number of all different words for given word size. Frequencies are calculated by

$$f_i^x = \frac{c_i^x}{\sum_{j=1}^{m} c_j^x}$$

Equation 1 *Frequency calculation*

with x = given sequence

4.8.1 COSINE

COSINE metric was taken from [62] but initially used by [101]:

$$COSINE(q,t) = \frac{\sum_{i=1}^{m}(c_i^q \cdot c_i^t)}{\sqrt{\sum_{i=1}^{m}(c_i^q)^2} \cdot \sqrt{\sum_{i=1}^{m}(c_i^t)^2}}$$

Equation 2 *COSINE metric*

4.8.2 D2

I implemented the D2 metric as described in [62] using word sizes k-1,k,k+1:

$$D2(q,t) = \sum_{i=k-1}^{k+1} \sum_{j=1}^{m_i} (c_{ij}^{q} - c_{ij}^{t})$$

Equation 3 *D2 metric*

with i = current word size, m_i = all words of size i, and c_{ij} = counts for word j of size i.

4.8.3 POISSON

The four different POISSON metrics were initially published by [79]. As they need background frequencies for the individual words, these were calculated based on the genome-wide word profiles given to the program using ***Equation 1***. Furthermore, for each word the expected number of occurrences is calculated by

$$\lambda_i = f_i^t \cdot (L_W - k + 1)$$

Equation 4 *Expected amount per word in given sequence*

with L_w = window length.

Word-based similarity for “POISSON:Additive” and “POISSON:Product is calculated by

$$c_i^{qt} = min(c_i^q, c_i^t)$$

Equation 5 *Common counts per word in query and target sequence*

$$P(x \geq c_i^{qt}) = \begin{cases} (1 - F(c_i^{qt} - 1, \lambda_i))^2 & if\ c_i^{qt} > 0 \\ 1 & otherwise \end{cases}$$

Equation 6 *Probability for a word to occur as or more often than expected in target sequence (F = Poisson distribution function)*

$$s_i^{qt} = 1 - P(x \geq c_i^{qt})$$

Equation 7 *Contribution of a single word to overall similarity*

Finally, the four metrics are calculated by:

$$Additive(q,t) = \frac{1}{m}\sum_{i=1}^{m} s_i^{qt}$$

Equation 8 *POISSON:Additive metric*

$$Product(q,t) = 1 - \sqrt[m]{\prod_{i=1}^{m} P(x \geq c_i^{qt})}$$

Equation 9 *POISSON:Product metric*

$$Distinct(q,t) = \frac{1}{m}\sum_{i=1}^{m} \left| F(c_i^t, \lambda_i) - F(c_i^q, \lambda_i) \right|$$

Equation 10 *POISSON:Distinct metric*

$$Overrepresented(q,t) = \frac{1}{m}\sum_{i=1}^{m} \left| F(c_i^t - 1, \lambda_i) - F(c_i^q - 1, \lambda_i) \right|$$

Equation 11 *POISSON:Overrepresented metric*

4.8.4 HEXDIFF

The HEXDIFF metrics was initially published by [64]. It uses word ratios describing how many times more or less than expected a specific word is contained in the query sequence. These ratios are than used to weight each individual word.

$$r_i = \frac{f_i^q}{f_i^T}$$

Equation 12 *Ratio calculation for all words in i = {1 … m}*

Only those words that are contained in the target window and the query profile can contribute to the score:

$$HEXDIFF(q,t) = \sum_{i=1}^{m}(c_i^t \cdot r_i)$$

Equation 13 *HEXDIFF metric*

4.8.5 Modified metric

The modified alignment-free metric I tested here generates word profiles in a bit different way than classical alignment-free metrics. Instead of dissecting both sequences independently into word profiles of a fixed size, both sequences are directly compared to extract perfect matching words of variable size. For this, the query sequence is first dissected in the classical way into words of fixed size, which form anchor points or “seeds” that are further extended in both directions afterwards. Extension proceeds as long as adjacent nucleotides perfectly match between query and target. This results in a set of words of various sizes. Words of low complexity are then further filtered. For this, each word is dissected further into overlapping 5-mers. These 5-mers are then tested for self overlap. Each word that contains more self-overlapping than non-overlapping 5-mers is excluded from the profile. This procedure is similar to the one used in [71]. Afterwards, words are mapped to the target sequence starting with the longest one. Only words of equal size are allowed to overlap during this process. Shorter words are truncated to their non-overlapping core and assessed again later. In case their length drops below seed size they are discarded. Overlap in query sequence is not assessed, which allows words in the query to occur repeatedly in the target but not vice versa. After all words are mapped, the profile is split into sub-profiles for the different contained word sizes and two parameters assessed per sub-profile:

$$WC_k = \frac{N_k^t}{L^t}$$

Equation 14 *Word coverage*

$$WI_k = \frac{N_k^t}{k \sum_{i=1}^{m_k} c_i^t}$$

Equation 15 *Word independence*

With N_k^t = number of nucleotides in window mapped to words of size k

These values are multiplied with a weighting factor (a^k, default: a =1.5, k = word size) that allows adjusting the contribution of the individual sub-profiles to the final score (for a = 1 all sub-profiles would be treated equally). Furthermore, the average frequency value for each sub profile is calculated by

$$WAF_k = \frac{1}{|I_k^t|} \sum_{i \in I_k^t} SCat_{nt,i}$$

Equation 16 *Word average frequency*

with I_k^t = indices of all nucleotides contributing to N_k^t ($SCat_{nt,i}$ see ***Figure 20***).

The final sub-profile score is then calculated by

$$SC_k = \frac{WC_k \cdot WI_k \cdot a^k}{WAF_k}$$

Equation 17 *Final score for profile formed by all words of size k*

to reduce the contribution of highly repetitive genomic regions. The final similarity score for query and target window is then the sum of all sub-profile scores:

$$SCORE = \sum_{k_{min}}^{k_{max}} SC_k$$

Equation 18 *Final similarity score for given window*

The full metric is implemented in the “WurmTDS.pl” script. For calling and evaluation of peaks see “Peak calling & evaluation” for NASCAR.

4.9 NASCAR

4.9.1 Profile generation

NASCAR motif profiles are generated in a similar way as described for the modified alignment-free metric. As before, query sequence is dissected into a word profile of given word size (default: k=8nt). Then, the same procedure is repeated for the full target sequence for performance reasons. Each word in the query is then mapped to every instance in the target and overlapping words of the same shift in query and target (i.e. these words overlap in the same diagonal of a dotplot) are collapsed into one word. The resulting profile is identical to that for the alignment-free approach and serves as “seeds” for subsequent mismatch extension. Extension is performed independently in both directions (upstream/downstream) using a simple additive match-mismatch scoring function until the motif score drops below 0 (i.e. accumulated mismatches score higher than all matching nucleotides). Both extensions are then truncated to the shortest, highest scoring region starting at the seed and merged to form the final motif. Each motif is thereby regarded as a mismatch-containing word similar to perfect matching words in alignment-free algorithms. The same procedure is repeated for the reverse complement target sequence and both generated profiles merged. This profile is then filtered to remove overlapping words in the target sequence. Like for the modified alignment-free metric, motifs are first ranked by their score and then mapped to the target, starting with the highest scoring. The score of each motif is determined by

$$s_i = p_i \cdot s_{match} + q_i \cdot s_{mismatch} \cdot L_i$$

Equation 19 *NASCAR motif score*

with i = {1 … M}, M = all motifs in target window, pi, qi = matching/mismatching nucleotides in motif, s = score of match/mismatch, and Li = length of motif i. Overlapping motifs are truncated to their next matching nucleotide and rescored as long as they still contain a perfect matching region of at least seed size. The fully filtered word set forms the final word profile (see ***Suppl. Figure 5***).

4.9.2 Score calculation

NASCAR score is calculated in a windowed fashion along the given target sequence as done for alignment-free algorithms (default: window size = enhancer length, stepping 25%). For this, all words within the current target window are extracted from the profile in a "fuzzy" fashion, allowing words to overlap the window boundaries. The extracted profile is then filtered for word overlap in the query sequence identical to the target filtering procedure. This results in a final profile of completely independent words. All words above or equals the minimal score cutoff (default: score of a perfect matching 12-mer) are then used to calculate the final similarity score for the target window ("PURE-score", M_{Valid} = all motifs above threshold after filtering):

$$Score_{PURE} = \sum_{i \in M_{valid}} s_i$$

Equation 20 *NASCAR "PURE score"*

4.9.3 Pattern detection

In a parallel approach, clusters of collinear motifs within a certain distance and query-target shift are traced (default: distance = 200, shift in diagonal = 25). For this, all motifs above the score threshold are ranked again in decreasing order. Then, starting at the strongest motif, two elliptical motif-spaces along the current motif diagonal and overlapping in one focus are computed, with the motif placed in the overlapping focus ***(Figure 22)***.

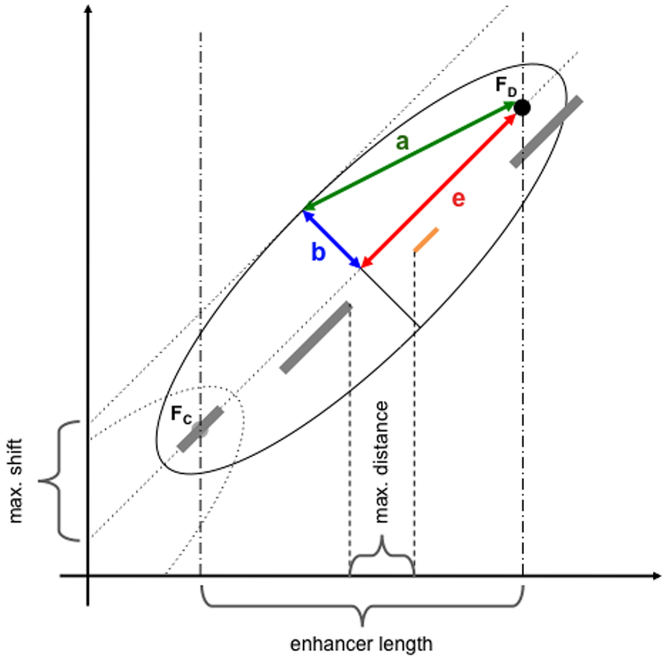


Figure 22 *Pattern detection technique. For each motif above threshold (grey bars) two elliptic spaces (upstream ellipse drawn only dotted; a,b,e are standard values of ellipses) are computed with the motif being located in the overlapping focus point (F_C). All motifs within those spaces that are also within the set intermotif distance form a pattern even if they are below the score threshold (orange motif). At least 3 motifs have to be combined in that way, 2 of them above score threshold.*

The distance between the two foci in query or target corresponds thereby to the enhancer length. This allows detection of clusters spanning the full window size. All motifs that are with their centre within these motif spaces are selected and tested for their inter-motif distance (end ←→ start). Motifs with an inter-motif distance smaller or equals the specified distance are combined into a motif cluster. All clusters of three motifs containing at least two words scoring at or above the threshold level are valid. This allows inclusion of motifs even below threshold as long as the previous requirement is met. Combined score of all motifs in all identified clusters is finally added to the previously computed PURE-score to form the final score ("COMB-score").

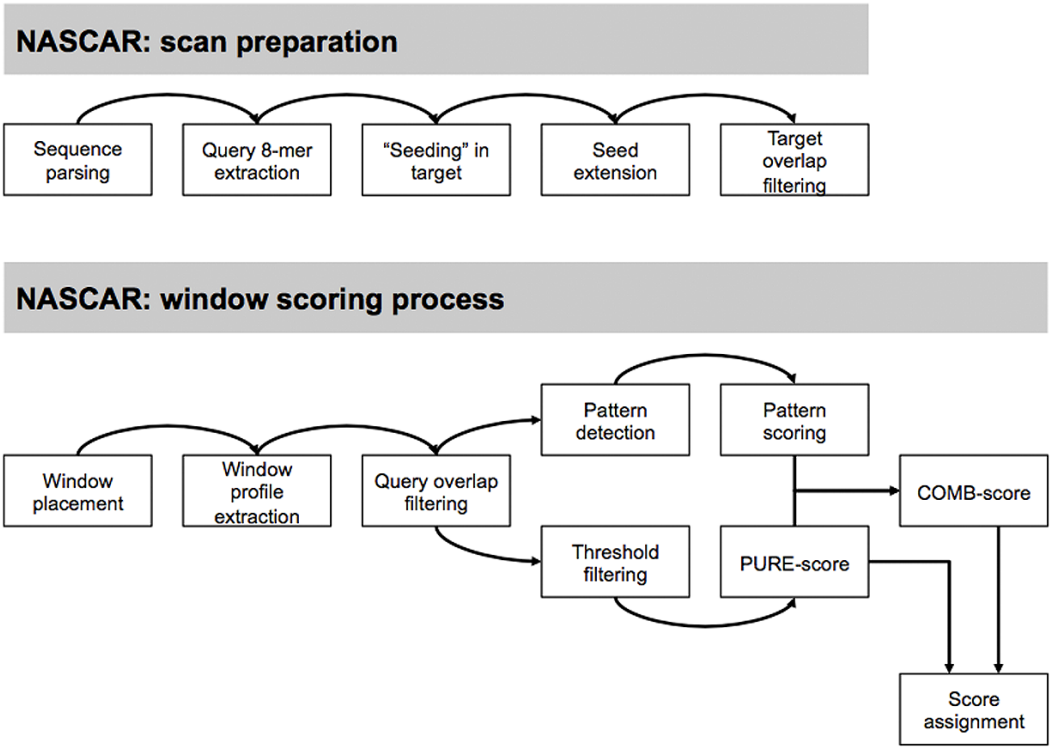


Figure 23 *NASCAR process diagram*

4.9.4 Peak calling & evaluation

All regions that have a NASCAR score above a set threshold are considered as peaks. These peaks are continuous intervals in the target sequence starting at the first window scoring above the threshold and ending as soon as the score drops below this limit. Default threshold is 3 times the median absolute deviation (MAD) above the median NASCAR score for the corresponding run. Median and MAD are independently calculated for each run of a human enhancer against the medaka genome. Called peaks are then evaluated based on 5 different categories: "double flanked" (orthologs of both flanking genes in human also flank the medaka peak), "single flanked", "near flank" (ortholog of a flanking gene in Human is still within the set distance and gene cuttoffs in Medaka but not directly flanking the peak), "not flanked" (orthologous gene within the cutoff but not flanking in human), and "not orthologous" (not a single orthologous gene near, but ortho-paralogous genes would be possible). The full procedure is done independently for "PURE" and "COMB" score and implemented in the "EvaluateRace.pl" script. Evaluation of the results for the modified alignment-free metric is performed in the same

way with the exception that the “near flank” category does not exists. This procedure is implemented in the “EvaluateCrawl.pl” script.

4.10 Random motif sets

The procedure for random motif generation is implemented in the “RandomizeMotifsInRegion.pl” script. Random sets were generated by randomly placing motifs of the same size as the real motifs within the given enhancer region in a non-overlapping fashion. In this way 10 independent random sets for each validated sequence (9 human and 10 medaka regions) were generated, which were subsequently used for conservation and TFBS analysis.

4.11 Conservation

The conservation of NASCAR motifs was analyzed by averaging across the compiled conservation information of all nucleotides forming a motif. The conservation data (PhastCons scores) for placental mammals (mammal sub set of the 46-way MultiZ vertebrate alignment) and teleosts (5-way MultiZ alignment) was obtained from the UCSC Genome Browser *[LINK3, LINK4]*. Resulting data was then further analyzed using R 2.8.1. The conserved blocks within the individual enhancer-candidate pairs were obtained using the UCSC TableBrowser (PARAMETERS: clade=”Mammal”, genome=”Human”, assembly=”hg19”, group=”Comparative Genomics”, track=”Conservation”, table=”phastConsElements46wayPlacental”, output format=”BED – browser extensible data”; clade=”Vertebrate”, genome=”Medaka”, assembly=”ol2”, group=”Comparative Genomics”, track=”Most Conserved”, table=”phastConsElements5way”, output format=”BED – browser extensible data”).

4.12 TFBSs

To analyse NASCAR motifs for the presence of putative TFBSs I modified a script existing in the lab (Juan Mateo, personal communication). The PWM scoring function of that script was adapted from [106], PWMs for the analysis were obtained from JASPAR and TRANSFAC. The threshold for calling

TFBSs was set to 90% percentage identity to the given PWM (command-line parameter: --tfbs_threshold 0.9). Whole procedure was implemented in the "TFBSscan.pl" script.

4.13 Cloning & in vivo validation

All constructs were amplified from the human or medaka genome by PCR using Phusion DNA Polymerase (Finnzyme/Biozym, F-530L) and the primers specified in ***Suppl. Table 1*** (programs see Appendix). Resulting fragments were extracted and purified from 1% Agarose gel (Agarose: Biozym, 840004) using the Analytik Jena innuPREP DOUBLEpure kit (REF: 845-KS-5050250). Constructs created via sticky-end ligation ***(Suppl. Table 2)*** were digested using HindIII (Fermentas; ER0501) after PCR and ligated by T4 ligase (Fermentas, EL0014) into the p339:HSP70:eGFP expression vector (Plasmid stock no. 1955). Remaining constructs were cloned into blunted p339:HSP70:eGFP vector. For this, vector was cut using ClaI (Fermentas, ER0141), sticky ends filled by Klenow enzyme (Roche, Cat.No.11008404001), and dephosphorylated using FastAP (Fermentas, EF0654). For blunt ligations, PEG4000 contained in the T4 ligase kit (Fermentas, EL0014) was used in addition (protocol see Appendix). Ligated constructs were used to transform in-house generated chemical competent MachT1 cells. Injection constructs were generated following the MidiPrep protocol of the QIAGEN Plasmid Purification kit. Purified constructs were subsequently mixed with I-SCE (NEB, R0694S) and injected into 100+ medaka embryos. All injected embryos were monitored from 1 – 10dpi (days post injection) and analysed for a consistent GFP expression pattern. Lens activity of the HSP70 promoter served as injection control allowing discrimination between unsuccessful injections and inactive putative enhancer elements. Due to the usually mosaic activity in injected fish, an enhancer was considered to be active if at least 25% of all lens positive fish showed a consistent activity pattern. Images of injected embryos were taken on an OLYMPUS MVX10 binocular at 4x magnification using a LEICA DFC500 camera.

5. References

1. Darwin C, Wallace A (1858) On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. Journal of the Proceedings of the Linnean Society of London Zoology 3: 45–62.
2. Morgan TH, Sturtevant A, Muller H, Bridges C (1915) The mechanism of Mendelian heredity.
3. McClintock B (1929) A Cytological and Genetical Study of Triploid Maize. Genetics 14: 180–222.
4. Griffith F (1928) The Significance of Pneumococcal Types. J Hyg (Lond) 27: 113–159.
5. McCarty M, Avery OT (1946) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J Exp Med 83: 97–104.
6. Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol 36: 39–56.
7. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737–738.
8. Wilkins MHF, Stokes AR, Wilson HR (1953) Molecular structure of deoxypentose nucleic acids. Nature 171: 738–740.
9. Franklin RE, Gosling R (1953) Molecular configuration in sodium thymonucleate. Nature 171: 740–741.
10. Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, et al. (1965) RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc Natl Acad Sci USA 53: 1161–1168.
11. Beadle GW, Tatum EL (1941) Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci USA 27: 499–506.
12. Jacob F, Perrin D, Sánchez C, Monod J, Edelstein S (2005) [The operon: a group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729]. C R Biol 328: 514–520.
13. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188: 107–116.
14. Jacob F (1977) Evolution and tinkering. Science 196: 1161–1166.
15. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
16. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
17. King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164: 788–798.
18. Fields C, Adams MD, White O, Venter JC (1994) How many genes in the human genome? Nat Genet 7: 345–346.
19. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, et al. (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet 25: 239–240.

20. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.
21. Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, et al. (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440: 1050–1053.
22. Marcellini S, Simpson P (2006) Two or four bristles: functional evolution of an enhancer of scute in Drosophilidae. PLoS Biol 4: e386.
23. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, et al. (2010) Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. Science 327: 302–305.
24. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61.
25. Tung J, Primus A, Bouley AJ, Severson TF, Alberts SC, et al. (2009) Evolution of a malaria resistance gene in wild primates. Nature 460: 388–391.
26. Sims RJ, Reinberg D (2009) Processing the H3K36me3 signature. Nat Genet 41: 270–271.
27. Khoueiry P, Rothbächer U, Ohtsuka Y, Daian F, Frangulian E, et al. (2010) A cis-regulatory signature in ascidians and flies, independent of transcription factor binding sites. Curr Biol 20: 792–802.
28. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377–1419.
29. Butler JE, Kadonaga JT (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev 15: 2515–2519.
30. Kadonaga JT (2002) The DPE, a core promoter element for transcription by RNA polymerase II. Exp Mol Med 34: 259–264.
31. Tokusumi Y, Ma Y, Song X, Jacobson RH, Takada S (2007) The New Core Promoter Element XCPE1 (X Core Promoter Element 1) Directs Activator-, Mediator-, and TATA-Binding Protein-Dependent but TFIID-Independent RNA Polymerase II Transcription from TATA-Less Promoters. Molecular and Cellular Biology 27: 1844–1858.
32. Anish R, Hossain MB, Jacobson RH, Takada S (2009) Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements. PLoS ONE 4: e5103.
33. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, et al. (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Research 17: 545–555.
34. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502.
35. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, et al. (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol 3: e157.
36. Romano LA, Wray GA (2003) Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting

components of transcriptional regulation. Development 130: 4187–4199.
37. Brown CD, Johnson DS, Sidow A (2007) Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science 317: 1557–1560.
38. Arnosti DN, Kulkarni MM (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94: 890–898.
39. Panne D, Maniatis T, Harrison SC (2007) An atomic model of the interferon-beta enhanceosome. Cell 129: 1111–1123.
40. Swanson CI, Evans NC, Barolo S (2010) Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. Dev Cell 18: 359–370.
41. Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.
42. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet 4: e1000106.
43. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854–858.
44. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell 10: 1453–1465.
45. Roeder RG (2005) Transcriptional regulation and the role of diverse coactivators in animal cells. FEBS Lett 579: 909–915.
46. Visel A, Akiyama JA, Shoukry M, Afzal V, Rubin EM, et al. (2009) Functional autonomy of distant-acting human enhancers. Genomics 93: 509–513.
47. Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, et al. (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat Genet: 1–10.
48. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462: 58–64.
49. Vázquez AV, Blanco M, Zaborowska J, Soengas P, González-Siso MI, et al. (2011) Two proteins with different functions are derived from the KlHEM13 gene. Eukaryotic Cell 10: 1331–1339.
50. Fujita T, Fujii H (2012) Transcription start sites and usage of the first exon of mouse Foxp3 gene. Molecular biology reports.
51. Castro DS, Martynoga B, Parras C, Ramesh V, Pacary E, et al. (2011) A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes Dev 25: 930–945.
52. Wunderlich Z, Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet 25: 434–440.
53. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, et al. (2010) ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet 42: 806–810.

54. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106–1117.
55. Jin F, Li Y, Ren B, Natarajan R (2011) Enhancers: multi-dimensional signal integrators. Transcription 2: 226–230.
56. de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. Genes Dev 26: 11–24.
57. Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2: 100–109.
58. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302: 413.
59. Visel A, Bristow J, Pennacchio LA (2007) Enhancer identification through comparative genomics. Semin Cell Dev Biol 18: 140–152.
60. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, et al. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40: 158–160.
61. Hide W, Burke J, Davison DB (1994) Biological evaluation of d2, an algorithm for high-performance sequence comparison. J Comput Biol 1: 199–215.
62. Vinga S, Almeida J (2003) Alignment-free sequence comparison-a review. Bioinformatics 19: 513–523.
63. Su J, Teichman S, Down T (2010) Assessing Computational Methods of Cis-Regulatory Module Prediction. PLoS Comput Biol: 1–15.
64. Chan BY, Kibler D (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. BMC Bioinformatics 6: 262.
65. Sosinsky A, Honig B, Mann RS, Califano A (2007) Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci USA 104: 6305–6310.
66. Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, et al. (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. Dev Cell 17: 568–579.
67. Arunachalam M, Jayasurya K, Tomancak P, Ohler U (2010) An alignment-free method to identify candidate orthologous enhancers in multiple Drosophila genomes. Bioinformatics 26: 2109–2115.
68. Taher L, Mcgaughey DM, Maragh S, Aneas I, Bessling SL, et al. (2011) Genome-wide identification of conserved regulatory function in diverged sequences. Genome Research 21: 1139–1149.
69. He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. PLoS Comput Biol 5: e1000299.
70. Swanson CI, Schwimmer DB, Barolo S (2011) Rapid Evolutionary Rewiring of a Structurally Constrained Eye Enhancer. Curr Biol 21: 1186–1196.
71. Leung G, Eisen MB (2009) Identifying cis-regulatory sequences by word profile similarity. PLoS ONE 4: e6901.
72. Göke J, Schulz MH, Lasserre J, Vingron M (2012) Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. Bioinformatics 28: 656–663.

73. Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res 35: D88–92.
74. Harris RS (2007) Improved pairwise alignment of genomic DNA.
75. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15: 1034–1050.
76. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research 14: 708–715.
77. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. (2003) Human-mouse alignments with BLASTZ. Genome Research 13: 103–107.
78. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
79. van Helden J (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20: 399–406.
80. Burke J, Davison D, Hide W (1999) d2_cluster: a validated method for clustering EST and full-length cDNAsequences. Genome Research 9: 1135–1142.
81. Davidson DB, Burke JF (2001) Brute force estimation of the number of human genes using EST clustering as a measure. IBM Journal of Research and Development 45: 439–447.
82. Hufton AL, Mathia S, Braun H, Georgi U, Lehrach H, et al. (2009) Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. Genome Research 19: 2036–2051.
83. Frankel N, Davis GK, Vargas D, Wang S, Payre F, et al. (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature: 1–5.
84. Mongin E, Auer TO, Bourrat F, Gruhl F, Dewar K, et al. (2011) Combining computational prediction of cis-regulatory elements with a new enhancer assay to efficiently label neuronal structures in the medaka fish. PLoS ONE 6: e19747.
85. Wittbrodt J, Meyer A, Schartl M (1998) More genes in fish? BioEssays 20: 511–515.
86. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20: 110–121.
87. Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L (2004) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Research 14: 472–477.
88. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100: 11484–11489.
89. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, et al. (2010) Homotypic clusters of transcription factor binding sites are a key

component of human promoters and enhancers. Genome Research 20: 565–577.
90. Koohy H, Dyer NP, Reid JE, Koentges G, Ott S (2010) An alignment-free model for comparison of regulatory sequences. Bioinformatics 26: 2391–2397.
91. Bourque G, Leong B, Vega VB, Chen X, Lee YL, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Research 18: 1752–1762.
92. Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nat Methods 4: 563–565.
93. Gordân R, Narlikar L, Hartemink AJ (2010) Finding regulatory DNA motifs using alignment-free evolutionary conservation information. Nucleic Acids Res 38: e90.
94. Cande J, Goltsev Y, Levine MS (2009) Conservation of enhancer location in divergent insects. Proc Natl Acad Sci USA 106: 14414–14419.
95. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet.
96. Barrière A, Gordon KL, Ruvinsky I (2011) Distinct Functional Constraints Partition Sequence Conservation in a cis-Regulatory Element. PLoS Genet 7: e1002095.
97. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, et al. (2003) Glocal alignment: finding rearrangements during alignment. Bioinformatics 19 Suppl 1: i54–62.
98. Karlin S, Altschul SF (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci USA 90: 5873–5877.
99. Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W (2003) WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res 31: 3795–3798.
100. Brudno M, Morgenstern B (2002) Fast and sensitive alignment of large genomic sequences. Proc IEEE Comput Soc Bioinform Conf 1: 138–147.
101. Stuart GW, Moffett K, Baker S (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18: 100–108.
102. Ritter DI, Li Q, Kostka D, Pollard KS, Guo S, et al. (2010) The importance of being cis: evolution of orthologous fish and mammalian enhancer activity. Mol Biol Evol 27: 2322–2332.
103. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Research 15: 1451–1455.
104. Goecks J, Nekrutenko A, Taylor J, Team G (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86.

105. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol Chapter 19: Unit 19.10.1–21.
106. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576–3579.

Web links:

LINK1: http://enhancer.lbl.gov/frnt_page_n.shtml

LINK2: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.25/

LINK3: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/

LINK4: http://hgdownload.cse.ucsc.edu/goldenPath/oryLat2/phastCons5way/

LINK5: http://www.ensembl.org/index.html

LINK6: http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html#options_scoring

LINK7: http://genomewiki.ucsc.edu/index.php/Human/hg19/GRCh37_46-way_multiple_alignment

LINK8: https://main.g2.bx.psu.edu/

Appendix

A. PCR-Programs:

Standard program:

10” @ 98°C	→	20” @ 68°C	→	2’ @ 72°C	x5
10” @ 98°C	→	20” @ 63°C	→	2’20” @ 72°C	x25
10’ @ 72°C	➔	END		

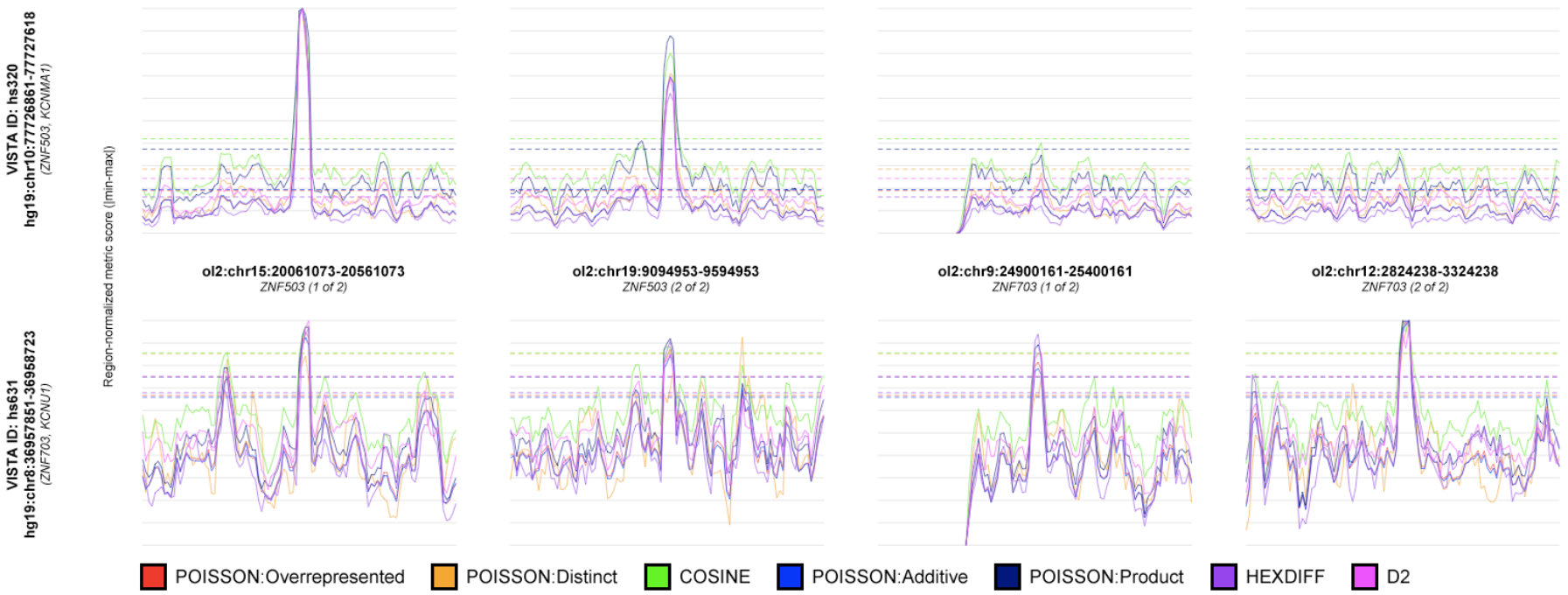
Program for long (~4kb) fragments:

10” @ 98°C	→	20” @ 63°C	→	3’ @ 72°C	x5
10” @ 98°C	→	20” @ 60°C	→	3’30” @ 72°C	x25
10’ @ 72°C	➔	END		

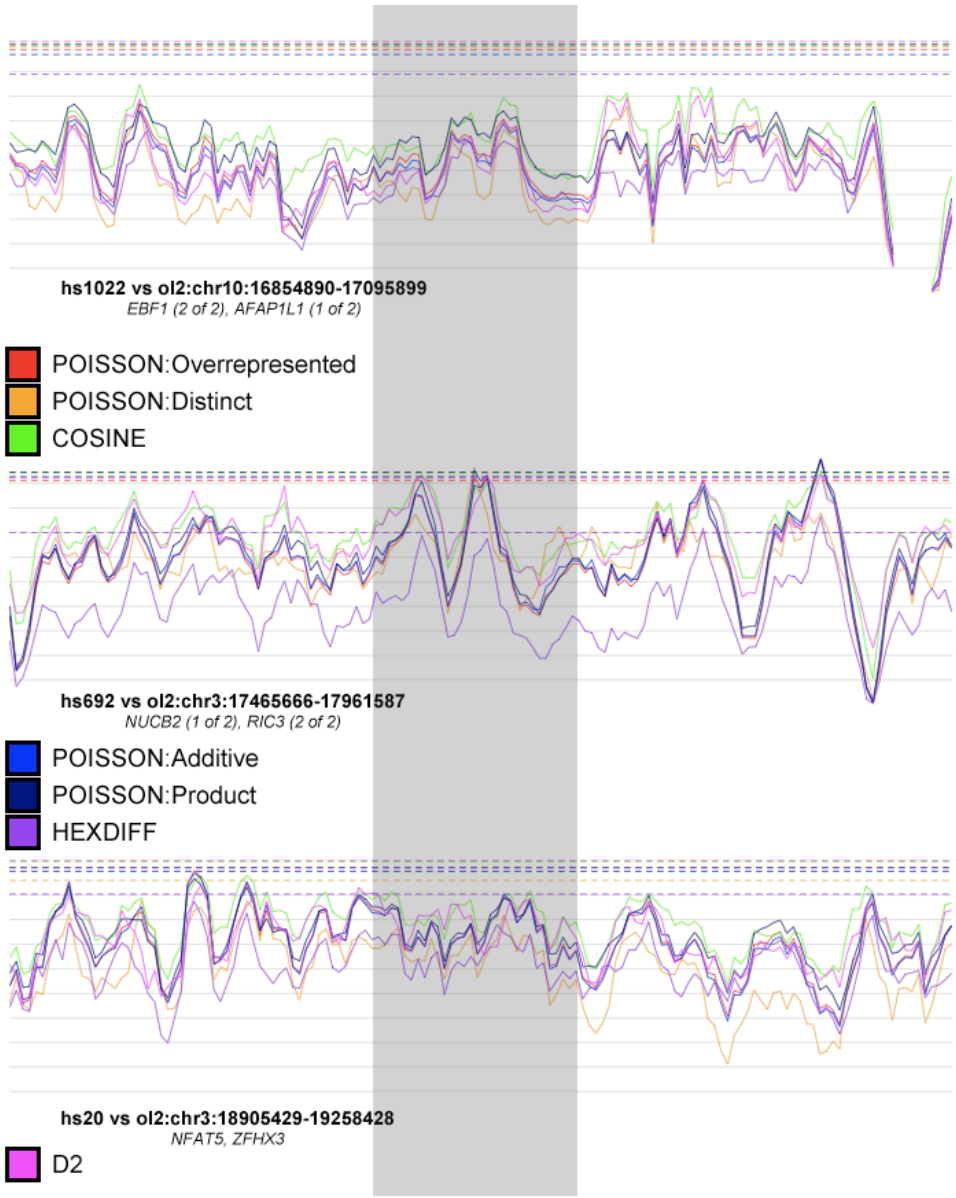
B. Blunt ligation protocol:

14µl	PCR product
1µl	blunted Vector
2µl	10x ligation buffer
1µl	PEG4000
1µl	T4 Ligase
1µl	dH2O

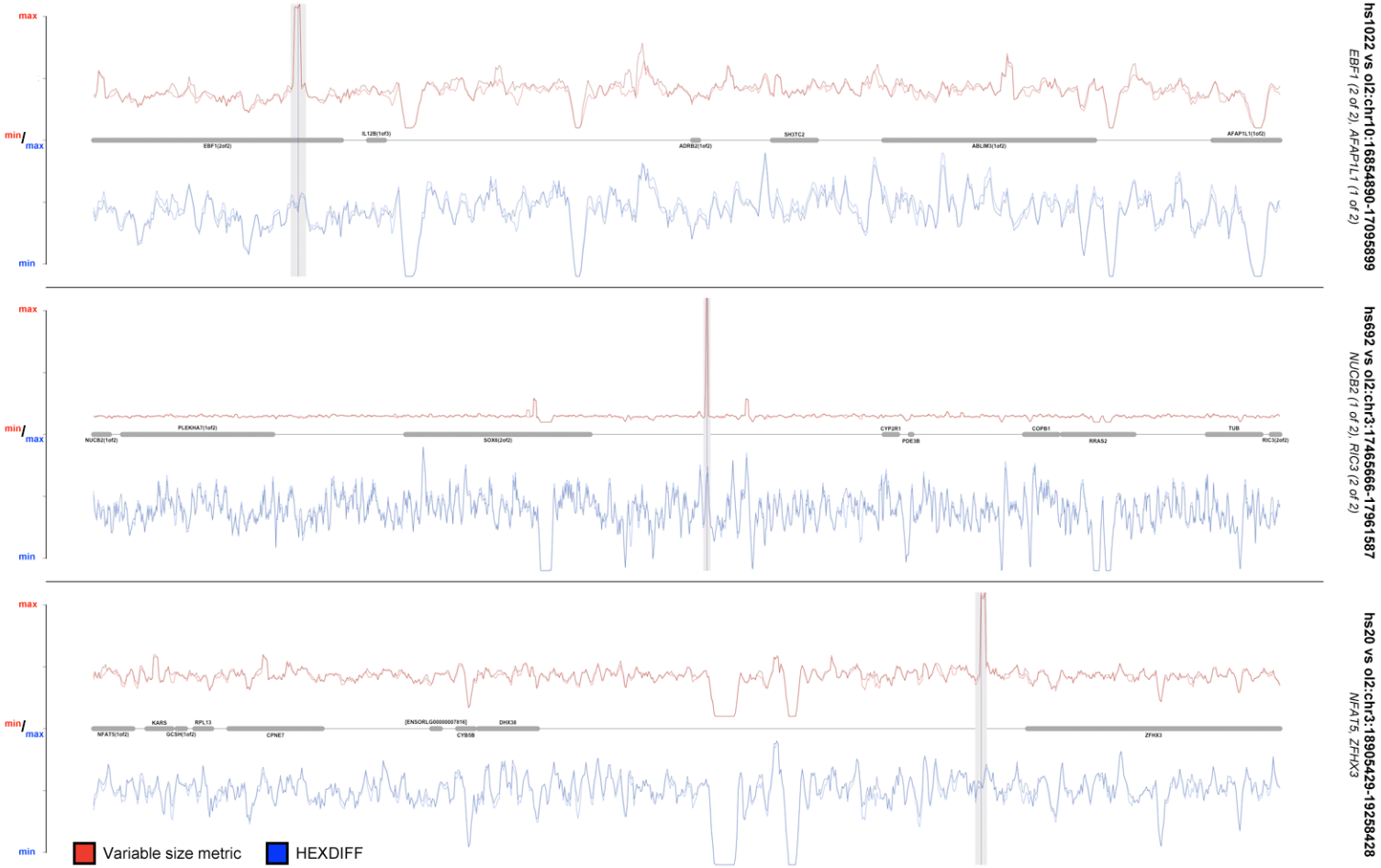
→ @ RT for 2h



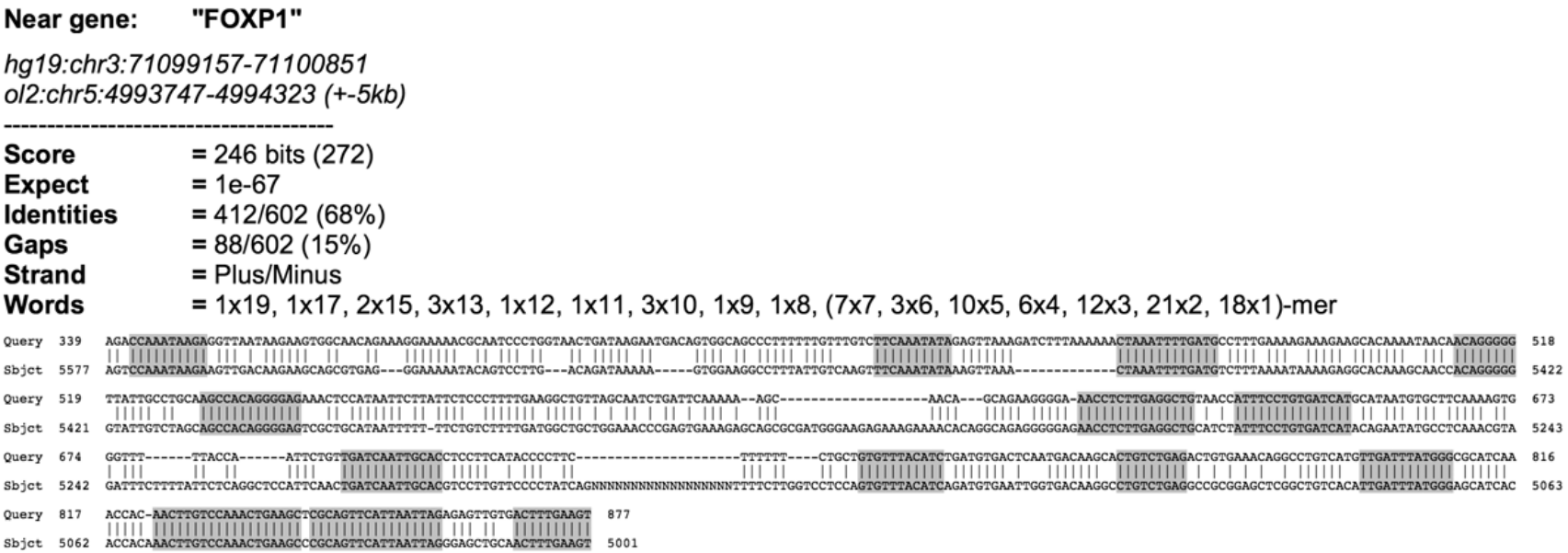
Suppl. Figure 1 *Score plots for all alignment-free metrics on each putative orthologous/ortho-paralogous target region in Medaka. Scores for hs320 show clear peak above threshold (dashed line) in both orthologous loci (2 upper left plots) but none in the ortho-paralogous regions (2 upper right plots). Scores for hs631 (lower row) are generally weaker and more noisy than for hs320 but in all regions a peak above threshold is visible.*



Suppl. Figure 2 *Score plots for all alignment-free metrics focused on the aligning regions (vertical grey bar) in the orthologous loci for all three candidates identified by the alignment pipeline (hs1022, hs692, hs20). Only for hs692 (middle plot) some metrics score above threshold (dashed line). HEXDIFF is very close to the threshold but does not fully reach it. Interestingly, most metrics peak at a close by position in that locus*



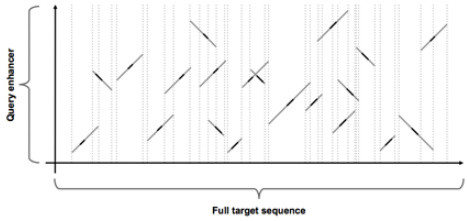
Suppl. Figure 3 *Score plots for modified alignment-free metric using variable words sizes (red) compared to the best-scoring classical metric (HEXDIFF; blue). For all three alignment-pipeline candidates, clear peaks at the alignment positions (grey bars) are visible in the modified metric whereas HEXDIFF reports only noise. Genes in those loci are drawn in scale as horizontal grey blocks. Double lines per metric show scores on masked and unmasked sequence indicating there is no big difference in those loci.*



Suppl. Figure 4 *Additional example for word pattern in significantly aligning region*

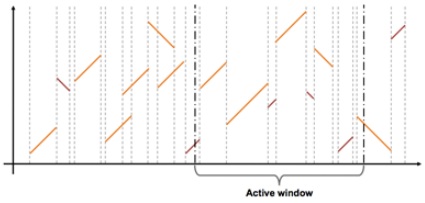
Step 1: raw profile extraction

Query sequence is dissected into 8-mers, followed by word-mapping to the target sequence, forming **"seeds"**.These are subsequently extended as far as possible, resulting the "raw" motif profile of the full target sequence.



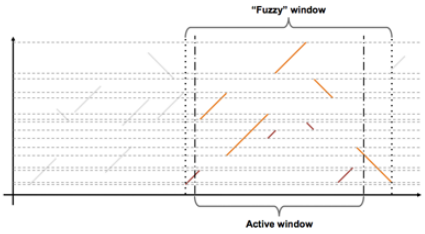
Step 2: target overlap filtering

After the profile is generated, all motifs are ranked by their score and mapped to the target sequence in decreasing order. Smaller motifs overlapping larger ones are either truncated or deleted if too short after truncation.This results in a filtered set of motifs above and below threshold.



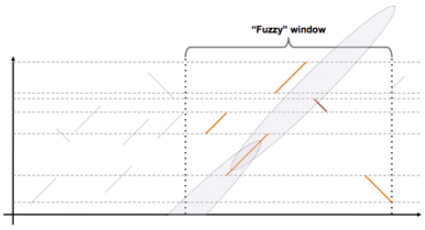
Step 3: window profile extraction

For each target window, the window profile is extracted from the full target profile. Motifs overlapping the window boundaries are also selected, generating a "fuzzy" window. Overlap in the query sequence is subsequently resolved like for the full target profile.



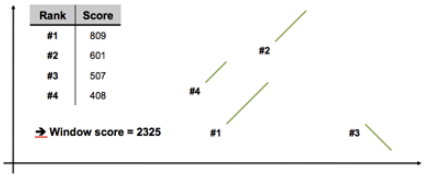
Step 4: pattern detection

Fully filtered (query & target) window profile is finally analysed for patterns of at least 3 collinear motifs, 2 have to be above threshold. These motifs need to be located in the search space around the motif on which the pattern is started. Each motif above threshold, that is not already contained in a, pattern can serve as starting point.



Step 5: profile scoring

If not contained in a pattern, all motifs below threshold are discarded prior to scoring. Scores of remaining motifs are summed and form the final window score **("PURE-score")**. In case one or more patterns are detected, the summed scores of each pattern are added to the PURE-score, resulting in the **"COMB-score"**.



Suppl. Figure 5 *Graphical display of NSACAR motif filtering process*

VISTA ID	Primer no.	Restriction site	Insert name	Organism	Target region	FWD primer	REV primer
hs394	#00037	HindIII	hg19_VAP#73_E	Homo sapiens	hg19:chr2:59746377-59746992	GACCTTAAGCTTCTTCTTCAGGAAATTCAAAA	GACCTTAAGCTTAAGCATATCCTGCCAAGGAA
	#00057	HindIII	ol2_VAP#73_T25_OrthP1	Oryzias latipes	ol2:chr15:7016932-7017682	GACCTTAAGCTTATCCCAACTTGACTCCATAATAAGC	GACCTTAAGCTTGAACCAGCGAGCTTCTACACC
hs1535	#00049	HindIII	hg19_VAP#76_E	Homo sapiens	hg19:chr2:60498057-60502013	GACCTTAAGCTTCTTGCATCCTAGGCTGTTCC	GACCTTAAGCTTTAGCGTAACAGAGGCCAAGC
	#00070	ClaI	ol2_VAP#76_T25_OrthP1	Oryzias latipes	ol2:chr15:7226761-7230869	GACCTTATCGATAAGGGACTTGTGAGCATCTTTTGG	GACCTTATCGATTTTGATTTCTGAGCCTCATTTAGCC
hs1344	#00050	HindIII	hg19_VAP#164_E	Homo sapiens	hg19:chr3:193660817-193662478	GACCTTAAGCTTTCAAGCCAGCCTCACTTTTT	GACCTTAAGCTTAAATCTGAGAGCCCCACCTT
	#00059	HindIII	ol2_VAP#164_T25_OrthP1	Oryzias latipes	ol2:chr4:13568963-13570775	GACCTTAAGCTTAGCGATTATGGGTTGGAATTAAGG	GACCTTAAGCTTCCTTTCAGGTGGTATCACAGC
	#00073	HindIII	ol2_VAP#164_T25_OrthoP2	Oryzias latipes	ol2:chr22:2742505-2744249	GACCTTAAGCTTCCTTTGAAATCATAATCAGACAGCTGCC	GACCTTAAGCTTGCCAGTGAAGAGTCAACCAGCTGCG
	#00077	HindIII	ol2_VAP#164_T25_OrthoP2	Oryzias latipes	ol2:chr22:2742401-2744352	GACCTTAAGCTTTATGAAATCCAGGACTGCTGCTCTGC	GACCTTAAGCTTCAGTATTTAACAGAAGAATGGTCCGC
	#00078	None		Oryzias latipes	DELETION of ol2:chr4:13570261-13570662 on hs1344:ol2-1	ATCAGGCCCAATCAGAATGATGTTTGTGTTTTTAGGGTGTCTGCG	AACACAAACATCATTCTGATTGGGCCTGATGGACTGAAGGG
hs865	#00052	HindIII	hg19_VAP#250_E	Homo sapiens	hg19:chr6:50685244-50686237	GACCTTAAGCTTGAATGTCCTTTTTCTCTTTTATTAC	GACCTTAAGCTTGAGGAATCCCTAGAGCTGGAAA
	#00061	HindIII	ol2_VAP#250_T25_OrthP1	Oryzias latipes	ol2:chr24:19553758-19554746	GACCTTAAGCTTAGCAAAGAGGCCTGGAGATCC	GACCTTAAGCTTCATCTCTGAAAGGTTAATTGACTGG
	#00074	ClaI	ol2_VAP#250_T25_OrthoP2	Oryzias latipes	ol2:chr12:12323733-12324922	GACCTTATCGATTTTGGGAAAGATTGTTTGATGAAAAAGGG	GACCTTATCGATTGATGGAGCTTTCAGGAGGAAACTAATCC
	#00079	None		Oryzias latipes	DELETION of ol2:chr24:19553879-19554005 on hs865:ol2-1	CTTCAGCTGCGAACGCGACCGGGGCTCCACGAGACCGC	AGCCCCGGTCGCGTTCGCAGCTGAAGAAAAGGAAGAGAGC
hs882	#00054	HindIII	hg19_VAP#469_E	Homo sapiens	hg19:chr13:71533037-71534195	GACCTTAAGCTTCACCAAAGAGAACTGCAAAGGATATTTCAA	GACCTTAAGCTTTGTGCTAAAAATTCCCATCAAA
	#00071	ClaI	ol2_VAP#469_T25_OrthP1	Oryzias latipes	ol2:chr21:9408900-9410191	GACCTTATCGATGCACATGAGCTATTGTTTTTATCGG	GACCTTATCGATTGTTTTAAAGTGTTTTTCCAGAGGG
	#00072	ClaI	ol2_VAP#469_T25_OrthP2	Oryzias latipes	ol2:chr21:9414626-9415925	GACCTTATCGATAAAATAGACCATTTATTGATGGTGC	GACCTTATCGATCATGAATCCAAACATAAAAGTAACC
hs848	#00055	HindIII	hg19_VAP#529_E	Homo sapiens	hg19:chr16:51491799-51493025	GACCTTAAGCTTCACCATGGGCTCTTTCTTTTCCTCAC	GACCTTAAGCTTCAAAATAACAGCAACGACGACA
	#00065	HindIII	ol2_VAP#529_T25_OrthP1	Oryzias latipes	ol2:chr3:29251568-29252933	GACCTTAAGCTTGTAACCTCTGTATATAGTCTAAATTGG	GACCTTAAGCTTGTGAATGATTTATATACTTTTATCTCCC
hs590	#00056	HindIII	hg19_VAP#580_E	Homo sapiens	hg19:chr18:34719386-34720720	GACCTTAAGCTTCCAAGTATGCCCAGAAATGGTA	GACCTTAAGCTTAGCATCTGATGGAGCTGTTAAA
	#00066	HindIII	ol2_VAP#580_T25_OrthP1	Oryzias latipes	ol2:chr5:16698574-16700002	GACCTTAAGCTTACTCACCCATCAATAGTCAATCC	GACCTTAAGCTTACATGCCCTTGTTTTACTCTTTTGC
hs1049	#00068	EcoRV	hg19_VAP#219_E	Homo sapiens	hg19:chr5:92314781-92316083	GACCTTGATATCCCACACCTTTTACAATAGAAAGGAAA	GACCTTGATATCGAGAGTGGGGTATGTTATAATCTGGA
	#00060	HindIII	ol2_VAP#219_T25_OrthP1	Oryzias latipes	ol2:chr9:15278072-15279513	GACCTTAAGCTTCTGGAGGAAACGGAGAAGTGG	GACCTTAAGCTTTGATCGCTGAGTCAGTTTGTAACG
hs1831	#00069	ClaI	hg19_VAP#302_E	Homo sapiens	hg19:chr7:95236622-95240458	GACCTTATCGATCACCAGCCTTGGTTTGTGTAGCA	GACCTTATCGATAAAGGGGGAGAAGGAGACAA
	#00062	HindIII	ol2_VAP#302_T25_OrthP1	Oryzias latipes	ol2:chr11:9757716-9761696	GACCTTAAGCTTCAGTTCCGCTCTTCAGTGAGG	GACCTTAAGCTTGGATGCATATTTGTTTGACCAAGC

Suppl. Table 1 *Primer table for all primers used to generate the tested NASCAR constructs*

VISTA ID	Construct	Plasmid Stock ID	FWD/REV Primer	Cloning	Organism
hs394	hs394:hg19	**3279**	*#37*	HindIII; sticky-end	Homo sapiens
	hs394:ol2-1	**3280**	*#57*	HindIII; sticky-end	Oryzias latipes
hs1535	hs1535:hg19	**3285**	*#49*	Blunt PCR ligation	Homo sapiens
	hs1535:ol2-1	**3286**	*#70*	Blunt PCR ligation	Oryzias latipes
hs1344	hs1344:hg19	**3271**	*#50*	HindIII; sticky-end	Homo sapiens
	hs1344:ol2-1	**3272**	*#59*	HindIII; sticky-end	Oryzias latipes
	hs1344:ol2-2	**3359**	*#73/#77*	Blunt PCR ligation	Oryzias latipes
	hs1344:ol2-1Δ	**3360**	*A: #59/#78* *B: #78/#59* *C: #59*	Fusion of A +B using C; blunt ligation	Oryzias latipes
hs865	hs865:hg19	**3273**	*#52*	HindIII; sticky-end	Homo sapiens
	hs865:ol2-1	**3274**	*#61*	HindIII; sticky-end	Oryzias latipes
	hs865:ol2-2	**3358**	*#74*	Blunt PCR ligation	Oryzias latipes
	hs865:ol2-1Δ	**3361**	*A: #61/#79* *B: #79/#61* *C: #61*	Fusion of A +B using C; blunt ligation	Oryzias latipes
hs882	hs882:hg19	**3335**	*#54*	HindIII; sticky-end	Homo sapiens
	hs882:ol2-1	**3336**	*#71*	Blunt PCR ligation	Oryzias latipes
	hs882:ol2-2	**3337**	*#72*	Blunt PCR ligation	Oryzias latipes
hs848	hs848:hg19	**3275**	*#55*	HindIII; sticky-end	Homo sapiens
	hs848:ol2-1	**3276**	*#65*	HindIII; sticky-end	Oryzias latipes
hs590	hs590:hg19	**3277**	*#56*	HindIII; sticky-end	Homo sapiens
	hs590:ol2-1	**3278**	*#66*	HindIII; sticky-end	Oryzias latipes
hs1049	hs1049:hg19	**3283**	*#68*	Blunt PCR ligation	Homo sapiens
	hs1049:ol2-1	**3284**	*#60*	HindIII; sticky-end	Oryzias latipes
hs1831	hs1831:hg19	**3287**	*#69*	Blunt PCR ligation	Homo sapiens
	hs1831:ol2	**3288**	*#62*	Blunt PCR ligation	Oryzias latipes

Suppl. Table 2 *All cloned and injected NASCAR constructs*